第5章

数据挖掘原理

5.1 数据挖掘综述

5.1.1 数据挖掘与知识发现

知识发现(Knowledge Discovery in Database,KDD)被认为是从数据中发现知识的整个过程。数据挖掘被认为是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(Pattern)。

KDD过程定义(Fayyad、Piatetsky-Shapiror和Smyth,1996):KDD是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的高级处理过程。

其中,数据集:事实F(数据库元组)的集合;模式:用语言L表示的表达式E,它所描述的数据是集合F的一个子集F_E,它是F_E的精炼表达,我们称E为模式;有效、新颖、潜在有用、可被人理解:表示发现的模式有一定的可信度,应该是新的,将来有实用价值,能被用户所理解。

KDD过程图如图5.1所示。

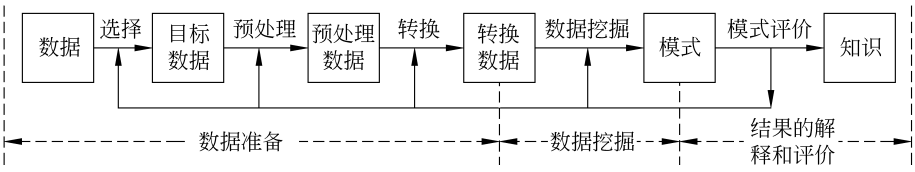

图5.1 KDD过程图

KDD过程可以概括为3部分:数据准备(Data Preparation)、数据挖掘(Data Mining)及结果的解释和评价(Interpretation & Evaluation)。

1. 数据准备

数据准备又可分为3个子步骤:数据选择(Data Selection)、数据预处理(Data Preprocessing)和数据转换(Data Transformation)。

数据选择的目的是确定发现任务的操作对象,即目标数据(Target Data),是根据用户的需要从原始数据库中选取的一组数据。数据预处理一般包括消除

噪声、推导或计算缺值数据、消除重复记录等。数据转换的主要目的是完成数据类型转换(如把连续型数据转换为离散型数据,以便于符号归纳,或是把离散型数据转换为连续型数据,以便于神经网络计算),尽量消减数据维数或降维(Dimension Reduction),即从初始属性中找出真正有用的属性以减少数据挖掘时要考虑的属性的个数。

2. 数据挖掘

数据挖掘是利用一系列方法或算法从数据中获取知识。按照数据挖掘任务的不同,数据挖掘方法分为聚类、分类、关联规则发现等。聚类方法是在没有类别的数据中,按"距离"的远近聚集成若干类别,典型的方法有 k 均值聚类算法。分类方法是对有类别的数据,找出各类别的描述知识,典型的方法有 ID3、C4.5、IBLE 等分类算法。关联规则发现是对多个数据项重复出现的概率,超过指定的阈值时,建立这些数据项之间的关联规则,典型的方法有 Agrawal 提出的关联规则挖掘方法等。

利用数据挖掘方法获得的知识是对这些数据的高度浓缩。

3. 结果的解释和评价

数据挖掘阶段获取的模式,经过评价,可能存在冗余或无关的模式,这时需要将其剔除;也有可能模式不满足用户要求,这时则需要回退到发现过程的前面阶段,如重新选取数据、采用新的数据变换方法、设定新的参数值,甚至换一种挖掘算法等。另外,KDD 由于最终是面向人类用户的,因此可能要对发现的模式进行可视化,或者把结果转换为用户易懂的另一种表示,如把分类决策树转换为 if…then…规则。

数据挖掘仅仅是整个过程中的一个步骤。数据挖掘质量的好坏有两个影响要素:一是所采用的数据挖掘技术的有效性;二是用于挖掘的数据的质量和数量(数据量的大小)。如果选择了错误的数据或不适当的属性,或对数据进行了不适当的转换,则挖掘的结果是不会好的。

整个挖掘过程是一个不断反馈的过程。例如,用户在挖掘途中发现选择的数据不太好,或使用的挖掘技术产生不了期望的结果。这时,用户需要重复先前的过程,甚至从头重新开始。

可视化技术在数据挖掘的各个阶段都扮演着重要的角色。特别是在数据准备阶段,用户可能要使用散点图、直方图等统计可视化技术来显示有关数据,以期对数据有一个初步的了解,从而为更好地选取数据打下基础。在数据挖掘阶段,用户则要使用与领域问题有关的可视化工具。在表示结果阶段,则可能要用到可视化技术,使发现的知识更易于理解。

5.1.2　数据挖掘任务与分类

1. 数据挖掘任务

数据挖掘任务有 6 项:关联分析、时序模式、聚类、分类、偏差检测、预测。

1) 关联分析

关联分析是从数据库中发现知识的一类重要方法。若两个或多个数据项的取值之间重复出现且概率很高时,它就存在某种关联,可以建立起这些数据项的关联规则。

例如,买面包的顾客有90%的人还买牛奶,这是一条关联规则。若商店中将面包和牛奶放在一起销售,将会提高它们的销量。

在大型数据库中,这种关联规则是很多的,需要进行筛选,一般用支持度和可信度两个阈值来淘汰那些无用的关联规则。

支持度表示该规则所代表的事例(元组)占全部事例(元组)的百分比,如既买面包又买牛奶的顾客占全部顾客的百分比。

可信度表示该规则所代表的事例占满足前提条件事例的百分比,如既买面包又买牛奶的顾客占买面包顾客中的90%,称可信度为90%。

2) 时序模式

通过时间序列搜索出重复发生概率较高的模式。这里强调时间序列的影响。例如,在所有购买了激光打印机的人中,半年后60%的人再购买新硒鼓,40%的人用旧硒鼓装碳粉。

在时序模式中,需要找出在某个最小时间内出现比率一直高于某一最小百分比(阈值)的规则。这些规则会随着形式的变化做出适当的调整。

在时序模式中,一个有重要影响的方法是相似时序。用相似时序的方法,要按时间顺序查看时间事件数据库,从中找出另一个或多个相似的时序事件。例如,在零售市场上找到另一个有相似销售的部门,在股市中找到有相似波动的股票。

3) 聚类

数据库中的数据可以划分为一系列有意义的子集,即类。简单地说,在没有类的数据中,按"距离"的远近聚集成若干类。在同一类别中,个体之间的距离较小,而不同类别上的个体之间的距离偏大。聚类增强了人们对客观现实的认识,即通过聚类建立宏观概念。例如,将鸡、鸭、鹅等都聚类为家禽。

聚类方法包括统计分析方法、机器学习方法、神经网络方法等。

(1) 在统计分析方法中,聚类分析是基于距离的聚类,如欧氏距离、汉明距离等。这种聚类分析方法是一种基于全局比较的聚类,它需要考察所有的个体才能决定类的划分。

(2) 在机器学习方法中,聚类是无导师的学习。在这里距离是根据概念的描述来确定的,故聚类也称为概念聚类,当聚类对象动态增加时,概念聚类则称为概念形成。

(3) 在神经网络方法中,自组织神经网络方法用于聚类,如ART模型、Kohonen模型等,这是一种无监督学习方法。当给定距离阈值后,各样本按阈值进行聚类。

4) 分类

分类是数据挖掘中应用最多的任务。分类是在聚类的基础上,对已确定的类找出该类别的描述知识,它代表了这类数据的整体信息,即该类的内涵描述,一般用规则或决策树模式表示。该模式能把数据库中的各元组映射到给定类别中的某个。

一个类的内涵描述分为特征描述和辨别性描述。

特征描述是对类中对象的共同特征的描述,辨别性描述是对两个或多个类之间的区

别的描述。特征描述允许不同类中具有共同特征，而辨别性描述对不同类不能有相同特征。辨别性描述用得更多。

分类是利用训练样本集（已知数据库元组和类别所组成的样本）通过有关算法而求得的。

建立分类决策树的方法，典型的有 ID3、C4.5、IBLE 等算法。建立分类规则的方法，典型的有 AQ 算法、粗糙集方法等。

目前，分类方法的研究成果较多，判别方法的好坏，可从 3 方面进行：①预测准确度（对非样本数据的判别准确度）；②计算复杂度（方法实现时对时间和空间的复杂度）；③模式的简洁度（在同样效果的情况下，希望决策树小或规则少）。

在数据库中，往往存在噪声数据（错误数据）、缺损值、疏密不均匀等问题。它们对分类算法获取的知识将产生坏的影响。

5）偏差检测

数据库中的数据存在很多异常情况，从数据分析中发现这些异常情况也是很重要的，以便引起人们对它更多的注意。

偏差包括很多有用的知识：

（1）分类中的反常实例；

（2）模式的例外；

（3）观察结果对模型预测的偏差；

（4）量值随时间的变化。

偏差检测的基本方法是寻找观察结果与参照之间的差别。观察常常是某个域值或多个域值的汇总。参照是给定模型的预测、外界提供的标准或另一个观察。

6）预测

预测是利用历史数据找出变化规律，建立模型，并用此模型来预测未来数据的种类、特征等。

典型的方法是回归分析，即利用大量的历史数据，以时间为变量建立线性或非线性回归方程。预测时，只要输入任意的时间值，通过回归方程即可求出该时间的预测值。

近年来发展起来的神经网络方法，如 BP 模型，它实现了非线性样本的学习，能进行非线性函数的判别。

分类也能进行预测，但分类一般用于离散数值；回归预测用于连续数值；神经网络方法预测既可用于连续数值，也可用于离散数值。

2. 数据挖掘分类

数据挖掘涉及多个学科，主要包括数据库、统计学和机器学习三大主要技术。

数据库技术经过 20 世纪 80 年代的大发展，除关系数据库外，又陆续出现面向对象数据库、多媒体数据库、分布式数据库以及 Web 数据库等。数据库的应用由一般查询到模糊查询和智能查询，数据库计算已趋向并行计算。从以上各类数据库中挖掘知识正在兴起并已得到迅速发展。

统计学是一门古老的学科，现已逐渐走向社会。它已成为社会调查、了解民意以及制

定决策的重要手段。

机器学习是人工智能的重要分支。它是在专家系统获取知识出现困难后发展起来的。机器学习的大部分方法和技术已演变为数据挖掘方法和技术。

数据挖掘可按数据库类型、数据挖掘对象、数据挖掘任务、数据挖掘方法和技术等方面进行分类。

1) 按数据库类型分类

数据挖掘主要是在关系数据库中挖掘知识。随数据库类型的不断增加,逐步出现了不同数据库的数据挖掘,现有关系数据挖掘、模糊数据挖掘、历史数据挖掘、空间数据挖掘等多种不同数据库的数据挖掘类型。

2) 按数据挖掘对象分类

数据挖掘除对数据库这个主要对象进行挖掘外,还有文本数据挖掘、多媒体数据挖掘、Web 数据挖掘。由于对象不同,挖掘的方法相差很大,文本、多媒体、Web 数据均是非结构化数据,挖掘的难度将很大。

目前,Web 数据挖掘已逐步引起人们的关注。

3) 按数据挖掘任务分类

数据挖掘的任务有关联分析、时序模式、聚类、分类、偏差检测、预测等。按任务分类有关联规则挖掘、序列模式挖掘、聚类数据挖掘、分类数据挖掘、偏差分析挖掘和预测数据挖掘等类型。

各类数据挖掘由于任务不同,将会采用不同的数据挖掘方法和技术。

4) 按数据挖掘方法和技术分类

数据挖掘方法和技术较多,在 5.2 节中将详细讨论。在此对其分类进行说明。

(1) 归纳学习类。

归纳学习类又分为基于信息论方法挖掘类和基于集合论方法挖掘类。基于信息论方法挖掘类是在数据库中寻找信息量大的属性来建立属性的决策树。基于集合论方法挖掘类是对数据库中各属性的元组集合之间关系(上、下近似关系,覆盖或排斥关系,包含关系等)来建立属性间的规则。各类中又包括多种方法,主要用于分类问题。

(2) 仿生物技术类。

仿生物技术类又分为神经网络方法类和遗传算法类。神经网络方法类是在模拟人脑神经元而建立的 MP 数学模型和 Hebb 学习规则的基础上,提出了一系列的算法模型,用于识别、预测、联想、优化、聚类等实际问题。遗传算法类是模拟生物遗传过程,对选择、交叉、变异过程建立了数学算子,主要用于问题的优化和规则的生成。

(3) 公式发现类。

在科学实验与工程数据库中,用人工智能方法寻找和发现连续属性(变量)之间的关系,建立变量之间的公式,已引起人们的关注,该类中有多种数据挖掘方法,如 BACON 和 FDD 等。

(4) 统计分析类。

统计分析是一门独立的学科,由于能对数据库中数据求出各种不同的统计信息和知识,因此它也构成了数据挖掘中的一大类方法。

(5) 模糊数学类。

模糊数学是反映人们思维的一种方式。将模糊数学应用于数据挖掘各项任务中,形成了模糊数据挖掘类,如模糊聚类、模糊分类、模糊关联规则等。

(6) 可视化技术类。

可视化技术是一种图形显示技术。对数据的分布规律进行可视化显示或对数据挖掘过程进行可视化显示,会明显提高人们对数据挖掘的理解和挖掘效果。该技术已形成了可视化数据挖掘类的多种方法。

本书的内容将按数据挖掘的方法和技术分类的各种方法进行详细和深入的介绍,以便读者学习和使用这些方法和技术,对实际问题完成数据挖掘任务。

5.1.3　不完全数据处理

对不完全数据(Incomplete Data)的处理是知识发现过程中数据预处理的主要内容。在现实领域中,人们所拥有的数据常常是不完全的。在这种情况下,知识发现应该具有处理这种不完全数据并提供相应合理的近似结果的能力。

现实世界的数据库(例如,商业数据库和医院数据库)中的数据很少是完全的:丢失的数据、观察不到的数据、隐藏的数据、录入过程中发生错误的数据等在现实中是经常发生的。在知识发现领域中,对不完全数据的研究比较多的在于丢失的数据。

例如,在对个人调查时,被调查的对象可能会拒绝提供他的收入情况,在一项实验过程中,某些结果可能会因为某些故障而丢失,这些情况都会产生数据丢失。

关于两个变量 X 和 Y 的采样。其中 X 是独立变量,总有观测值;Y 是响应变量,可能涉及丢失值。以 $Y=?$ 代表丢失值,以($X=i$,$Y=?$)代表不完全的记录。由这种简单的两个变量模型,可以推广到更一般的情况,即一个不含丢失值的变量的集合总是影响着可能具有丢失值的另一个变量。这种情况在统计学、机器学习、数据挖掘和知识发现领域里是相当常见的。

丢失数据模式分类取决于 $Y=?$ 的概率是否依赖于 Y 与 X 的状态。如果这一概率依赖于 X 但不依赖于 Y,则认为数据是随机丢失(Missing at Random)的;如果 $Y=?$ 的概率既不依赖于 Y 也不依赖于 X 的状态,则认为数据是完全随机丢失(Missing Completely at Random)的。对于数据随机丢失和数据完全随机丢失两种情况,如果数据挖掘方法都不受影响,那么丢失数据的模式是可以忽略的。但当 $Y=?$ 的概率既依赖于 Y 又依赖于 X 时,则丢失数据的模式就是不可忽略的。

处理丢失数据的方法有以下 5 种。

1. 基于已知数据的方法

忽略掉丢失的数据而只对得到的数据进行挖掘和分析。这种方法最为简单,在数据量不太大且数据是完全随机丢失的情况下可以得到令人满意的结果。但是如果数据不是随机丢失的,这种方法就不是很有效,会导致严重的偏差,这时可以采用删除有丢失数据的属性方法。

2. 基于猜测的方法

首先猜测被丢失的值,从而得到完全的数据,然后再运用标准的统计学和机器学习的方法进行数据挖掘和分析。具体方法如下。

(1) 均值替换法。用含有丢失值属性的已知值的平均值来代替丢失的值。

(2) 概率统计法。先求丢失值的所在属性的各取值的出现概率 $P(v_i^a)$,即表示属性 a 的取值 v_i 出现的概率。丢失值用出现最大概率的值 v 来代替。

(3) 回归猜测。采用回归分析的方法,用未丢失的数据建立回归方程,用所依赖的变量 X 求出该丢失值 Y。

3. 基于模型的方法

对于丢失值构造出一个适当的模型(非回归模型),然后在此模型下采用恰当的方法猜测丢失的值,这是一种较为灵活的方法。

4. 基于贝叶斯理论的方法

利用贝叶斯分类技术和贝叶斯网络处理丢失的数据。

5. 基于决策树的方法

利用决策树和规则归纳的技术来处理丢失的数据。

以上主要讨论了对不完全数据的处理。另外,对未知的数据、隐藏的数据、错误的数据等以及这些数据和已知数据的关系,目前研究较少,还需要深入研究。

5.1.4 数据库的数据浓缩

数据浓缩就是在满足某种等价条件下,将复杂的难以理解的数据库,变换成简洁的、容易理解的高度浓缩的数据库。

数据浓缩包括属性约简和元组(记录)压缩两方面。

1. 属性约简

属性约简一般用于分类问题。属性约简的原则是保持数据库中分类关系不变。目前,属性约简一般采用粗糙集(Rough Set)方法,也可以采用信息论方法。

在数据库(S)的分类问题中,属性分为条件属性(C)和决策属性(D)。属性约简是在条件属性中删除那些不影响对决策属性进行分类的多余的属性。经过研究对条件属性一般分为可省略属性和不可省略属性。不可省略属性(Core(S))实质是对决策属性进行分类的核心属性;而可省略属性(Choice(S))并不是全部都可省略的属性,需要在可省略属性中挑选出部分属性与核心属性组合成等价原数据库的分类效果。

例如,如下汽车数据库(CTR),有 9 个条件属性,1 个决策属性(里程),如表 5.1 所示。

表 5.1　汽车数据库(CTR)

序号	类型 *a*	汽缸 *b*	涡轮式 *c*	燃料 *d*	排气量 *e*	压缩率 *f*	功率 *g*	换挡 *h*	重量 *i*	里程 *D*
1	小型	6	Y	1 型	中	高	高	自动	中	中
2	小型	6	N	1 型	中	中	高	手动	中	中
3	小型	6	N	1 型	中	高	高	手动	中	中
4	小型	4	Y	1 型	中	高	高	手动	轻	高
5	小型	6	N	1 型	中	中	中	手动	中	中
6	小型	6	N	2 型	中	中	中	自动	重	低
7	小型	6	N	1 型	中	中	高	手动	重	低
8	微型	4	N	2 型	小	高	低	手动	轻	高
9	小型	4	N	2 型	小	高	低	手动	中	中
10	小型	4	N	2 型	小	高	中	自动	中	中
11	微型	4	N	1 型	小	高	低	手动	轻	高
12	微型	4	N	1 型	中	中	中	手动	中	高
13	小型	4	N	2 型	中	中	中	手动	中	中
14	微型	4	Y	1 型	小	高	高	手动	中	高
15	微型	4	N	2 型	小	中	低	手动	中	高
16	小型	4	Y	1 型	中	中	高	手动	中	中
17	小型	6	N	1 型	中	中	高	自动	中	中
18	小型	4	N	1 型	中	中	高	自动	中	中
19	微型	4	N	1 型	小	高	中	手动	中	高
20	小型	4	N	1 型	小	高	中	手动	中	高
21	小型	4	N	2 型	小	高	中	手动	中	中

经过分析,可以得到:

Corse(S)={燃料,重量},Choice(S)={类型、汽缸、涡轮式、排气量、压缩率、功率、换挡}

保持数据库(S)分类关系不变的 7 种属性约简。

(1){类型,燃料,排气量,重量}4 个属性。

(2){燃料,排气量,压缩率,重量}4 个属性。

(3){类型,汽缸,燃料,压缩率,重量}5 个属性。

(4){类型,燃料,压缩率,功率,重量}5 个属性。

(5){类型,汽缸,燃料,功率,重量}5 个属性。

(6){汽缸,燃料,压缩率,功率,重量}5 个属性。

(7){类型,汽缸,涡轮式,燃料,换挡,重量}6个属性。

以上7种属性约简都等价于原数据库中9个属性的决策分类。

其中最小属性约简是(1)和(2),用4个属性就可以代替数据库中9个属性。利用最小属性约简(2),经过进一步处理,可以得到原数据库的等价数据库,如表5.2所示。

表5.2 约简后的数据库

	燃料	排气量	压缩率	重量	里程
1'	*	*	*	重	低
2'	*	*	*	轻	高
3'	*	小	中	*	高
4'	*	中	*	中	中
5'	1型	小	高	*	高
6'	2型	*	高	中	中

注:*表示可不考虑该属性的取值。

2. 元组(记录)压缩

元组(记录)压缩实质上是对数据库的元组(记录)进行合并、归并和聚类等。

1) 相同元组(记录)的合并

在进行属性约简后,会出现很多相同的元组(记录),这样就可以合并这些相同的元组(记录)。

2) 利用概念树进行归并

概念树是一种对概念的层次进行划分的树。概念树与数据库中特定的属性有关,它将各个层次的概念按从一般到特殊的顺序排列。在概念树中最一般的概念作为树的根节点;最特殊的概念作为叶节点,它对应数据库具体属性值。例如,反映某数据库中"籍贯"这个属性的概念树,如图5.2所示。

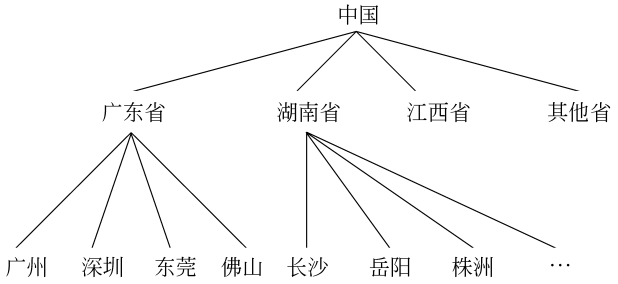

图5.2 "籍贯"概念树

利用概念树进行向上归纳,可以实现数据库元组(记录)归并。例如,对数据库中"籍贯"为广州、深圳、东莞、佛山等城市的所有学生的记录都归并为广东省,即放在"籍贯=广东省"的新记录中,这样就完成了广东省内学生的多个元组(记录)都归并到一个元组(记

录)中,实现了元组(记录)的压缩。对学生数据库这种元组(记录)压缩有利于学校对各省学生的生活习惯有概括的了解,便利了学校对他们的管理。

3) 对元组(记录)的聚类

为了对数据库中所有元组(记录)有一个概括的了解,在元组(记录)之间设定一种距离方法(如汉明距离),对数据库中所有元组(记录)进行聚类。这种聚类能完成对同一类的多个元组(记录)进行聚集,形成一个类元组(记录)。数据库按类元组(记录)重新组织,就完成了原数据库元组(记录)高度压缩的新数据库。

5.2　数据挖掘方法和技术

数据挖掘方法依据的基本原理:①信息论,主要是计算数据库中属性的信息量,如 ID3、IBLE 等算法;②集合论,利用集合之间的覆盖关系(如粗糙集方法、覆盖正例排斥反例的 AQ11 算法),或计算数据项在整个集合中所占的比例(如关联规则挖掘方法);③仿生物技术,把生物体的运转过程转换成数学模型,再用数学模型去解决现实世界的非生物问题,如神经网络、遗传算法等;④人工智能技术,主要是利用启发式搜索方法,如公式发现的 BACOM、FDD 等方法;⑤可视化技术,主要是利用图形显示技术。

数据挖掘方法和技术可以分为五大类:归纳学习的信息论方法、归纳学习的集合论方法、仿生物技术的神经网络方法、仿生物技术的遗传算法和数值数据的公式发现。

5.2.1　归纳学习的信息论方法

归纳学习方法是目前重点研究的方向,研究成果较多。从采用的技术上看,分为两大类:信息论方法(这也是常说的决策树方法)和集合论方法。每类方法又包含多个具体方法。

信息论方法是利用信息论的原理建立决策树。由于该方法最后获得的知识表示形式是决策树,因此一般文献中称它为决策树方法。该类方法的实用效果好,影响较大。

信息论方法中较有特色的方法有以下两种。

1. ID3 等算法(决策树方法)

Quiulan 研制的 ID3 算法是利用信息论中互信息(Quiulan 称其为信息增益)寻找数据库中具有最大信息量的字段,建立决策树的一个节点,再根据字段的不同取值建立树的分支,再由每个分支的数据子集重复建树的下层节点和分支的过程,这样就建立了决策树。这种算法的数据库愈大效果愈好。ID3 算法在国际上影响很大。ID3 算法以后又陆续开发了 ID4、ID5、C4.5 等算法。

2. IBLE 算法(决策规则树方法)

钟鸣、陈文伟研制了 IBLE 算法,是利用信息论中信道容量,寻找数据库中信息量从大到小的多个字段的取值建立决策规则树的一个节点,根据该节点中指定字段取值的权值之和与两个阈值比较,建立左、中、右三个分支,在各分支子集中重复建树节点和分支的