

数据挖掘中决策树算法的探讨

唐华松, 姚耀文

(华南理工大学 计算机系, 广东 广州 510640)

摘要: 决策树算法是 DM 的一个活跃的研究领域。首先给出了 DM 中决策树算法的基本思想, 然后讨论了决策树算法中的难点问题, 提出了利用熵与加权熵的思想来选择取值的算法。

关键词: 数据挖掘, 决策树, 熵

中图分类号: TP301.6

文献标识码: A

文章编号: 1001-3695(2001)08-0018-02

Research on Decision Tree in Data Mining

TANG Hua-song, YAO Yao-wen

(Dept. of Computer Science, South China University of Technology, Guangzhou Guangdong 510640, China)

Abstract: Decision Tree is one of heated fields in Data Mining in recent years. This paper first gives the main thoughts of algorithm of Decision Tree in Data Mining, then discusses the difficult problem of selecting value on division in Decision Tree, and put forward an algorithm using the thoughts of entropy and weighted entropy to solve the problem with the examples.

Key words: DM, Decision tree, Entropy

1 引言

数据库技术的迅速发展以及数据库管理系统的广泛应用, 导致人们积累了越来越多的数据。巨增的数据背后蕴藏着丰富的知识, 而目前的数据库技术虽可以高效地实现数据的查询、统计等功能, 但却无法发现数据中存在的关系和规则, 无法根据现有的数据预测未来的发展趋势。数据库中存在着大量的数据, 却缺乏挖掘数据背后隐藏的知识的手段, 出现了“数据爆炸而知识贫乏”的现象。

在此背景下, 数据库知识发现(KDD)及其核心技术—数据挖掘(DM)便应运而生了。KDD的研究内容是, 能自动地去处理数据库中大量的原始数据, 从中挖掘搜索出具有规律、富有意义的模式。它的发现过程主要有三个步骤: 定义要发现的问题; 根据问题进行数据搜索、模式抽取; 评价所发现的知识的好坏。三者之中, 核心技术是第二步, 即数据搜索及模式抽取方法。KDD = 问题处理 + DM + 解释评价。由于问题处理和解释评价的研究较成熟, 所以目前 KDD 的研究和实现难点重点都集中在核心的 DM 上。

DM 的核心技术算法主要有统计分析方法、神经网络、决策树方法、遗传算法等。其中, 决策树是一种常用于预测模型的算法, 它通过将大量数据有目的地分类, 从中找到一些具有商业价值的、潜在的信息。

2 决策树的基本思想

决策树的结构, 顾名思义, 就像一棵树。它利用树的结构将数据记录进行分类, 树的一个叶结点就代表某个条件下的一个记录集, 根据记录字段的不同取

值建立树的分支, 在每个分支子集中重复建立下层结点和分支, 便可生成一棵决策树。

例如, 我们要分析一个网站的用户接受某项新服务的情况, 可以从中选取 100 个用户, 其中 50 个接受这项新服务的, 50 个拒绝这项新服务的, 然后通过建立决策树来分析用户的情况, 寻找一些潜在的规则信息。

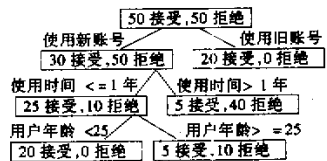


图 1 网站某项新服务的决策树结构

利用决策树进行分析, 可以容易地找到一些具有商业价值的潜在的规则信息。如在上例中, 从决策树结构图可以看出: 在接受这项新服务的用户中有 60% 是使用新帐号的, 在拒绝这项新服务的用户中有 100% 是使用旧帐号的; 也就是说, 如果用户是使用新帐号的, 那么他就有 60% 的可能接受这项新服务, 如果用户是使用旧帐号的, 那么他就有 100% 的可能拒绝这项新服务。当然, 还可以从决策树中找到其它的规则信息, 这里就不再举例说明了。

3 决策树的技术难点

建决策树, 就是根据记录字段的不同取值建立树的分支, 以及在每个分支子集中重复建立下层结点和分支。建决策树的关键在于建立分支时对记录字段不同取值的选择。选择不同的字段值, 会使划分出来的记录子集不同, 影响决策树生长的快慢以及决策树结构的好坏, 从而导致找到的规则信息的优劣。

可见 ,决策树算法的技术难点也就是选择一个好的分支取值。利用一个好的取值来产生分支 ,不但可以加快决策树的生长 ,而且最重要的是 ,产生的决策树结构好 ,可以找到较好的规则信息。相反 ,如果根据一个差的取值来产生分支 ,不但减慢决策树的生长速度 ,而且会使产生的决策树分支过细 ,结构性差 ,从而难以发现一些本来可以找到的有用的规则信息。

怎样的取值才算一个好的取值呢？一个好的取值 ,就是决策树根据此值来分裂时 ,产生的分支子集中的记录在预测内容上尽可能一致。何谓子集中的记录在预测内容上尽可能一致呢？举个例子 ,我们对学生的思想品德情况进行分析 ,预测内容是在思想品德上是优或良 ,抽取 10 个学生记录 ,其中 5 个学生的思想品德是优 ,另 5 个的是良 ,那么我们可以得到以下两个不同的划分：

学号	成绩	学号	成绩	学号	成绩	学号	成绩
01	优	03	良	01	优	02	优
02	优	05	良	03	良	04	优
04	优	06	良	05	良	06	良
07	优	08	良	07	优	08	良
10	优	09	良	09	良	10	优
(A)				(B)			

图 2 学生思想品德情况的两个划分

在 (A) 方案中 ,划分后的左分支子集的记录的思想品德都是优 ,右分支子集的记录的思想品德都是良 ,即左分支 100% 优、0% 良 ,右分支 100% 良、0% 优 ,子集中的记录在预测内容上已尽可能一致。我们就可以从两个分支中寻找记录的共性 ,以得到和学生思想品德情况有关的信息。

在 (B) 方案中 ,划分后的左分支子集中 3 条记录的思想品德是优 ,2 条记录的思想品德是良 ,右分支子集中 2 条记录的思想品德是优 ,3 条记录的思想品德是良 ,即左分支 60% 优、40% 良 ,右分支 60% 良、40% 优 ,子集中的记录在预测内容上的一致性差 ,还不能有效地总结出和学生思想品德情况有关的信息。

怎样找到好的取值呢？从上例 ,可以看出 ,好的取值分支后子集的记录的一致性良好 ,也就是记录的有序性好。通常 ,在系统学上 ,经常使用“熵”来表示事物的无序度。所以在这里 ,也可以引用“熵”来估量分支后子集有序性的好坏。熵 $H = -\sum (P_i) \times \lg(P_i)$ 其中 P_i 是指分支子集中记录取某值的可能性。

如果子集的熵值越小 ,则子集记录的有序性越好 ,如果熵值越大 ,则有序性越差。因此 ,我们可以对根据不同取值进行分裂后的左右分支的子集分别计算各自的熵值 ,选择熵值最小所对应的记录字段的取值 ,这就是好的取值。如上例中 ,

$$H_{左,右} = (-5/5) \times \lg(5/5) + (-0/5) \times \lg(0/5) = 0.0$$
$$H_{左,右} = (-3/5) \times \lg(3/5) + (-2/5) \times \lg(2/5) = 0.3$$

要提出注意的是 ,如果左右分支子集的记录数相差太远 ,则可能导致新的情况 ,此时只用熵值来判断可能选择不到好的取值。如上例 ,也可以得到以下两个划分：

- (C) 左分支 : 优
右分支 : 优 , 优 , 优 , 良 , 良 , 良 , 良 , 良
(D) 左分支 : 优 , 优 , 优 , 优 , 良
右分支 : 优 , 良 , 良 , 良 , 良

$$H_{左} = (-1/1) \times \lg(1/1) + (-0/1) \times \lg(0/1) = 0.00$$
$$H_{右} = (-4/9) \times \lg(4/9) + (-5/9) \times \lg(5/9) = 0.99$$
$$H_{左} = (-4/5) \times \lg(4/5) + (-1/5) \times \lg(1/5) = 0.72$$
$$H_{右} = (-4/5) \times \lg(4/5) + (-1/5) \times \lg(1/5) = 0.72$$

取左右分支的和平均值 , 则

$$H_{平} = (0 + 0.99) / 2 = 0.50$$
$$H_{平} = (0.72 + 0.72) / 2 = 0.72$$

选择小值 ,可能就选择 (C) 方案 ,但从例子可以看出 (D) 方案才是较好的 ,因为它把同类的基本划分到一起了 ,而且如果像 (C) 方案那样 ,每次都只把一个数据划分出去 ,只会导致决策树结构的层次变得复杂 ,同类型的记录无法有效地归到一起 ,不利于从中发掘潜在的信息。

要解决这个问题 ,可以利用“加权”的思想 ,根据分支子集所占的比重来给它一个权值 ,然后计算加权熵 ,通过比较加权熵的大小来选择取值。

加权熵 $H_{加} = \sum X_i \times H_i$
其中 X_i 是指分支子集所占的比重 , H_i 是指分支子集的熵 ,则

$$H_{加} = 9/10 \times 0.99 + 1/10 \times 0.0 = 0.89$$
$$H_{加} = 1/2 \times 0.72 + 1/2 \times 0.72 = 0.72$$

4 实验事例

在实验事例中 ,我们构造一个包括 10 条记录的学生总评成绩的模型 ,以分析学生总评成绩在 85 分以上与何因素有关。我们提出两个方案 (I) 方案每次选择第一个取值来产生分支 (II) 方案利用加权熵算法选择取值来产生分支。通过对两个方案产生的决策树进行比较 ,可以了解加权熵算法的优点。

学号	01	02	03	04	05	06	07	08	09	10
性别	F	M	M	F	M	M	M	F	F	M
年龄	21	20	21	22	23	20	21	23	20	21
体育成绩	A	A	B	A	A	A	B	B	B	A
思想品德	优	良	优	优	良	优	良	优	优	良
发表论文	2	0	0	1	0	2	0	0	1	0
平时成绩	95	90	80	80	75	85	95	80	70	80
总评成绩	95	85	80	85	70	85	90	75	70	75

图 3 学生总评成绩的情况

在图 4 中 ,Y 表示学生的总评成绩在 85 分以上 ,N 表示学生的总评成绩在 85 分以下。由图 4 可见 ,方案 (II) 构造的决策树结构好 ,基本上将总评成绩在 85 分以上或以下的同类学生划分到一起 ,容易得出“如果学生的平时成绩在 85 分以上 ,他的总评成绩就有 100% 的可能在 85 分以上”、“如果学生的平时成绩在 85 分以下 ,他的总评成绩就有 5/6 即 83.3% 的可能在 85 分以下”等规则信息。 (下转第 22 页)

(上接第 19 页)

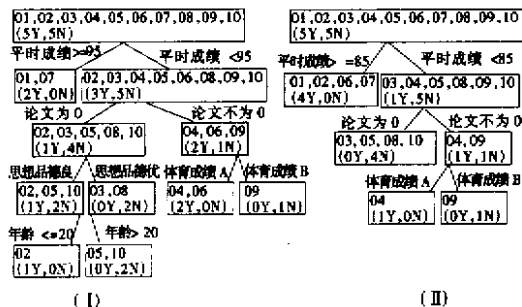


图 4 两个方案的决策树结构

实验中两个方案的比较表明,利用加权熵算法来选择决策树的分支取值,不但可以加快决策树的生长,而且最重要的是可以得到结构好的决策树,便于从中挖掘好的规则信息。虽然计算加权熵需要一定的时间开销,但随着记录数据的增大,这点开销不但因为决策树的快速生长得到弥补,而且会使总的运算时间减少,从而提高算法的效率和性能。

万方数据

5 小结

为了在数据挖掘的决策树算法中,解决决策树分支取值的难点问题,我们提出了结合事例利用熵、加权的思想来选择取值的算法。实验表明,利用这个算法,可以提高决策树的生长速度,优化决策树的结构,发掘较好的规则信息。特别是在使用决策树算法来挖掘的数据越多,算法的效率和性能就越好,算法的优越性就越明显。

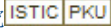
参考文献:

- [1] [美] Harjinder S GILL. 数据仓库—客户/服务器计算指南 [M]. 王仲谋, 刘书舟. 北京: 清华大学出版社, 1997.
- [2] Alex Berson, Setphen J Smith. Data Warehousing, Data Mining, & OLAP [M]. McGraw - Hill Book Co., August, 1999.

作者简介:

唐华松(1976-),女,硕士研究生,研究方向为分布式计算。

数据挖掘中决策树算法的探讨

作者: [唐华松, 姚耀文](#)
作者单位: [华南理工大学计算机系](#)
刊名: [计算机应用研究](#) 
英文刊名: [APPLICATION RESEARCH OF COMPUTERS](#)
年, 卷(期): 2001, 18(8)
被引用次数: 82次

参考文献(2条)

1. [HarjinderSGill 数据库-客户/服务器计算指南](#) 1997
2. [Alex Berson;Setphen J Smith Data Warehousing, Data Mining, & OLAP McGraw - Hill Book Co](#) 1999

相似文献(10条)

1. 学位论文 [赵翔 数据挖掘中决策树分类算法的研究](#) 2005

决策树方法是数据挖掘中一种重要的分类方法。本课题从新的建树准则、决策树修剪、多变量决策树、多决策树组合、不完备信息系统下的模型建立等几个方面对决策树方法进行了研究和探讨。

本课题的主要研究工作和成果有:

- 1、针对传统决策树算法的不足(如ID3、C4.5),提出了基于协方差及高阶相关系数的决策树生成算法,避免了经典的以信息熵作为建树准则的决策树生成算法盲目地偏向于属性值较多的属性的缺点。
- 2、针对决策树的构造和修剪通常不能同时进行所产生的效率低下的问题,提出了基于粗糙集的决策树构造方法。利用优先策略,将知识相依性同时作为属性约简和建树的准则,在决策树预修剪的同时进行节点生成,大大提高了决策树构造的效率。
- 3、针对单变量决策树忽视信息系统中广泛存在的属性间的关联作用,而且修剪时往往代价很大的缺陷,提出了一种基于主成分分析的多变量决策树构造方法,提取信息系统中的若干主成分来构造决策树。
- 4、探讨了用Boosting方法组合多决策树,构造决策森林的方法。
- 5、在不完备信息系统中的模型拓展。提出了一种加权联系度容差关系,在各属性重要性排序的前提下对不完备信息系统进行进一步的粗糙集模型拓展,使其更加符合人的主观要求和客观现实。从而为进一步探讨在不完备信息系统中构造分类器模型打下基础。

2. 期刊论文 [田苗苗 数据挖掘之决策树方法概述](#) -[长春大学学报](#)2004, 14(6)

数据挖掘在科研和商业应用中正发挥着越来越重要的作用。分类器是数据挖掘的一种基本方法,决策树是一种最重要的分类器。本文介绍了分类器中的决策树方法及其优点,决策树表示法,决策树构造思想,并比较了各种重要的决策树算法。介绍了决策树算法的实现工具,决策树与数据仓库的结合,决策树的适用范围及应用,最后探讨了决策树的发展趋势。

3. 学位论文 [程向前 基于决策树的数据挖掘算法和可视化研究](#) 2007

数据挖掘是一种可以从海量数据中智能地和自动地抽取一些有用的、可信的、有效的、可以理解的模式的过程,也被称之为数据库中的知识发现。分类是数据挖掘的一种非常重要的方法。分类的概念是在已有数据的基础上学习一个分类函数或构造出一个分类模型(即分类器)进行类型的划分。该函数或模型能够把数据库中的数据记录映像到给定类别中的某一个。分类方法应用领域广泛,如金融市场走向分析、顾客信用度分析、医疗诊断等。

决策树是数据挖掘中一种应用最为广泛的分类器。其原因主要有:(1)决策树分类的直观表示方法较容易转化为标准的数据库查询;(2)决策树分类归纳的方法行之有效、尤其适合于大型数据集;(3)决策树在分类过程中,除了数据集中已经包括的信息外,不再需要其他额外的信息;(4)决策树分类模型的预测准确度较高。由于决策树本身具有建树思想简单、易于提取规则、贴近人类思维、便于理解等优点,使其在分类数据挖掘中得到了广泛应用。决策树算法的研究可以扩大算法的应用范围,提高算法的运行效率以及分类的准确率。本文从属性离散化、降维、属性选择标准、剪枝、与其它数据挖掘方法的结合等几个方面对目前决策树在分类数据挖掘中的研究状况进行了阐述。

本文在介绍了一些典型的决策树分类算法的基础上,重点描述了一种基于决策树的数据挖掘新算法,即基于属性相似度的决策树分类器的研究成果。不同测试属性在决策中的地位也不相同,部分测试属性甚至对决策不起任何作用,完全可进行约简。实验也证明数据集中无关的、干扰的属性会影响分类器的性能,导致性能变差。因而本文首先进行了属性选择,只保留与决策最为相关的属性,而将其他属性都去除。然后通过计算测试属性与决策属性的相似度作为启发规则来构造决策树。算法还使用了分类阈值设定方法简化决策树的生成过程。新算法在对UCI实验数据库中的四个数据集的实验中,运行效率明显高于ID3算法,预测精度在某些数据集中也优于ID3。

Weka数据挖掘平台是新西兰怀卡托大学开发的基于Java语言的开源的数据挖掘平台。它提供了一个Java库形式的框架,这个框架支持嵌入式及其学习的应用,以及新的学习方案的实现。本文在熟悉其API的基础上,成功地在此平台上实现了自己的新的算法。数据挖掘结果的可视化可以使用户和决策者非常形象和直观地分析得到的知识,本文在Weka平台上将新算法模型得到的决策树成功地以图形的方式展示。

4. 期刊论文 [朱娟, 杨丰华 改进的决策树算法在教务管理数据挖掘系统中的应用](#) -[软件导刊](#)2010, 09(4)

决策树算法是数据挖掘系统中一个重要的分类算法,选择合理而有效的测试属性以及对决策树进行适当的修剪是决策树算法的关键内容之一。将决策树算法引入教务管理挖掘系统,并对决策树测试属性的选择算法以及预剪枝算法进行改进,以九江学院学生四级考试信息为例,结果表明改进的决策树算法对于数据挖掘更具可靠性和有效性。

5. 学位论文 [宋广玲 基于多关系决策树算法的研究](#) 2009

多关系数据挖掘是近年来快速发展的重要的数据挖掘领域之一。高效性和可扩展性一直是数据挖掘领域的重要研究课题。考虑多关系数据挖掘,这个问题尤为重要。多关系数据挖掘任务的复杂性对算法的性能提出了更高的要求。与传统的数据挖掘算法相比,多关系数据挖掘算法的搜索空间变得更复杂,更大。对于多关系数据学习算法,提高算法效率的主要瓶颈在于假设空间。针对以上问题,本文主要做了以下工作:

首先, 本文对数据挖掘理论、关系数据挖掘理论进行了研究, 尤其是多关系数据挖掘的分类算法-多关系决策树算法及多关系数据挖掘的最新技术-元组传播技术进行了深入的研究。

其次, 本文提出了多关系决策树的改进算法。多关系决策树主要从两方面进行改进: 1为了多关系决策树算法可扩展性, 本文将虚拟连接元组传播技术应用到改进的多关系决策树算法中; 2为了减少系统独自摸索的时间、减少系统搜索有用属性的时间和提高用户的满意程度, 本文提出了在用户指导下完成分类任务的背景属性传递技术, 并将该技术应用到改进的多关系决策树中。

最后, 本文对改进的多关系决策树算法进行了理论证明和实验验证。本文的实验主要利用了PKDD CUP' 99中的Loan、Account、Transaction三个关系, 采用两种方法对一般多关系决策树算法和改进的对关系决策树算法进行比较实验。第一种方法, 固定三个关系的记录数不变, 每个关系分别增加属性个数进行实验, 第二种方法, 固定三个关系中的属性个数不变, 改变关系记录条数进行实验。

通过上面的实验结果, 本文研究认为, 当改进的多关系决策树在搜索数据项未达到背景属性传递阈值时, 改进多关系决策树算法的运行效率较低; 当改进的多关系决策树在搜索数据项达到背景属性传递阈值时, 改进的多关系决策树算法的效率相对很高且受属性个数增加(或记录数增加)影响较小。

6. 期刊论文 [邹媛, ZOU Yuan 基于决策树的数据挖掘算法的应用与研究 -科学技术与工程2010, 10\(18\)](#)

数据挖掘是指从数据库中抽取隐含的、具有潜在使用价值信息的过程, 是一种新型的数据分析技术。研究数据挖掘中的决策树算法以及决策树算法在具体的客户关系管理系统中的研究与分析, 对数据挖掘中的决策树技术做了详细的描述。

7. 学位论文 [但小容 数据挖掘中决策树分类算法的研究与改进 2008](#)

数据库技术的迅速发展以及数据库管理系统的广泛应用, 导致人们积累了越来越多的数据。巨增的数据背后蕴藏着丰富的知识, 而目前的数据库技术虽可以高效的实现数据的查询、统计等功能, 却无法发现数据中存在的关系和规则, 无法根据现有的数据预测未来的发展趋势。数据库中存在着大量的数据, 却缺乏挖掘数据背后隐藏的知识的手段, 出现了“数据爆炸而知识贫乏”的现象。

在此背景下, 数据库知识发现(KDD)及其核心技术——数据挖掘(DM)便应运而生。数据挖掘(Data Mining)是信息处理技术研究领域的一项重要课题。数据挖掘是利用分析工具从大量的、不完整的、有噪声的、模糊的、随机的数据中, 提取出隐含在其中、事先未知、潜在有用的信息和知识的过程, 建立数据间关系模型, 用其做出预测, 从而为决策者提供辅助。它是一种新型的数据分析技术, 已被广泛应用于金融、保险、政府、教育、运输以及国防等领域。

数据分类是数据挖掘中一个重要的内容。常用的分类模型有决策树、神经网络、遗传算法、粗糙集、统计模型等。决策树是分类应用中采用最广泛的模型之一。与神经网络和贝叶斯方法相比, 决策树无须花费大量的时间和进行上千次的迭代来训练模型, 适用于大规模数据集, 除了训练数据中的信息外不再需要其他额外信息, 表现了很好的分类精确度。并且决策树算法是以实例为基础的归纳学习算法, 以其易于提取显示规则、计算量相对较小、可以显示重要的决策属性和较高的分类准确率等优点而得到广泛的应用。据统计, 目前决策树算法是利用最广泛的数据挖掘算法之一。然而在实际应用过程中, 现存的决策树算法也存在着很多不足之处, 如计算效率低下、多值偏向等。因此, 进一步改进决策树, 提高决策树的性能, 使其更加适合数据挖掘技术的应用要求具有重要的理论和实际意义。

本文主要介绍如何利用决策树方法对数据进行分类挖掘。文中详细的阐述了决策树的基本知识和相关算法, 并对几种典型的决策树算法进行了分析比较, 如: 核心经典算法-ID3算法; 能够处理不完整的数据、对连续属性的数据离散化的C4. 5算法; 利用GINI系数判别数据集分裂属性并形成二叉树的CART算法; 使数据的分类不受机器主存的限制, 有着良好的伸缩和并行性的SLIQ和SPRINT算法。文中分析并比较了它们各自的优缺点。在决策树算法中属Quinlan于1986年提出的ID3算法最有名, 它是非递增算法, 并且采用信息熵作为属性选择的标准, 可是这个标准易偏向于属性值数较多的属性, 而属性值较多的属性却不总是最优的属性。为了解决取值偏向的问题, 本文提出了一种基于ID3算法的加权简化信息熵算法, 该算法的思想是首先将泰勒公式的原理与ID3算法的属性选择标准——信息熵的求解相结合, 对ID3算法信息熵的求解进行简化, 改变了决策树算法中属性选择的标准, 减小了算法的计算复杂度, 提高了算法的运行效率; 然后再赋予每个属性的信息简化熵一个权值, N的取值取决于每个属性的取值个数, 用以平衡每个属性对数据集的不确定程度, 使得属性的选择更加合理化, 避免选择的属性与实际不相符。最后在Visual studio6. 0平台上利用C++语言分别实现改进前后的ID3算法。实验结果表明, 改进后的加权简化信息熵算法提高了决策树的构建速度, 减少了算法的计算运行时间, 同时也克服了ID3算法往往偏向于选择取值较多的属性作为测试属性的缺陷。并且随着数据规模的增大, 决策树的分类性能表现得越好。理论分析和实验结果表明, 本文提出的改进算法改善了决策树的ID3算法的性能, 表现出了良好的分类效果。

8. 学位论文 [卢东标 基于决策树的数据挖掘算法研究与应用 2008](#)

数据挖掘是指从数据库中抽取隐含的、具有潜在使用价值信息的过程, 是一种新型的数据分析技术, 已经被广泛应用于金融、保险、政府、教育、运输以及国防等领域。

数据分类是数据挖掘中一个重要的内容。分类存在很多方法, 常见的分类模型有决策树、神经网络、遗传算法、粗糙集、统计模型等。其中决策树算法是以实例为基础的归纳学习算法, 以其易于提取显示规则、计算量相对较小、可以显示重要决策属性和较高的分类准确率等优点而得到广泛的应用。据统计, 目前决策树算法是利用最广泛的数据挖掘算法之一。

然而在实际的应用过程中, 现存的决策树算法也存在很多不足之处, 如计算效率低下、多值偏向等。因此, 进一步改进决策树, 提高决策树的性能, 使其更加适合数据挖掘技术的应用要求具有重要的理论和现实意义。

本文针对上述数据库知识发现的不足, 进行深入的研究, 探索数据挖掘中决策树分类的优化算法, 以便更好地提高分类的准确性, 更好地应用于实际工作中。本文主要的研究工作如下:

第一, 从宏观上介绍了数据挖掘和分类技术的理论基础, 并重点对几种常见决策树算法进行了分析和比较, 例如ID3、C4. 5、CART算法。

第二, 详细地分析了利用决策树方法对数据进行分类挖掘时常见的几个问题: 属性值空缺、连续属性的处理、过度拟合数据等。这些问题都会导致决策树的分类精度下降, 因此在构建决策树时必须选择合理的策略, 提高决策树的分类精度。

第三, 本文对决策树算法进行了优化研究, 对属性值空缺、属性选择标准多值化、属性选择标准等问题提出了具体的解决办法。本文还提出了加权简化熵的概念, 并对ID3算法进行了改进, 经过比较, 改进算法在总体性能上优于目前广泛应用的ID3算法。

第四, 利用新的决策树算法在一个棉纺厂的设备管理系统中进行数据挖掘, 为厂家的决策支持提供了科学、准确的根据。

9. 期刊论文 [李琳, 陈德钊, 束志恒, 叶子青, Li Lin, Chen Dezhaoh, Shu Zhiheng, Ye Ziqing 基于预处理的决策树在化学数据挖掘中的应用 -分析化学2005, 33\(8\)](#)

化学数据挖掘可从海量数据中提取蕴含的知识, 决策树方法是一种重要的挖掘工具。鉴于决策树在处理连续数据上的局限性, 本研究提出先进行预处理, 将连续属性离散化, 通过特征选择删除其冗余量, 以此为基础构建决策树。该方法可防止决策树模型“过细”, 使之具有良好的预报性能。将此方法应用于两个化学样品分类实例, 效果良好。与贝叶斯分析和单一的决策树方法相比, 其预报正确率有显著提高, 且表达形式直观明确, 易于理解和分析, 适用于化学分类知识模式的挖掘。

10. 学位论文 [刘振宇 数据挖掘算法分析及其在铁路员工培训系统中的应用 2006](#)

数据库知识发现(KnowledgeDiscoveryinDatabase, KDD)是从大量数据中发现潜在规律、提取有用知识的方法和技术。近年来, KDD受到了国内外普遍关注, 已经成为信息系统和计算机科学领域研究中最活跃的部分。KDD被认为是从数据中发现有用知识的整个过程, 而数

据挖掘(DataMining, DM)被认为是KDD过程中的一个特定步骤,它用专门算法从数据中抽取模式。

数据挖掘作为一种高效、深层次的数据分析处理技术,其目的在于从大量的数据中提取出隐含在其中的潜在信息,这些信息将为人们进行各种决策分析提供有力依据。如何利用数据挖掘技术对现有的大量数据进行分析处理,具有重要的实际应用价值。目前数据挖掘的研究主要集中在如何完成各种知识发现任务,如分类知识发现、聚类知识发现、关联规则发现等。研究的重点在具体的数据挖掘算法,算法研究的目的在于提高挖掘的效率及挖掘结果的实用性。

本文以实现铁路员工培训系统中培训资源和培训模式选择的优化为目标。首先在初步调研与分析知识发现与数据挖掘相关理论与应用的基础上,归纳了该领域的主要研究内容和关键技术。进而结合数据挖掘的应用现状和理论基础,重点分析了分类、聚类算法的理论、方法和实现技术。研究的主要内容有数据挖掘的过程模型、数据预处理、决策树分类和聚类的常用算法等。然后介绍了目前铁路员工培训资源与培训模式的现状及现有铁路员工培训系统的作用和意义。并着重分析了系统中存在的问题,在培训资源与培训模式方面提出了改进方案。最后利用聚类与分类算法对培训资源与培训模式进行优化,并对所搜集的现有培训资源与培训模式进行了聚类和分类挖掘,分析了已有数据的规律,期望对未知类别的数据进行预测。本文所提出的培训资源与培训模式优化选择方案对铁路员工培训具有一定的指导及帮助作用。

本文主要研究工作如下:1、介绍数据挖掘算法中基本分类算法—决策树分类算法,进行了系统的总结,给出了决策树算法的处理流程以及决策树生成过程,对经典的决策树算法进行了比较,分析了各自的优缺点。

2、针对经典决策树与人的思维及感知认识上的不相符,对连续属性处理的缺陷,引入模糊决策树算法,深入研究了模糊决策树算法的实现策略,在此基础上提出了一种新的模糊决策树算法—模糊基尼系数法。

3、对聚类算法中的经典K均值法进行描述,指出该算法的不足之处,提出了一种改进的K均值算法,并对二者的性能进行了比较,证明了改进后的K均值算法优于经典K均值算法。

4、基于本文所阐述的决策树算法和聚类算法,设计了一个关于铁路员工培训资源与培训模式的优化选择方案,对培训资源与培训模式进行分析与预测,可以提高员工培训质量。

本文针对上述研究内容,进行了大量的实验研究和论证。结果表明,本文的理论、方法与技术基本正确有效,所涉及的铁路员工培训系统培训资源与培训模式优化方案对实际员工培训可提供一定的指导作用,具有良好的实际应用前景。

引证文献(82条)

1. [梁柱森 改进的ID3算法构造应聘结果决策树](#)[期刊论文]-[现代计算机（专业版）](#) 2010(14)
2. [王苗,柴瑞敏 一种改进的决策树分类属性选择方法](#)[期刊论文]-[计算机工程与应用](#) 2010(8)
3. [王莉 ID3算法的研究与应用](#)[期刊论文]-[福建电脑](#) 2010(1)
4. [熊蜀峰,聂黎明 基于C5.0算法的学生成绩分析决策树构造](#)[期刊论文]-[科技信息](#) 2010(8)
5. [胡国华,赵青杉 ID3算法的改进和优化](#)[期刊论文]-[福建电脑](#) 2010(7)
6. [韩煜 数据挖掘技术在医院信息系统中的应用](#)[期刊论文]-[医学信息学杂志](#) 2010(10)
7. [郑凯明,李义杰,姜鑫 基于决策树的超市客户数据挖掘与仿真](#)[期刊论文]-[世界科技研究与发展](#) 2009(2)
8. [梁柱森 数据挖掘技术在高校人力资源管理中的应用现状及展望](#)[期刊论文]-[现代计算机\(专业版\)](#) 2009(2)
9. [戴秋霞 基于客户感知的TSPA网络质量提升模型](#)[期刊论文]-[科技创新导报](#) 2009(8)
10. [魏焕新 浅谈数据挖掘技术及数据挖掘方法](#)[期刊论文]-[科技信息](#) 2009(32)
11. [刘莺迎 决策树分类算法的分析和比较](#)[期刊论文]-[科技情报开发与经济](#) 2008(2)
12. [李萍,段富 基于OLAP和决策树分析结合的教学评估系统](#)[期刊论文]-[沈阳师范大学学报（自然科学版）](#) 2008(3)
13. [杨贵明 分类数据挖掘算法在开发区经济决策方面的应用](#)[期刊论文]-[科技创新导报](#) 2008(11)
14. [王海珍,刘艳梅 浅析数据挖掘技术](#)[期刊论文]-[电脑知识与技术](#) 2008(36)
15. [韦小铃 浅谈数据挖掘技术](#)[期刊论文]-[电脑知识与技术](#) 2008(34)
16. [安颖 基于Apriori算法的兴趣集加权关联规则挖掘](#)[期刊论文]-[北京联合大学学报\(自然科学版\)](#) 2008(4)
17. [严铁 基于Web日志挖掘应用研究](#)[期刊论文]-[科技信息（学术版）](#) 2008(10)
18. [湛宁,徐杰 决策树算法的改进](#)[期刊论文]-[电脑知识与技术](#) 2008(15)
19. [冯明军,郭剑毅,赵蕴智 IID3算法在CRM数据分类中的应用研究](#)[期刊论文]-[计算机与数字工程](#)

2007(1)

20. 臧志彭, 来鹏, 张俊琴 基于决策树的科技人员薪酬满意度影响因素实证研究——以江苏省南通市为例 [期刊论文] - 中国科技论坛 2007(10)

21. 徐澜 数据挖掘在成人高校管理中的应用 [期刊论文] - 福建电脑 2007(7)

22. 王明哲 基于数据挖掘技术的信用卡客户的信用评价 [期刊论文] - 商场现代化 2007(22)

23. 戴泳 知识发现与知识挖掘技术及其应用 [期刊论文] - 科技情报开发与经济 2007(26)

24. 鲁为, 王枫 决策树算法的优化与比较 [期刊论文] - 计算机工程 2007(16)

25. 陈黎力 基于数据挖掘的电信客户流失模型分析与设计 [学位论文] 硕士 2007

26. 龙际珍, 任海叶, 易华容 一种改进决策树算法的探讨 [期刊论文] - 株洲师范高等专科学校学报

2006(2)

27. 张明霞 数据挖掘技术及其应用 [期刊论文] - 苏盐科技 2006(3)

28. 曹宇, 刘晓君 数据挖掘在智能化企业竞争情报系统的应用研究 [期刊论文] - 情报杂志 2006(3)

29. 张滢, 张新卫 电路实验数据分析挖掘技术研究 [期刊论文] - 现代电子技术 2006(24)

30. 胡勇, 胡玲 基于C4.5算法在水利水电建筑工程专业成绩分析中的应用 [期刊论文] - 高等建筑教育

2006(4)

31. 王兵 基于多策略的学生成绩挖掘与分析系统的研究与实现 [学位论文] 硕士 2006

32. 林亚丽 数据挖掘技术在纳税评估系统中的研究与应用 [学位论文] 硕士 2006

33. 周刚 数据挖掘中决策树算法在客户流失中的应用研究 [学位论文] 硕士 2006

34. 由军平 基于粗糙集理论的决策树剪枝 [学位论文] 硕士 2006

35. 陈雪峰 恶性血液病数据库分析系统的建立 [学位论文] 硕士 2006

36. 李守华 中文信息过滤技术的研究 [学位论文] 硕士 2006

37. 姜安琦 基于数据挖掘的数据可视化系统的设计与实现 [学位论文] 硕士 2006

38. 李文硕 信息挖掘技术在教务系统等级成绩评估中的应用研究 [学位论文] 硕士 2006

39. 任丽君 数据挖掘在大学生心理问题中的应用研究 [学位论文] 硕士 2006

40. 彭玉楼, 刘亚辉 利用决策树和聚类理论对XML文档数据挖掘的研究 [期刊论文] - 株洲工学院学报

2005(4)

41. 唐海兵, 秦怀青 利用决策树改进基于特征的入侵检测系统 [期刊论文] - 微机发展 2005(4)

42. 数据挖掘在消费者生活形态细分中的应用研究 [期刊论文] - 市场研究 2005(10)

43. 刘慧巍, 张雷, 翟军昌 数据挖掘中决策树算法的研究及其改进 [期刊论文] - 辽宁师专学报(自然科学版) 2005(4)

44. 张群峰, 王静红, 李笔 基于属性约简的决策表算法 [期刊论文] - 河北省科学院学报 2005(3)

45. 熊家军, 张丽, 李庆华 入侵检测中数据挖掘模式的编码与检测研究 [期刊论文] - 计算机应用与软件

2005(11)

46. 姚岳 基于数据挖掘实现中国电信精确化营销的探讨 [学位论文] 硕士 2005

47. 骆庆 基于数据挖掘的信用卡客户细分应用研究 [学位论文] 硕士 2005

48. [陈飞](#) [神经网络和决策树进行数据分类的对比研究](#)[学位论文]硕士 2005
49. [车德文](#) [数据挖掘在政府数据中心中的应用](#)[学位论文]硕士 2005
50. [张允](#) [数据仓库和数据挖掘技术在钻井生产信息中的研究与应用](#)[学位论文]硕士 2005
51. [徐金哲](#) [入侵检测系统中过滤器的设计与实现](#)[学位论文]硕士 2005
52. [王峰](#) [基于关联规则数据挖掘的研究与应用](#)[学位论文]硕士 2005
53. [水静](#) [数据挖掘技术及其在电信CRM中的应用研究](#)[学位论文]硕士 2005
54. [徐春荣](#) [决策树分类在交通数据分析系统中的应用研究](#)[学位论文]硕士 2005
55. [丁元明](#) [数据挖掘技术在高校教学质量评估中的应用研究](#)[学位论文]硕士 2005
56. [赵基](#) [基于数据挖掘的银行客户分析管理关键技术研究](#)[学位论文]博士 2005
57. [迟庆云](#) [基于决策树的分类算法研究和应用](#)[学位论文]硕士 2005
58. [屈娅玲](#) [基于数据挖掘一对一营销分类系统设计与实现](#)[学位论文]硕士 2005
59. [屈娅玲](#) [基于数据挖掘一对一营销分类系统设计与实现](#)[学位论文]硕士 2005
60. [梅尼亚](#) [数据挖掘中的高速可伸缩分类算法](#)[学位论文]硕士 2005
61. [李文](#) [公安执法监督管理中的文本理解技术的研究及其应用](#)[学位论文]硕士 2005
62. [谷宏群](#) [数据挖掘中可视化方法研究](#)[学位论文]硕士 2005
63. [胡保坤](#) [石化安装工程预算智能系统的设计与开发](#)[学位论文]硕士 2005
64. [李昕](#) [智能授导系统中学习者特征分析的研究](#)[学位论文]硕士 2005
65. [金琼](#) [电信客户忠诚度的分析与预测](#)[学位论文]硕士 2005
66. [张毅](#) [数据挖掘技术在ERP内部数据仓库中的应用](#)[期刊论文]-[浙江万里学院学报](#) 2004(2)
67. [马秀红](#), [宋建社](#), [董晟飞](#) [数据挖掘中决策树的探讨](#)[期刊论文]-[计算机工程与应用](#) 2004(1)
68. [郭秀娟](#) [数据挖掘方法综述](#)[期刊论文]-[吉林建筑工程学院学报](#) 2004(1)
69. [陈晴光](#) [决策树在汽车ERP系统中的应用探索](#)[期刊论文]-[桂林电子工业学院学报](#) 2004(1)
70. [付成宏](#), [傅明](#), [肖如良](#), [唐贤瑛](#) [基于决策树的快速入侵检测方法](#)[期刊论文]-[长沙电力学院学报\(自然科学版\)](#) 2004(1)
71. [姚晔](#), [李翔](#) [决策树算法的教育应用探讨](#)[期刊论文]-[江西师范大学学报\(自然科学版\)](#) 2004(4)
72. [林志雄](#), [欧伟强](#), [郑见纯](#) [数据挖掘技术在电信营销的应用](#)[期刊论文]-[广东通信技术](#) 2004(12)
73. [郭秀娟](#) [基于关联规则数据挖掘算法的研究](#)[学位论文]博士 2004
74. [齐金鹏](#) [数据挖掘模型可视化研究及其应用实例](#)[学位论文]硕士 2004
75. [王文利](#) [基于数据挖掘的金融时间序列的小波理论应用](#)[学位论文]硕士 2004
76. [杨晓松](#) [应用数据挖掘技术构建云南移动客户离网预测模型](#)[学位论文]硕士 2004
77. [张艳丽](#) [数据挖掘技术在数字化校园的教务系统中的应用](#)[学位论文]硕士 2004
78. [饶丹](#) [带否定关联规则挖掘算法的研究](#)[学位论文]硕士 2004
79. [熊家军](#) [基于数据挖掘的入侵检测关键技术研究](#)[学位论文]博士 2004
80. [蔡凌卿](#) [银行卡客户细分系统分析与设计](#)[学位论文]硕士 2004
81. [李利](#) [基于数据库的数据挖掘研究](#)[学位论文]硕士 2003

82. 路应金. 徐谟. 周宗放 应用数据挖掘技术分析技术MBA培养模式 [期刊论文] - 电子科技大学学报(社会科学版) 2002 (3)

本文链接: http://d.wanfangdata.com.cn/Periodical_jsjyyyj200108006.aspx

授权使用: 温州大学图书馆(wzdxstg), 授权号: 62c917e4-d78a-4bce-b0ff-9efc016ec217

下载时间: 2011年6月8日