教育教学研究

常用的决策树生成算法分析

于 莉

天津财贸管理干部学院，天津 300170

[摘要] 数据分类是数据挖掘中的一个重要课题，被有效地应用于科学实验、医疗诊断、气象预报、商业预测等领域。常用的分类方法包括决策树、神经网络、遗传算法等。其中决策树是分类方法中的一个重要研究方向，由于其结构简单、可以清晰的生成便于人们理解的规则、效率高、以及适用大数据量等优点而被广泛使用。本文就几种常用的决策树生成算法进行较深入地分析和比较。

[关键词] 数据挖掘；决策树；决策树算法

[中图分类号] TP301.6 [文献标识码] B [文章编号] 1008-9055(2008)02-0019-02

Analyses on Establishing Calculate Ways of Decision Trees in Common Use

YU Li

(Tianjin Commercial and Financial Management Institute, Tianjin 300170)

[Abstract] Data classification is an important topic in digital excavation, and being used for science tests, medical treatments, weather forecasts, and business predictions. The classification method in common use includes decision tree, neural net, the genetic calculation etc. Among them decision tree is a method of much more importance in the research direction. It is in a simple structure, easy to be established and understood, with high efficiency, and suitable to a great deal of data etc. It is used extensively due to all these advantages. This article carries on analyzing and comparing more and thoroughly for a few decision trees' establishing calculate ways in common use.

[Key Words] digital excavation; decision tree; calculate ways of decision trees

一、数据挖掘及分类

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程，其任务是从大量的数据中发现有价值的模式。

分类是数据挖掘中应用极其广泛的重要技术之一，其目的是分析输入数据，通过数据表现的特性构造一个分类模型，用该模型对类别未知的数据进行分类。分类过程分为训练和测试两个步骤：在训练阶段，分析训练数据，为每个类别产生一个对应的数据集的描述规则；在测试阶段，用上述产生的规则对测试数据进行分类，以此来测试分类的准确性。训练阶段用于产生分类模型，是关键步骤。分类模型的构造方法有多种，其中最为典型的是基于决策树的分类方法。

二、决策树分类方法

决策树方法广泛用于解决与分类相关的各种问题，属于有指导的归纳学习算法。该方法着眼于从一组无次序、无规则的数据中归纳出一个分类描述，从中发现潜在的、具有商业价值的信息。

决策树(Decision Tree)是一个类似树结构的表示法，每个内部节点表示一个属性的测试，分支表示一个测试的输出，而叶节点就表示类或类的分布。使用训练样本构建决策树时，通常采用自顶向下的递归方式，即从代表全部训练样本的根节点开始，为每个内部节点选择一个测试属性，并根据该属性的取值将样本划分为若干分支，直到叶节点将样本划分为某一类。在决策树的构建过程中，关键问题是测试属性的选择以及分割点的确定。不同的决策树算法采用的属性分割方法不同，常用的决策树算法主要有ID3、C4.5、CART、SLIQ、SPRINT等。

三、常用的决策树算法分析比较

(一) ID3算法

最早出现的决策树算法是1966年由Hunt等人提出的CLS算法，其主要思想是从一棵空的决策树开始，通过添加结点逐步求精，直到产生一棵能正确分类训练实例的决策树为止。CLS算法在构造决策树的过程中没有给出选择测试属性的具体标准。因此，在该算法的基础上，Quinlan在1979年提出了在国际上最有影响力的以信息熵的下降速度作为选择测试属性标准的ID3算法。

该算法的基本思想是：从代表训练样本的单个节点开始树的构造，如果样本属于同一类，则该节点成为叶节点，并用该类标记；否则采用窗口的采样方法，随机地从训练数据集中选择一个子集，通过计算每个属性的信息增益，选择增益最大且从根节点到当前节点尚未被选择的属性作为节点，并

[收稿日期] 2007-12-08

[作者简介] 于莉(1977—)，女，汉族，天津市人，天津市财贸管理干部学院会统系讲师。研究方向：计算机教学。

根据该属性的不同取值创建不同的分支。直至节点中的所有记录属于同一类或节点中的记录数小于规定的最小记录数。

从ID3算法构造决策树的过程可以看出,ID3算法属于一种自顶向下、分而治之的递归构造决策树的贪心算法。它采用不可返回的策略,每次搜索全部样本空间的一个子集生成决策树,以确保决策树建立最简单,每次分析的训练数据最少。其优点是在测试属性的选择上,利用了信息增益的概念,描述简单,构造的决策树平均深度较小,分类速度快,学习能力强。但其也存在许多缺陷,如不能处理连续属性、可伸缩性差、容易产生过度拟和等。

(二)C4.5算法

C4.5算法是针对ID3算法存在的不足,由Quinlan于1993年提出的决策树主流算法之一。该算法与ID3算法相辅相成,都属于启发式的探索属性空间的贪心算法。在ID3的基础上,C4.5算法对其缺点进行了改进:

1.当每条记录的属性值都不同时,为避免ID3算法倾向于优先选择多值的属性,C4.5用增益比例取代信息增益作为选择分割属性的标准,避免过度拟和的产生。

2.既可以处理离散属性也可以处理连续属性。对于连续属性值通过自动离散化的方式进行处理,即先对连续属性A的值进行递增排序,排序后如果相邻的两个值不属于同一类,则用这两个值的中点m将数据划分为两部分,一部分落入该值范围内即A<=m,另一部分大于该值即A>m。将所有的中点值m作为可能的分割点计算信息增益比,选择最大信息增益比对应的中点值作为分割点划分样本空间。由于C4.5采用深度优先建树,所以为找到对于连续属性而言最佳的分割点,在每一节点处需对数据进行反复的划分。

虽然C4.5算法继承了ID3算法的全部优点,而且由于其思想简单,结果可靠等优点,更加巩固了其主流算法的地位。但其本身也存在决策树性能改善困难、达不到全局最优的结果、评价决策树主要依据错误率以及没有考虑树的深度和结点的个数等不足。

(三)CART算法

CART(Classification and Regression Trees)算法即分类与回归树算法,是由L.Breman等人于1984年提出的生成二叉决策树的算法。它依据GINI系数作为测试属性的选择标准,每次将能够降低数据无序度的预测属性选择出来,按照深度优先的策略构造决策树,属于有指导的学习算法。

该算法可对连续型和离散型属性进行处理。对于离散属性将其所有的属性值看作可能的分割点,求出对应的gini参数,最后选择所有分割点中基尼指数最小者对应的属性作为节点的分割属性。对于连续属性首先将属性值进行排序,可能的分割点为相邻属性值的中点。从所有可能分割点中找出gini指数最小的分割点对应的属性作为分割属性。由于CART算法采用深度优先建树,所以对于连续属性为了找到最佳分割点,需要在节点上对数据反复进行划分。

(四)CHAID算法

CHAID (CHi square Automatic Interaction Detector)算法,即卡方自动交互检测算法,是由Gordon B.Kss博士在1976年提出的快速生成多层决策树的算法。该算法主要对离散型变量进行处理,有时也可以对连续型变量进行处理。但由于选择分割属性的算法不是针对连续型变量设计的,所以对于该类型的输入变量需要先进行离散化的操作。

应用CHAID算法建立决策树时,首先为分类变量的每个取值建立一个分支,如果测试属性存在缺失值,则将缺失值单独分支。然后依据卡方分布的P值,来决定是否进行节点的分裂操作。如果节点中类别的p值小于预先指定的阈值,则节点被分割,直到所有节点的p值均大于阈值,则树的构造结束。

上述四种算法存在的一个共同弱点就是在决策树生成的过程中要求训练集全部或部分一直驻留在内存。所以在数据量急剧增长的情况下,由于数据集不能扩展,致使这些算法不能处理大容量的数据。因此迫切需要具有可伸缩性的算法来解决这一问题。

(五)SLIQ算法

SLIQ(Supervised Learning In Quest)即Quest上的有监督学习,是由IBM Almaden研究中心的Mehta等人在1996年提出的一种快速可扩展的分类算法。该算法在树的构建阶段针对数据量远大于内存容量的情况,利用驻留在磁盘上的属性列表和驻留在内存的类列表两种数据结构,通过采用预排序技术和宽度优先的决策树生长方法,使SLIQ算法能够对驻留在磁盘上的大数据集进行分类,而且在改进学习的时间的同时没有降低精确度。虽然SLIQ算法能以更快的速度生成较小的树,而且不限制训练数据的数量及属性的数量,但由于类表需要一直驻留在内存,当类表不能一次装入内存时,SLIQ算法需要额外进行内外存数据交换,所以处理的数据量仍有限。

(六)SPRINT算法

SPRINT(Scalable Parallelizable Induction of Classification Tree)算法即可扩展的、可并行的归纳决策树算法,是由IBM的J.Shafer于1996年提出的。它完全不受内存的限制,而且处理速度很快,且可扩展。该算法在设计上兼顾了并行处理,允许多个处理器相互合作生成一致的模型。SPRINT算法使用属性列表和类统计矩形表,通过一次排序寻找最佳分割点。由于在SPRINT中将属性列表平均分配到多个处理器上,使得它可以处理大规模的数据集。但随着训练集的增长,它所使用的HASH表也成正比例增长,从而使其运行性能受到较大影响。

通过分析,每种算法各有优势和适用范围。没有一种算法对于所有的数据都适用,也没有一种算法完全优于其他方法。因此需要根据特定问题和特定的数据选择适合的算法。

参考文献:

[1]1.Hunt E B,J Marin,P T Stone.Experiments in Induction.Academic Press. 1966.

[2]Quinlan J R.Induction of decision trees.Machine learning,1986,1.

[3]Quinlan J R.Discovering rules from large collections of examples: A case study.In:Michie D,ed.Expert Systems in the Micro Electronic Age,Edinburgh University Press,1979.

[4]Jiawei Han ,Micheline Kamber.,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.

[5]数据挖掘资料汇编 http://www.dmgroup.org.cn.

[6]陈文伟,黄金才,赵新昱.数据挖掘技术[M].北京:北京工业大学出版社,2002.

[7]邵峰晶,于忠清.数据挖掘原理与算法[M].北京:中国水利水电出版社,2001.

责任编辑:周晓丰

常用的决策树生成算法分析

作者： 于莉
作者单位： 天津市财贸管理干部学院,天津,300170
刊名： 天津市财贸管理干部学院学报
英文刊名： JOURNAL OF TIANJIN INSTITUTE OF FINANCIAL AND COMMERCIAL MANAGEMENT
年，卷(期)： 2008，10(2)
引用次数： 1次

参考文献(7条)

1. 1.Hunt E B,J Matin,P T Stone.Experiments in Induction.Academic Press.1966.

2. Quinlan J R.Induction of decision trees.Machine learning.1986.1.

3. Quinlan J R.Discovering rules from large collections of examples:A case study.In:Michie D,ed.Expert Systems in the Micro Electronic Age,Edinburgh University Press,1979.

4. Jiawei Han,Micheline Kamber.,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.

5. 数据挖掘资料汇编http://www.dmgroup.org.cn.

6. 陈文伟,黄金才,赵新昱.数据挖掘技术[M].北京:北京工业大学出版社.2002.

7. 邵峰晶,于忠清.数据挖掘原理与算法[M].北京:中国水利水电出版社.2001.

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. 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类库形式的框架，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。

3. 学位论文 刘振宇 数据挖掘算法研究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。

4.学位论文 胡小刚 数据挖掘中决策树分类算法的研究 2002

数据挖掘,也称之为数据库中知识发现是一个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、有效的和可以理解的模式的过程.分类是数据挖掘的重要内容之一.目前,分类已广泛应用于许多领域,如医疗诊断、天气预测、信用证实、顾客区分、欺诈甄别. 现已有多种分类的方法,其中决策树分类法在海量数据环境中应用最为广泛.其原因如下:1、决策树分类的直观的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,尤其适合大型数据集.3、决策树在分类过程中,除了数据集中已包括的信息外,不再需要额外的信息.4、决策树分类模型的精确度较高. 该文首先研究了评估分类模型的方法.在此基础上着重研究了决策树分类方法,并对决策树算法的可伸缩性问题进行了具体分析,最后给出了基于OLE DB for DM开发决策树分类预测应用程序.

5.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)便应运而生了。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、不完全的、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，以其易于提取显式规则、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系数判别数据集中的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，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。

6.学位论文 周刚 数据挖掘中决策树算法在客户流失中的应用研究 2006

数据挖掘是从大量的数据中抽取出潜在的、不为人知的有用信息、模式和趋势。其目的是提高市场决策能力、检测异常模式、在过去的经验基础上预言未来趋势等等。它致力于数据分析和理解、揭示数据内部蕴藏知识的技术，已成为未来信息技术应用的重要目标之一。经过20多年的发展，数据挖掘产生了许多新概念和方法。特别是最近几年，一些基本概念和方法趋于清晰，它的研究正向着更深入的方向发展。　　分类模式挖掘是数据挖掘中的一种非常重要的方法，可以应用于数据预测，可划为决策树学习、贝叶斯分类、遗传算法和粗糙集等等。决策树学习是以实例为基础的归纳学习算法。它着眼于从一组无次序、无规则的事例中推理出决策树表示形式的分类规则。它采用自顶向下的递归方式，在决策树的内部结点进行属性值的比较并根据不同的属性值判断从该结点向下的分支，在决策树的叶结点得到结论。　　本文主要是研究数据挖掘中的决策树算法以及决策树算法在具体的小灵通流失分析中的研究与分析。首先对数据挖掘技术的产生和发展现状做了概括性的阐述，介绍了数据挖掘的概念、主要内容、模式和主要问题，以及数据挖掘的应用和发展；接着对数据挖掘中的决策树技术做了详细的描述，介绍了决策树中的经典挖掘算法ID3算法，在分析和总结了ID3，C4.5算法的基本性质、性能和特点的基础上，本文作者对经典的C4.5算法进行了一些改进，并分析了改进后的特点和效果。最后，针对电信业小灵通客户流失的问题，通过数据挖掘技术在大量的历史数据中进行挖掘分析，使用SAS等工具，结合ID3、C4.5以及改进后的C4.5算法对样本集进行分析和比对，进行客户细分，挖掘出不同客户群的业务特征，向公司建议，针对流失倾向较高的群体，并结合这些客户对应的客户群特征，采取有针对性的客户挽留策略。在理论知识商业化应用方面，本论文进行了一次有意义的探索和尝试。

7.学位论文 王惠坡 基于决策树的货票数据挖掘系统的研究 2003

随着铁路信息化技术的发展,作为铁路信息系统子系统的货票系统已经积累了丰富的数据.如何以较少的人力和技术成本,合理利用现有的货票信息资源获取有价值的决策信息,成为货运营销和信息技术部门的一个工作重点.数据挖掘技术的迅速发展,为铁路货运营销工作的深入分析奠定了良好的基础,但现有的数据挖掘工具大都基于数据仓库、OLAP Server或数据文件等,无法直接应用于现有的货票系统中.该课题针对目前铁路信息系统不具有数据仓库的现状和应用人员数据库技术有限的特点,紧密结合铁路货运营销分析问题,采用数据挖掘技术的决策树归纳方法,研究、设计了一个基于决策树的以OLTP数据库为数据源的数据挖掘系统——HPMiner.基本系统的研究和设计力图集预处理、决策树生成、分类规则提取、统计分析与预测为一体,能直接进行连续属性的动态离散化,该离散化过程基于OLTP数据库,是面向具体的挖掘问题,从而降低了对源数据的要求;另一方面,离散化可直接面向应用领域人员,可由用户指定离散区间个数和设定阈值,从而极大地方便了用户的使用,较好地适应了货票信息系统中数据的复杂性.HPMiner系统基于决策树分类算法ID3和C4.5的基本思想,系统的基本平台是Client/Server结构,前台使用VB.NET语言开发,后台通过ADO.NET连接Oracle或SQL Server数据库,基本系统的设计便于和货票信息系统的集成,界面友好.该系统应用于铁路货运营销分析,解决了保价运输收入分析和货流去向分析等多个具体问题.HPMiner系统的研究将决策树分类技术与现有货票信息系统有机地结合起来,使得应用领域分析人员可以方便地挖掘出所希望的知识,用于指导生产;另一方面也为决策树分类技术的应用研究开辟了新的领域.

8.学位论文 周燕 基于ID3决策树算法的医疗数据挖掘研究 2004

医学领域已成为数据挖掘的一个重要领域.在当前医学中,存在大量的可以使用的历史成功案例数据,这些数据中蕴含着很有实用价值的规则,医生可以利用这些规则对新的病人进行辅助诊断,以提供其工作速度、准确度与可靠性,并增强对问题的理解,或者用来训练没有经验的学生或相关人员.因此,研究适用于医学领域的数据挖掘具有重要的意义.决策树(Decision Tree),就是一棵规则树.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.决策树的学习是以实例为基础的归纳学习算法.它着眼于从一组无次序、无规则的事例中推理出决策树表示形式的分类规则.基于决策树的学习算法的一个最大优点就是它在学习过程中不需要使用者了解太多的背景知识,只要训练例子能够用属性—结论式的方式表达出来,就能使用该算法来学习.在决策树学习算法的各种算法中,最为有影响的是Ouinlan于1979年提出的以信息熵(entropy)取测试属性的标准的ID3算法.ID3算法采取信息熵原理选择测试属性分割样本集,处理具有离散型属性和属性值齐全的样本,常常能生成结构比较好,效率比较高的决策树.本文主要分析设计了基于决策树ID3算法的医疗数据挖掘的方法,重点研究了决策树ID3算法,并建立了一个基于决策树ID3算法的医疗数据挖掘系统原型,将本系统应用于医学数据集上可以获得较好的效果,证实本文所探讨的方法具有一定的应用价值.数据挖掘是一个处于不断发展和完善的多学科研究领域,其理论本身及其在医学领域中的应用还存在很多问题值得探讨.本文的研究工作主要是针对是已经转化好的关系数据库数据,避开了复杂类型数据到关系数据库数据这一转化过程.因此,相关研究工作还需要进一步深入.

9.学位论文 林海 基于基因表达式编程的决策树研究 2006

随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，在大量的数据背后隐藏着许多重要的信息。数据挖掘就是利用分析工具从大量的、不完全的、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、但又潜在有用的信息和知识的过程，建立数据间关系模型，并用其做出预测。近年来，数据挖掘受到了国内外的普遍关注，已经成为信息系统和计算机科学领域研究中最活跃的前沿领域。数据挖掘已广泛应用于生物医学、金融、零售业、电信业等领域，并产生了巨大的效益。　　分类是数据挖掘中的一种非常重要的方法。它是在已有数据的基础上学会一个分类函数或构造出一个分类模型(即通常说的分类器)。该函数或模型能够把数据库中的数据项映射到给定类别中的某一个，从而可以应用于数据预测。目前，分类已广泛应用于许多领域，如医疗诊断、天气预测、信用证实、顾客区分、欺诈甄别。　　现已有多种分类的方法，其中决策树分类法在海量数据环境中应用最为广泛。其原因如下：1、决策树分类的直观的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，尤其适合大型数据集。3、决策树在分类过程中，除了数据集中已包括的信息外，不再需要额外的信息。4、决策树分类模型的精确度较高。　　决策树分类器是一个类似流程图的树型结构，其中树的每个内部结点代表对一个属性(取值)的测试，其分支就代表测试的每个结果，而树的每个叶结点就代表一个类别。决策树很容易用IF-THEN规则进行表达。决策树模型是数据挖掘中最常用的一种方法。它能够直接体现数据的特点，便于理解，具有较好的分类预测能力，并能方便提取决策规则。决策树的生成过程也就是知识发现的过程，决策树模型的复杂度和预测精度决定了决策树的好坏。决策树是根据启发规则生成的，常见的决策树生成算法有基于信息论的ID3、C4．5算法以及基于最小GINI指标的CART、SILQ、PUBLIC方法。
演化计算中最重要的分支是遗传算法。遗传程序设计是遗传算法的一个变体。遗传算法和遗传程序设计这两个算法虽然都遵循自然界优胜劣汰的基本原理，但是它们最初在工程应用领域具有不同的功能：遗传算法主要用于函数优化，而遗传程序设计则主要用于建模。一般而言，这两者都要优于传统的统计学方法。今年来，演化计算以及成功应用于数据挖掘，尤其是分类规则挖掘。演化计算已成为数据挖掘的一种不可或缺的工具。　　基因表达式编程是C.Ferreira发明的一种新的遗传算法。基因表达式编程结合了遗传算法和遗传程序设计的优点，克服了它们的缺点，在数学建模方面取得了非常好的效果。正因为其优点和良好的效果，使得基因表达式编程在并不漫长的时间里引起了演化计算领域的广泛关注甚至争议。本文简要介绍了基因表达式编程的基本技术，分析了其具有较高效率的根本原因在于其编码方式所具有的独特优势。　　本文以基因表达式编程和决策树作为主要对象，研究如何利用先进的基因表达式编程技术来构造决策树，以及这种决策树在实际分类中效果如何。本文在第一章首先介绍了论文的选题及其研究意义、选题的国内外研究现状、主要的研究内容。然后在第二章中概述了数据挖掘和分类技术，内容包括分类的主要方法、分类的比较和评估以及分类技术中存在的若干问题。在第三章中首先介绍了有关决策树的基本概念，然后介绍了基本ID3决策树算法，以及针对决策树算法的有关讨论。第四章概述了当前两种主要的基因表达式编程分类器，以及它们各自的特点。第五章是本文的主要工作，对现有的基因表达式编程分类器的优劣进行分析，提出了一种基于基因表达式编程技术的新的决策树算法，并通过试验结果说明了该方法与现有的基因表达式编程分类器方法相比的优势。在第六章结论中，总结了论文的主要工作和后续工作。

10.学位论文 高明 基于决策树的流数据挖掘分类算法研究 2007

随着信息技术的飞速发展和广泛应用，目前许多组织都拥有非常庞大的数据库，并且数据量仍然以每天数百万条记录的速度快速增长。传统的统计和机器学习算法大多是以数据从静态分布中随机抽取样本为假设前提的，然而当前可得到的用来进行数据挖掘的大型数据库一般都违反这一假设。这些数据的产生经过了数月或者数年的时间，而数据生成过程在这段时间又发生了改变，有时甚至是根本性改变，从而使得传统的统计和机器学习算法不再适用。因此有必要对流数据挖掘算法进行研究。　　本文重点研究了基于决策树的两种流数据挖掘分类算法VFDT(Very FastDecision Tree learner)和CVFDT(Concept-adapting Very Fast Decision Treelearner)。VFDT可以进行实时分析，它对每个样本使用固定的内存和时间来处理，并在此基础上建立决策树。VFDT能够使用现有的硬件设备来合并每秒成千上万的样本数据，并使用Hoeffding边界来保证它的输出结果收敛于传统学习器得到的结果。CVFDT在VFDT的基础上做出了一些调整和改进，它以生成一棵派生树的方式来利用绝大多数的旧数据从而保持决策树的更新，一旦旧的决策树变得不可靠而新的决策树变得更准确的时候，就用新的决策树替换旧的决策树。　　基于上述的算法研究，本文对网络与信息安全领域的入侵检测系统进行了研究，根据通过入侵检测系统数据的特征以及流数据挖掘分类算法的目的，分析将流数据挖掘分类算法应用到入侵检测系统的必要性和可行性，并尝试利用UCI(University of California，Irvine)KDD Archive中用于入侵检测领域的测试数据集进行实证研究，从而为算法开辟了新的应用领域，并且也可以从不同的角度来检验算法的适用性，为下一步的研究工作奠定基础。

引证文献(1条)

1.洪雷 数据挖掘技术在推广教育规划中的应用研究[期刊论文]-企业技术开发（学术版） 2009(6)