基于一种新的属性选择标准的 ID3 改进算法*

喻金平[1]，黄细妹[1]，李康顺[1,2]
(1. 江西理工大学 信息工程学院，江西 赣州 341000；2. 华南农业大学 信息学院，广州 510642)

摘 要：结合 ID3 算法的不足，提出一种基于属性重要度简化标准的 ID3 改进算法：a)简化 ID3 算法的信息熵从而降低算法的计算时间；b)引入属性重要度概念来弥补 ID3 算法属性选择标准的不足；c)综合 a)和 b)来实现新的属性选择标准即属性重要度简化标准。在开源的 Weka 数据挖掘软件环境下进行仿真实验，结果表明该改进算法是可行的，并且在算法的计算时间和准确度方面都优于 ID3 算法，尤其是在数据样本集规模达到一定数量时，效果更加明显。

关键词：简化；ID3 算法；重要度；数据挖掘软件
中图分类号：TP301.6 **文献标志码**：A **文章编号**：1001-3695(2012)08-2895-04
doi:10.3969/j.issn.1001-3695.2012.08.024

Improved ID3 algorithm based on new attributes selection criterion

YU Jin-ping[1], HUANG Xi-mei[1], LI Kang-shun[1,2]
(1. *School of Information Engineering, Jiangxi University of Science & Technology, Ganzhou Jiangxi* 341000, *China*; 2. *School of Information, South China Agricultural University, Guangzhou* 510642, *China*)

Abstract: Combining the defects of the ID3 algorithms, this paper proposed an improved ID3 algorithm based on importance of attributes simplified criterion. This algorithom a) simplified the information entropy of ID3 algorithm to reduce the computational time, b) introduced the importance of attribute concept to remedy the inadequacy of the attribute selection criterion of ID3 algorithm, c) combined in a) and b) to achieve a new attribute selection criterion which was the important degree of attributes reduction criterion. Simulation experiments were carried out in the open source Weka data mining software. The results show that the improved algorithm is feasible, and the computation time and accuracy are better than the ID3 algorithm, especially in the sample data set to achieve a certain amount, effect is more obvious.

Key words: simplify; ID3 algorithm; important degree; data mining software

随着计算机技术和网络技术的高速发展，信息数据量也以不可估量的倍数在增长，如何有效地从这些数据中找到潜在的有用知识，数据挖掘技术的目的正在于此。数据挖掘是 20 世纪 80 年代末开始逐步发展起来的一个研究领域，它是从大量的、不完全的、有噪声的、模糊的实际应用数据中，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[1,2]。

当前数据挖掘的分类方法有很多种，如决策树算法、遗传算法、贝叶斯网络、粗糙集、K-最近邻方法、关联规则方法等[3]，决策树分类算法是其中一种比较普遍典型的分类方法。早在 1979 年 Quinlan 提出了著名的 ID3 算法，之后很多学者对 ID3 算法进行了大量研究，针对其不足进行了大量改进，其中较为典型的改进算法是 C4.5 算法，该算法能够很好地处理连续属性，其他还有 ID3 算法的增量版本 ID4 等[4]。

本文研究的重点是决策树算法中的 ID3 算法，通过对其缺点的分析，提出了一种改进的算法。ID3 改进算法主要是从降低熵的计算时间和完善属性选择标准两方面进行，在实例验证比较中发现改进的算法能够很好地找到符合实际情况的分类规则。

1 ID3 算法

决策树是一种比较常用的分类方法，它是一种用于预测的分类型建模方法，采用的是“分而治之”的思想，将所有问题的搜索空间分为若干个子类[5]。应用这种方法建立一棵树来对分类过程进行建模，建立好了决策树就可以将其应用到数据集中的元组并得到分类结果。

决策树就是一个类似流程图的树结构，树的节点是样本的属性，属性的取值为树的分支。决策树的产生是根据信息论原理，对大量样本的属性进行分析归纳的结果[6,7]。决策树算法都是利用贪心非回溯的算法，自顶向下地建造决策树。建立决策树需要两个步骤：a)利用训练样本集来建立决策树，并对决策树进行优化；b)对前面建好的决策树进行分析，得到分类规则并对测试样本集进行分类。其中决策树的建立是相当重要的，其过程是：首先寻找初始分裂属性，即将训练集的每个属性进行分裂量化，从中找到一个最好的分裂属性；然后重复上一步，直到每个叶节点的记录都属于同一类，并形成一棵完整的树。常用的决策树算法是 ID3。

收稿日期：2011-12-25；**修回日期**：2012-02-07 **基金项目**：国家自然科学基金资助项目(70971043)；江西省自然科学基金资助项目(2008GZS0028)

作者简介：喻金平(1964-)，男，教授，硕导，硕士，主要研究方向为数据挖掘技术、ERP 系统、计算机信息管理系统；黄细妹(1987-)，女，硕士，主要研究方向为数据挖掘(429496342@qq.com)；李康顺(1962-)，男，教授，博导，博士，主要研究方向为演化计算、网络路由算法.

1.1 ID3 算法思路

ID3 算法的核心思想是利用信息增益度作为决策树节点属性选择标准。信息增益的原理来自信息论,它是使某个属性用来分割训练集而导致的期望熵降低,因此,信息增益越大的属性分裂的可能性越大[8,9]。

信息熵是用来度量整个训练实例集的不确定性,定义为

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

其中:X 为训练实例集。将训练实例集分为 n 类,第 i 类训练实例的个数为 $|X_i|$,训练实例总个数为 $|X|$,一个实例属于第 i 类的概率 $P(X_i) = |X_i|/|X|$。

如果选择属性 A 进行测试。设属性 A 具有属性值 $a_1, a_2, a_3, \cdots, a_n$,且在 $A = a_j$ 时属于第 i 类的实例个数为 $|X_{ij}|$,则有

$$P(X_i|A = a_j) = |X_{ij}|/|X_i| \tag{2}$$

当 $A = a_j$ 时的实例集记为 Y_j,此时决策树分类的不确定度就是训练集对属性 A 的条件熵,则有

$$H(Y_j) = -\sum_{i=1}^{n} P(X_i|A = a_j)\log_2 P(X_i|A = a_j) \tag{3}$$

节点 A 对于分类的信息熵记为

$$H(X|A) = \sum_{j=1}^{t}\sum_{i=1}^{n} P(A = a_j)P(X_i|A = a_j)\log_2 P(X_i|A = a_j) \tag{4}$$

属性 A 的信息增益为

$$\text{gain}(A) = H(X) - H(X|A) \tag{5}$$

运用 ID3 算法建立决策树是选择每次计算的信息增益最大的属性作为决策树的新节点,并对属性的每个属性值构建分支,依据此思路划分训练数据样本集。

1.2 ID3 算法的优缺点

ID3 算法在选择重要特征时利用的是互信息的知识,算法的基础理论比较简单、清楚[10,11],具有以下优势:

a)每一次的搜索都使用全部的训练样本,在一定程度上降低了个别噪声数据对构建的决策树的影响。

b)构建的决策树简单、直观,能够很好地从中提取出易于理解的分类规则。

c)擅长处理属性值为离散型的数据。

ID3 虽然是一种典型的决策树分类算法,但是仍然存在不足之处:

a)由于 ID3 算法是选择信息熵最大作为属性标准,因此,它将更偏向于取值较多的属性,但并不是属性值较多就是最优属性。

b)数据集越大,算法的计算量也增加得越大。

c)对数据噪声比较敏感。

d)处理的数据必须是离散型数据。

针对上述 ID3 算法的缺点,可以在提高算法计算时间、改进属性选择标准、处理连续型数据等方面对该算法进行改进。

2 算法改进

针对 ID3 算法偏向属性值较多的属性选择和算法计算量大这两个不足之处,将对此算法进行改进,主要是对 ID3 算法进行简化使其计算量减小。针对 ID3 算法偏向属性值较多的属性的不足而引入重要度的改进算法,最后结合这两种改进算法即在简化的基础上对 ID3 进行引入重要度处理得到简化改进算法。

2.1 ID3 算法简化研究

针对 ID3 算法的计算复杂度大的问题,提出了对信息量计算的改进方法,即利用数学上的泰勒公式和迈克劳林公式对 ID3 算法的信息增益进行简化计算,在很大程度上降低了算法计算的复杂度。

1)算法原理

ID3 简化方法的目的是使算法的计算量减小,这里主要是运用数学知识对 ID3 算法中的信息增益公式进行简化,使复杂的计算公式简化成为只有加、减、乘、除的四则运算公式,从而降低算法运算难度[12]。具体可将训练实例集化分为两类,即正例和反例。设正例实例集个数为 m,反例实例集个数为 n,属性 A 的取值为 $a_1, a_2, a_3, \cdots, a_n$,且当 $A = a_j$ 时,所在分支下的元素共有 $n_i + m_i$ 个。根据前面介绍的 ID3 算法的信息增益公式,其简化过程如下所示:

由式(1)可得

$$H(X) = -\frac{n}{n+m}\log_2\frac{n}{n+m} - \frac{m}{n+m}\log_2\frac{m}{n+m} \tag{6}$$

由式(3)可得

$$H(Y_j) = -\frac{m_j}{n_j+m_j}\log_2\frac{m_j}{n_j+m_j} - \frac{n_j}{n_j+m_j}\log_2\frac{n_j}{n_j+m_j} \tag{7}$$

由式(4)可得

$$H(X|A) = \sum_{j=1}^{t}\frac{n_j+m_j}{n+m}H(Y_j) \tag{8}$$

将式(7)代入式(8)并利用对数性质对其进行简化,由于 $\frac{1}{(n+m)\ln 2}$ 是一个固定值,所以可得

$$H^*(X|A) = \sum_{j=1}^{t}\left(-m_j\ln\frac{m_j}{n_j+m_j} - n_j\ln\frac{n_j}{n_j+m_j}\right) \tag{9}$$

根据数学上的等价无穷小原理,在 x 很小时,有 $\ln(1+x)\approx x$,可将式(9)简化为

$$H^*(X|A) = \sum_{j=1}^{t} -\frac{2m_jn_j}{n_j+m_j} \tag{10}$$

由此可得

$$\text{gain}^*(A) = \sum_{i=1}^{t}\frac{m_jn_j}{n_i+m_j} \tag{11}$$

可以将 $\text{gain}^*(A)$ 作为属性 A 简化处理后的信息增益。依据式(11)计算简化信息增益,并从中选择简化信息增益值最小的属性作为节点,根据标准重复选择属性直到分类结束。该简化方法相对于原来的 ID3 算法,在运算速度上较快,因为它只涉及到了加、乘、除运算。

2)算法不足

简化过程的最大优势是使 ID3 算法计算公式的对数运算转换为普通的四则运算,这大大地降低了算法的计算时间,提高了算法的执行效率,因此,简化后算法构成的决策树与原 ID3 算法构成的决策树应该完全一致,但结果并不是一致的。回顾简化的过程,可以发现在简化过程中存在误差问题。误差的产生可能使条件属性信息增益的值产生一定的误差,也可能使生成的决策树与 ID3 算法生成的决策树不相同,精度可能会有所降低。

2.2 引入属性重要度

针对 ID3 算法偏向选择属性取值多的属性,而这与实际情况往往不符的不足,引入一个属性重要度的概念。

1)算法原理

针对偏向属性值多这个不足,已经有很多学者对其进行了研究讨论,其中 C4.5 算法就很好地弥补了这个缺陷。C4.5 算法是通过信息增益率作为属性选择标准。对比信息增益和信息增益率的计算公式可以发现,信息增益率是通过引入分割信息量作为权衡因子对信息增益的多值取向进行弥补,使测试属性更符合现实情况。本节借鉴这一规律引入属性重要度概念来弥补这个缺陷。

属性重要度概念是指所有条件属性集中属性对信息增益贡献的重要性对比结果。其值可以定义为 A 属性的分支子集所占 A 实例的比重,此比重可以赋予一个值 a。a 值为增加此属性的重要度值,$-a$ 值可以用来降低多属性值的条件属性对信息增益的贡献,然后根据各个分支子集实例计算各个属性的熵值,通过比较此熵的大小来选择最优的属性。

重要度值 a 的作用是区分不同信息属性的重要性或依赖度。它是一个模糊的概念,可以认为 $\alpha \leqslant \min p_i, p_i$ 为属性 $A = a_i$ 时的条件概率。因此 $0 \leqslant \alpha \leqslant 1$,可将式(4)变化为

$$H_\alpha(X|A) = \sum_{j=1}^{t}\sum_{i=1}^{n} P((A=a_j) - \alpha) \times P(X_i|A=a_j)\log_2 P(X_i|A=a_j) \tag{12}$$

相应的信息增益由式(5)变化为

$$\text{gain}_\alpha(A) = H(X) - H_\alpha(X|A) \tag{13}$$

可以将 $\text{gain}_\alpha(A)$ 作为属性 A 的新的信息增益。依据式(13)计算信息增益并从中选择信息增益值最大的属性作为节点,根据标准重复递归选择属性直到分类结束。

通过对重要度概念的理解,应用中应该注意以下几个方面:a)对于符合实际情况的,在分类中起到重要作用的条件属性的重要值需要通过先验经验进行测试;b)通过对属性重要性的认识,可以很好地预防大数据掩盖小数据的现象;c)当条件属性中存在大量先验知识时,可根据实际的属性重要性来选择,逐步进行,提高效率。

2)算法不足

此算法虽然很好地解决了 ID3 算法的第二个缺点,但是未考虑其第一个缺点,算法的计算量并没有得到很好的解决。

2.3　ID3 改进算法

1)改进简化算法的不足

由于简化过程中是通过估算得到结果,因此简化过程必然引起误差。为了改善误差,可以假设每个属性的特征值个数为 L,通过大量实验证明在 $\text{gain}^{\&}(A)$ 上乘以 L 可以有效地弥补误差。

2)改进引入重要度算法的不足

简化方法可以有效地减少算法的计算量,因此,可结合简化方法与引入属性重要度方法来弥补引入重要度算法的不足,有效地结合简化方法和引入属性重要度方法,继承了这两种方法的优势。把上面介绍的式(9)~(12)结合变化为

$$H_\alpha{}^*(X|A) = \sum_{j=1}^{t}\left(\frac{m_j n_j}{n_j+m_j} + \frac{\alpha m_j n_j}{(n_j+m_j)^2}\right) \tag{14}$$

相应地式(13)变化为

$$\text{gain}_\alpha{}^*(A) = \sum_{j=1}^{t}\left(\frac{m_j n_j}{n_j+m_j} + \frac{\alpha m_j n_j}{(n_j+m_j)^2}\right) \tag{15}$$

可以将 $\text{gain}_\alpha{}^*(A)$ 作为属性 A 的改进的信息增益。依据式(15)计算改进信息增益并从中选择改进信息增益值最小的属性作为节点,根据标准重复递归的选择属性直到分类结束。该改进算法相对于原来的改进算法,在运算速度上较快,因为它只涉及到了加、乘、除运算;相对于算法和原 ID3 算法又引入了重要度概念,克服了这些算法的不足。

3)ID3 改进算法

结合 1)和 2)可以得到一种新的 ID3 改进算法,这种算法有效地克服了原 ID3 算法的不足之处,也弥补了简化算法的误差和引入重要度算法的计算复杂问题。该算法的具体情况是在式(16)中乘以属性的特征值个数 L,得到

$$\text{gain}_\alpha{}^*(A) = \sum_{j=1}^{t}\left(\frac{m_j n_j}{n_j+m_j} + \frac{\alpha m_j n_j}{(n_j+m_j)^2}\right)L \tag{16}$$

ID3 改进算法最终使用 $\text{gain}_\alpha{}^*(A)$ 的结果作为决策树分类属性选择标准。选取 $\text{gain}_\alpha{}^*(A)$ 取值最小的属性作为决策树的节点,根据标准重复递归的选择属性直到分类结束。

3　实例分析

3.1　举例

下面根据一个具体的实例来对上述介绍的分类算法进行比较分析。实例是根据气象部门报告的天气(outlook)、温度(temperature)、湿度(humidity)、风(windy)的记录来决定是否去参加户外运动(outdoor sport)。类别属性根据相关因素决定是否适合进行户外运动。样本训练集的实例如表 1 所示。

表 1　样本训练集

序号	outlook	temperature	humidity	windy	outdoor sport
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	moderate	high	false	yes
5	rainy	cold	medium	false	yes
6	rainy	cold	medium	true	no
7	overcast	cold	medium	true	yes
8	sunny	moderate	high	false	no
9	sunny	cold	medium	false	yes
10	rainy	moderate	medium	false	yes
11	sunny	moderate	medium	true	yes
12	overcast	moderate	high	true	yes
13	overcast	hot	medium	false	yes
14	rainy	moderate	high	true	no

3.2　实例结果分析

针对上面这些训练集,可以根据前面介绍的公式计算其信息增益、改进的信息增益,然后根据决策树算法建立树的规则,建立决策树。

通过计算表 1 的训练样本集的信息增益来建立决策树。具体的决策树如图 1 和 2 所示。

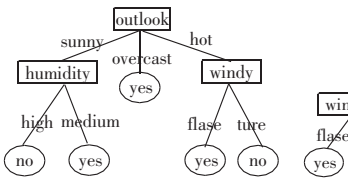

图 1　ID3 算法构造的决策树

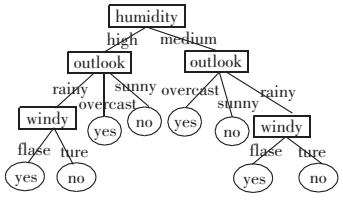

图 2　改进 ID3 算法建立的决策树

决策树构建成功了,接下来就是挖掘决策树的分类规则。通过图 1 可以得到五条决策树分类规则,如表 2 所示。

表 2　ID3 算法得到的分类规则

规则	outlook	temperate	humidity	windy	outdoor sport
1	overcast				yes
2	sunny		medium		yes
3	sunny		high		no
4	rainy			true	no
5	rainy			false	yes

在图 2 中,分别取四个属性(outlook、temperate、humidity、windy)的重要度值为 0.35、0、0、0,然后计算其改进的新的信息增益,结果得到的决策树如图 2 所示。从图 2 中可以得到八条决策树分类规则,如表 3 所示。

表 3 改进 ID3 算法得到的分类规则

规则	outlook	temperate	humidity	windy	outdoor sport
1	sunny		high		no
2	overcast		high		yes
3	rainy		high	true	no
4	rainy		high	false	yes
5	sunny		medium	false	no
6	sunny		medium		yes
7	rainy		medium	true	no
8	rainy		medium	false	yes

分析比较图 1 和 2 得到的分类规则可以发现,图 1 的决策树较复杂,生成的叶子节点较多,自然得到的决策树分类规则也较多,其中更加符合现实情况的分类规则也较多,比如 if (humidity = "medium" and outlook = "rainy" and windy = "true") then outdoor sport = "no"等。图 2 是通过降低 outlook 属性的重要性而得到的结果。图 2 的 outlook 属性更远离根节点,从而可以弥补 ID3 算法偏向选择属性值较多的属性的不足,即通过减少属性值多属性的重要性来克服 ID3 算法的属性取值偏向的不足。

4 实验结果分析

针对上面提出的 ID3 算法以及其改进算法,下面通过实验对其进行分析验证。实验主要是从算法的正确性和优化性两方面进行比较研究[13~15]。

4.1 实验条件

本实验主要是在 Weka 平台上进行的。Weka 的全名是 Waikato environment for knowledge analysis,是一款免费的、基于 Java 环境下的开源数据挖掘软件。该软件操作简单,由于是开源软件,可以很方便地将自己改进的算法添加其中,然后调试改进算法。

本实验是在 NetBeans IDE 7.0.1 RC1 中导入 Weka 的源代码,然后添加改进的算法源代码,最后在 NetBeans IDE 7.0.1 RC1 中编译 Weka 中的原 ID3 算法及其改进算法程序。

实验用到的计算机的基本配置是 CPU 为 AMD Athlon X2 Dual Core 1.81 GHz,内存为 1 GB;操作系统为 Windows XP。

4.2 实验内容

实验选择了 UCI 样本数据集中的 Car evaluation,将 Car evaluation数据集中的 car. data 数据导入 Weka 中,然后对其进行数据分类,最后对比实验结果来验证改进算法的正确性和优化性。

1)正确性验证

Car evaluation 数据集中的数据信息情况如图 3 所示。通过 ID3 的改进算法分析 car. data 数据,得到图 4 所示的结果。

比较图 3 和 4 可以发现,运用 ID3 改进算法分析的 car. data 数据集获得的实例数、属性、属性值等信息基本一致,从中可以验证 ID3 改进算法的正确性。

2)优化性分析

算法的优化性分析从算法所用的时间和算法的准确性方面进行比较。取 car. data 数据集中的不同记录数从时间和准确性两方面来分析,比较原 ID3 算法、改进的 ID3 算法和 C4.5 算法的结果。通过实验比较算法计算所用的时间如图 5 所示,准确性如图 6 所示。

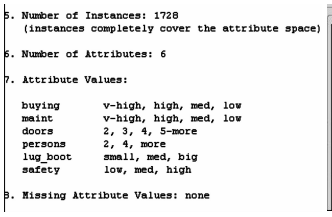

图 3 car.data 数据情况

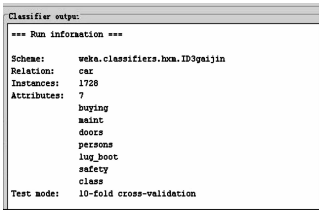

图 4 ID3 改进算法分析结果

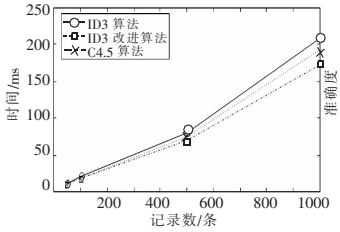

图 5 算法计算耗时比较

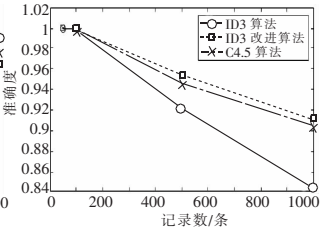

图 6 算法准确率比较

从图 5 中可以发现,改进的 ID3 算法在处理同样多的记录数时消耗的时间要比原 ID3 算法和 C4.5 算法少;从图 6 中可以发现,当数据集的记录数少时,这三种算法的准确率没有多大区别,到记录数达到一定数量后,改进算法的准确率明显要优于原 ID3 算法,但与 C4.5 算法的准确率相差不大。

下面进行的实验是选择 UCI 样本数据集中不同样本数据集进行分析,并且所选的样本数据记录数也不同,比较结果结果如图 7、8 所示。

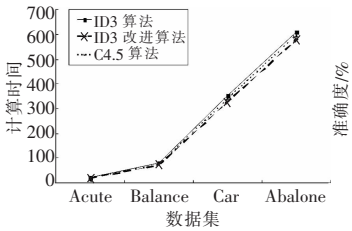

图 7 不同数据集下的算法计算时间比较

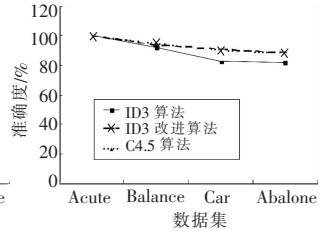

图 8 不同数据集下的算法准确度比较

图 7 给出了在不同数据集下原 ID3 算法、本文的改进 ID3 算法和 C4.5 算法的计算时间比较结果,可以发现本文算法在消耗时间方面要比原 ID3 算法和 C4.5 算法少;图 8 给出了原 ID3 算法、本文的改进 ID3 算法和 C4.5 算法的准确率比较结果,可以发现本文算法的准确率明显要优于原 ID3 算法,但与 C4.5 算法的准确率相差不大。

4.3 实验结论

改进 ID3 算法在减少原 ID3 算法运算时间的同时,还克服了原 ID3 算法偏向取属性值较多的属性的不足。实验结果分析也表明,改进的 ID3 算法具有更优的时间复杂度和准确率,尤其在数据集达到一定规模时,这种优越性体现得更明显。

5 结束语

ID3 算法是决策树算法中最典型的算法,有大量学者对其进行了研究分析,本文在前人的研究下,对 ID3 算法及其改进算法进行了详细介绍,通过具体实例系统地比较分析了这些算法,结果证明了改进算法的可行性,最后通过实验仿真进一步验证了改进算法的正确性和优化性。 (下转第 2908 页)

(上接第 2898 页)

通过本文的研究发现,改进算法对 ID3 算法是有效的,在以后的工作中也可以利用这些改进算法的思路去改进其他算法,如 C4.5 算法、贝叶斯算法等。

参考文献:

[1] 朱明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2008.

[2] HAN Jia-wei, KAMBER M. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2006.

[3] 王斌. 决策树算法的研究及应用[D]. 上海:东华大学,2008.

[4] 刘祺. 决策树 ID3 算法的改进研究[D]. 哈尔滨:哈尔滨工程大学,2009.

[5] 黄爱辉,陈湘涛. 决策树 ID3 算法的改进[J]. 计算机工程与科学,2009,31(6):109-111.

[6] 屈志毅,周海波. 决策树算法的一种改进算法[J]. 计算机应用,2008,28(6):141-143.

[7] 鲁为,王枞. 决策树算法的优化与比较[J]. 计算机工程,2007,33(16):189-190.

[8] 孙爱东,朱梅阶,涂淑琴. 基于属性值的 ID3 算法改进[J]. 计算机工程与设计,2008,29(12):156-160.

[9] QUINLAN J R. Simplifying decision trees[J]. International Journal Man-Machine Studies,1987,27(3):221-234.

[10] 冯少荣,肖文俊. 基于样本选取的决策树改进算法[J]. 西南交通大学学报,2009,44(5):643-646.

[11] 魏振钢,周翔. ID3 算法的一种改进算法[J]. 计算机科学,2010,37(7):104-107.

[12] MIR B O, DANIEL K, ASSAF S, *et al*. Hierarchical decision the induction in distributed genomic databases[J]. IEEE Trans on Knowledge and Data Engineering,2005,17(8):1138-1151.

[13] 韩松来,张辉,周华平. 决策树的属性选取策略综述[J]. 微计算机应用,2007,28(8):785-790.

[14] 韩松来,张辉,周华平. 基于关联度函数的决策树分类算法[J]. 计算机应用,2005,25(11):2655-2657.

[15] 张春来,张磊. 一种基于修正信息增益的 ID3 算法[J]. 计算机工程与科学,2008,30(11):46-47.