一种基于属性敏感度的决策树算法①

王　梅　于　京　(北京电子科技职业学院科技工程学院电子工程系 北京市 100016)

刘　光　(北京市电信工程设计院有限公司 北京市 100036)

摘　要: 决策树算法是数据挖掘中重要的分类算法。目前，已有许多构建决策树的算法，其中，ID3 算法是核心算法。本文首先对 ID3 算法进行研究与分析，针对计算属性的信息熵十分复杂的缺点，提出了一种新的启发式算法 SID3，它是基于属性对分类的敏感度的。文章最后通过实例对两种算法进行比较分析，结果表明，SID3 算法能够生成正确的决策树，并且使建树过程更简便，更快速。

关键词: 数据挖掘; 决策树; 分类; ID3 算法; 属性敏感度

Sensitive Attribute Algorithm for Decision Tree SID3

WANG Mei[1], YU Jing[1], LIU Guang[2]

(1. Dept. of Electronic Engineering, Institute of Science and Technology Engineering, Beiing Electronic Science and Technology Vocational College, Beijing 100016, China;

2. Beijing Telecom Engineering Design Institute Co, Ltd., Beijing 100036, China)

Abstract: Decision tree is the most important classification algorithm in data mining. At present, there are many decision tree algorithms, ID3 algorithm is the core one. This paper first studies and analyses the ID3 algorithm, then discusses the complicacy of computing the Information Entropy of attribute, and put forward a new heuristic based on the sensitive of attribute contributing to the classification. Finally, this paper compares the two algorithms by experiments, the results show that SID3 can generate the correct decision tree and the process is more simple, more quickly.

Keywords: decision tree; ID3 algorithm; sensitive of attribute

决策树方法是目前应用最广泛的归纳推理算法之一，是一种以实例为基础的归纳学习算法，通常用来形成分类器和预测模型，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成的分类规则，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，最后在决策树的叶结点得到结论。因此，从根到叶结点的一条路径就对应着一条合取规则，而整棵决策树就对应着一组析取表达式规则。

目前为止，决策树有很多实现算法。例如：由 Quinlan 在 1986 年提出的 ID3 算法[1]和在 1993 年提出的 C4.5 算法[2]等。这些算法多数是基于信息熵的，由于计算信息熵的公式复杂，过程繁琐，本文提出了一种新的基于属性对分类的敏感程度的启发式算法，通过对属性敏感度[3]的分析，选择敏感度最大的属性作为测试属性，来构造决策树，这种方法可以快速的构造出正确的决策树。

1　ID3算法

决策树方法的起源是概念学习系统(Concept Learning System, CLS)，然后发展到 ID3 方法而为高峰。Quinlan 提出的 ID3 算法通过对一个例子集进行学习生成一棵决策树，现假设一个例子仅属于两种分类之一：正例，即符合被学习的目标概念的例子；

① 收稿时间:2010-03-08;收到修改稿时间:2010-04-23

反例，即不符合目标概念的例子。另外，假设例子的所有属性都是离散属性。

1.1 ID3 算法描述

ID3 的基本概念[4]如下：

(1) 决策树中每一个非叶结点对应着一个非类别属性，树枝代表这个属性的值。一个叶结点代表从树根到叶结点之间的路径对应的记录所属的类别属性值。

(2) 每一个非叶结点都将与属性中具有最大信息量的非类别属性相关联。

(3) 采用信息增益来选择能够最好地将样本分类的属性。

信息增益是基于信息论中熵(Entropy)的概念[4]。设 S 是 s 个数据样本的集合，假定类标号属性具有 m 个不同值，定义 m 个不同类 $C_i(i=1, 2, \cdots, m)$。设 s_i 是类 C_i 中的样本数。对一个给定的样本分类所需的期望信息为：

$$I(s_1, s_2, \mathbf{L}\, s_m) = -\sum_{i=1}^{m} p_i lb(p_i)$$

其中，p_i 是任意样本属于 C_i 的概率，一般用 s_i/s 来估计。

设属性 A 具有 v 个不同值$\{a_1, a_2, \cdots, a_v\}$。可用属性 A 将 S 划分为 v 个子集$\{S_1, S_2, \cdots, S_v\}$，其中 S_j 包含了 S 中在 A 上具有值为 a_j 的一些样本。若 A 作为测试属性，设 s_{ij} 是子集 S_j 种类的样本数，根据属性 A 划分当前子集的熵为：

$$E(A) = -\sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \mathbf{L} + s_{mj}}{s} I(s_{1j}, s_{2j}, \mathbf{L}, s_{mj})$$

其中，$\frac{s_{1j} + s_{2j} + \mathbf{L} + s_{mj}}{s}$ 充当第 j 个子集的权，

$I(s_{j1}, s_{2j}, \mathbf{L}\, s_{mj}) = -\sum_{i=1}^{m} p_{ij} lb(p_{ij})$ 表示给定子集 S_j 的期望信息，$p_{ij} = \frac{s_{ij}}{|s_j|}$ 是 S_j 中的样本属于类 C_j 的概率。

由期望信息和熵值得到的在 A 上分支获得的信息增益为：

$$Gain(A) = I(s_1, s_2, \mathbf{L}\, , s_m) - E(A)$$

ID3 算法计算每个非类别属性的信息增益，并选取最高增益的属性(即最小熵值属性)作为给定集合 S 的测试属性。对被选取的属性创建一个结点，并以该属性标记，对该属性的每个值创建一个分支，并据此划分样本。此算法采用自上而下、分而治之的递归方式来构造一棵决策树。

1.2 ID3 算法的不足[5]

从上面的讨论可知，ID3 算法是把信息增益(信息熵)作为选择测试属性的标准，即树结点的选择策略。但在计算基于属性的信息熵时，公式比较复杂，计算量较大，相应的复杂度也高，当数据量很大的时候很耗费硬件资源，计算花费的时间较长。针对以上缺点，本文提出了一种基于属性敏感度的决策树生成算法 SID3，此算法中，不再把信息增益作为选择测试属性的标准，而是把属性敏感度作为度量标准，用更加简便的方法构造与 ID3 算法同样的决策树。

2 SID3算法

要构建一棵决策树，关键问题是如何选择测试属性。目前，已经有不少选择测试属性的度量标准[6-8]。本文主要是针对上述 ID3 算法的不足之处，引入了属性敏感度作为选择测试属性的标准。下面就来解释一下属性敏感度的定义以及算法的描述。

在用于构建决策树的样本集合中，有许多的特征属性，但是并不是所有的属性对于分类都起到关键作用，也就是说，每一个特征属性对于分类都有不同的敏感程度[3]。为了找出某些特征属性的敏感度，需要从表中删除一些属性以考察没有该属性分类会怎样变化。若去掉该属性分类亦相应改变，说明该属性的强度大，即敏感度高；反之，说明该属性的强度小，即敏感度低。由此引入属性敏感度的概念，定义如下：

定义 1. 设 S 是 s 个数据样本的集合，该样本集合中包含若干特征属性 A，B，…，若删除属性 A，样本集合中存在不一致的数据样本的个数为 m，则

称 M(A)为属性 A 的敏感度。

定义 2. 如果数据样本 X 所有的属性值与样本数据 Y 的相同，而分类不同，那么就称数据样本 X 与 Y 是不一致的。

本文提出的 SID3 算法的基本思想是以 ID3 算法为基础的：

从一棵空决策树开始，选择敏感度最大的属性作为测试属性，若两个属性的敏感度相同，则选择排序

靠前的属性作为测试属性。该测试属性对应决策树中的决策结点，根据该属性的值不同，可将训练样本分成相应的子集，如果该子集为空，或该子集中的样本属于同一类，则该子集为叶结点；否则该子集对应于决策树的内部结点，即测试结点，算法使用同样的过程，递归的选择敏感度最大的属性作为新的测试属性，并对该子集进行划分，直到所有的子集都为空或属于同一类。

SID3 算法与 ID3 算法的根本不同之处在于对测试属性的选择标准不同。虽然标准不同，但是 SID3 算法能够构建与 ID3 算法同样的决策树，并且在构建决策树的过程中，对于属性敏感度的计算非常简便，能够更快的确定测试属性，以加快决策树的生成。

3　算法实例分析

下面我们给出一个实例，分别采用 ID3 算法与 SID3 算法来构造决策树，并对结果进行分析。

表 1　实例数据集合

属性	穿衣指数	温度	湿度	风力	天气舒适度
1	较多	很高	很大	没有	不舒适（N）
2	较多	很高	很大	很大	不舒适（N）
3	较多	很高	很大	中等	不舒适（N）
4	正常	很高	很大	没有	舒适（Y）
5	正常	很高	很大	中等	舒适（Y）
6	很多	适中	很大	没有	不舒适（N）
7	很多	适中	很大	中等	不舒适（N）
8	很多	很高	正常	没有	舒适（Y）
9	很多	很高	正常	很大	不舒适（N）
10	较多	适中	很大	没有	不舒适（N）
11	较多	适中	很大	中等	不舒适（N）
12	很多	适中	正常	没有	不舒适（N）
13	很多	适中	正常	中等	不舒适（N）
14	较多	适中	正常	中等	舒适（Y）
15	较多	适中	正常	很大	舒适（Y）
16	正常	适中	很大	很大	舒适（Y）
17	正常	适中	很大	中等	舒适（Y）
18	正常	很高	正常	没有	舒适（Y）
19	很多	适中	很大	很大	不舒适（N）
20	正常	很高	正常	中等	舒适（Y）

表 1 给出了影响天气舒适度的几个相关指标的数据集合[9]，包含 4 个属性：穿衣指数，温度，湿度和风力。数据集被分为舒适(正例 Y)和不舒适(反例 N)两类。

首先，利用 ID3 算法对样本训练集进行分类，构建决策树。

由于 ID3 算法是选择最高增益的属性作为给定集合 S 的测试属性，通过信息增益的计算公式可知，属性的熵值越小，其信息增益越大，所以我们只需选择具有最小熵值的属性作为测试属性即可。下面计算出了各个属性的熵值。

由于穿衣指数在属性中具有最小的熵值，所以它首先被选作测试属性，并以此创建一个结点，用穿衣指数标记，对于每个属性值，各引出一个分支，

$$E(\text{穿衣指数}) = \frac{6}{20}\times 0 + \frac{7}{20}\times\left(\frac{2}{7}\text{lb}\frac{2}{7} + \frac{5}{7}\text{lb}\frac{5}{7}\right) + \frac{7}{20}\times\left(\frac{1}{7}\text{lb}\frac{1}{7} + \frac{6}{7}\text{lb}\frac{6}{7}\right) = 0.5092$$

$$E(\text{温度}) = \frac{11}{20}\times\left(\frac{4}{11}\text{lb}\frac{4}{11} + \frac{7}{11}\text{lb}\frac{7}{11}\right) + \frac{9}{20}\times\left(\frac{5}{9}\text{lb}\frac{5}{9} + \frac{4}{9}\text{lb}\frac{4}{9}\right) = 0.9661$$

$$E(\text{湿度}) = \frac{12}{20}\times\left(\frac{4}{12}\text{lb}\frac{4}{12} + \frac{8}{12}\text{lb}\frac{8}{12}\right) + \frac{8}{20}\times\left(\frac{5}{8}\text{lb}\frac{5}{8} + \frac{3}{8}\text{lb}\frac{3}{8}\right) = 0.9328$$

$$E(\text{风力}) = \frac{7}{20}\times\left(\frac{3}{7}\text{lb}\frac{3}{7} + \frac{4}{7}\text{lb}\frac{4}{7}\right) + \frac{8}{20}\times\left(\frac{4}{8}\text{lb}\frac{4}{8} + \frac{4}{8}\text{lb}\frac{4}{8}\right) + \frac{5}{20}\times\left(\frac{2}{5}\text{lb}\frac{2}{5} + \frac{3}{5}\text{lb}\frac{3}{5}\right) = 0.9880$$

数据集被划分为 3 个子集{穿衣正常，穿衣较多，穿衣很多}。其中穿衣正常子集中的样本属于同一类，于是该子集为叶结点，另外两个子集中的样本不属于同一类，于是为测试结点，下面对两个子集使用上述同样的方法，计算属性熵值，构建子树。最终得到的由算

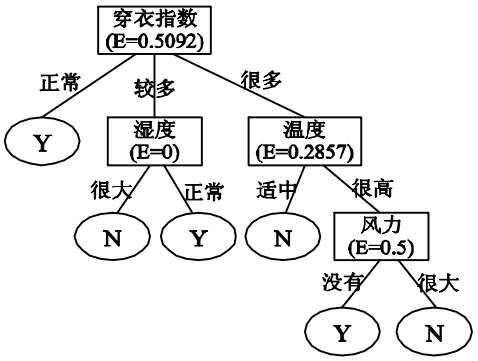

图 1　ID3 构建的决策树

法 ID3 构建的决策树如图 1 所示。

下面，利用 SID3 算法对样本训练集进行分类，构建决策树。

计算各属性的敏感度，对于穿衣指数属性，若删除此属性，可发现样本 1 和样本 4，样本 3 和样本 5，样本 7 和样本 17，样本 13 和样本 14，样本 16 和样本 19，这 5 对样本都具有相同的属性值，但是却属于不同的类。即删除穿衣指数属性后，样本集合中存在不一致的数据样本的个数为 5，所以穿衣指数属性的敏感度为 M(穿衣指数)=5。

同样可计算其他三个属性的敏感度分别为：

M(温度)=1；M(湿度)=1；M(风力)=1。

由于穿衣指数的敏感度最大，所以选择该属性作为测试属性，以此创建以穿衣指数为标记的结点，并且引出三个分支，分别为穿衣指数=正常，穿衣指数=较多，穿衣指数=很多。根据上述办法，计算不同子集中属性的敏感度，构建各个分支的子树 ，最终得到的由算法 SID3 构建的决策树如图 2 示。

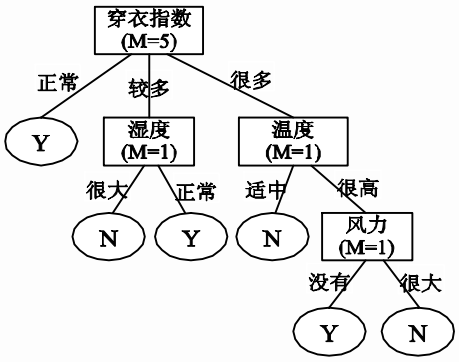

图 2 SID3 构建的决策树

比较两种决策树的生成算法,可以发现本文提出的算法 SID3 与 ID3 算法所生成的决策树一致，说明通过该算法可以构造出正确的决策树，并且可以得到正确的分类规则。并且通过实例证明，在分类问题中采用 SID3 算法生成决策树能够极大地降低计算复杂性，快速生成决策树，它是直接把属性和类别结果相联系，通过比较属性对分类结果敏感度大小，选择敏感度最大的属性作为测试属性，直观地构造出决策树，在这个计算过程中，计算的复杂性很大程度上得到了缓解，对计算机的硬件要求也相应降低，使得决策树思想在数据挖掘等领域的发展受计算机硬件的限制较小。

另一方面，由于把属性的敏感度作为选择测试属性的度量标准，整棵树的敏感程度将随着它本身的归纳过程而升高[3]。在构建决策树过程中，当敏感属性的值缺少或有噪音时，就会有较大的分类误差。实际上，数据的缺少或噪音广泛地存在，因此，在缺值情况很少或噪音较小的情况下，选择基于属性敏感度的方法构建决策树比较好，因为此方法要比较稳健，准确率高。然而缺值、噪音较多时，则需要增大训练集，对数据进行预处理并消除噪音干扰，则结果会更好。

4 应用案例

在证明算法的正确性之后，我们把该算法应用于某移动公司彩铃业务的客户价值分析案例中。该公司数据库中拥有大量的客户信息，但老客户流失与新客户获取问题十分严重，为了更好的发展该项业务，现需要从客户的基本信息与购买信息将客户进行分类，找出不同类别客户群的基本特征规律与购买规律，从而利用这些规律对不同类别的客户采取不同的经营策略，通过有效获取新客户与保留老客户，实现该业务用户数量增长，下载量增加，收入增长的目的。

该公司的数据库中共有 24 张表，其中用户表详细记录了公司用户的详细信息，多达 18 个属性，约有 360 万条记录，每条用户记录中包含了用户姓名，性别，年龄，婚姻状况等详细信息。用户订购铃音记录表详细记录了用户订购彩铃的详细信息，多达 21 个属性，约有 1406 万条记录，每条记录中包含了铃音订购时间，铃音资费，实际资费，收费类型等用户购买信息。在对数据进行清洗并对数据属性进行筛选后，我们选择如下属性用于进行客户分类：年龄、性别、家庭或个人年收入、学历、职业、铃音订购时间、实际资费、订购状态和操作渠道，其中前五个属性是用户表中的属性，后四个属性是订购铃音记录表中的属性。

彩铃业务中的客户被分为 5 组：活跃用户、沉默用户、流失用户、潜在用户和消极用户，对不同的客户分别赋以这五种不同的类别标记，选择三分之二的数据，采用 SID3 算法从客户购买行为与客户基本信息两个方面对客户进行分类，分别构建决策树，描述不同类别客户群的特征。对于得到的决策树，利用剩下三分之一的数据进行测试，发现改进算法得到你的决策树的分类规则正确率是 89.6%和 92.6%，达到之前定的 85%的目标,说明该算法构造的决策树是可靠、有效的，于是可以从得到的分类规则中归纳出不同类

(下接第 65 页)

(上接第 55 页)

别的客户特征描述，并根据这些描述对不同的客户群采取不同的销售态度与商业策略，从而推动彩铃业务长久健康的发展，为企业带来可观的经济效益。

5 结论

本文通过分析 ID3 算法的基本原理，针对该算法的计算较复杂的缺点提出了基于属性敏感度的决策树生成算法 SID3 来构造决策树。通过实例分析对两种算法进行分析与比较，展示了 SID3 算法的简洁、计算效率高等特性，从而可以在计算机硬件配置较低、资源消耗较少的条件下来快速生成决策树，得到相应的分类规则。最后通过具体案例再次验证了该算法的可实施性和有效性，但对于海量的、噪声较大的数据，只有先对数据进行有效的清洗与整理后，才能构建出正确率较高的决策树，从而得到有用的规则。

参考文献

1 Quinlan JR. Induction of Decision Tree. Machine Learing, 1986,1(1):81—106.

2 Quinlan JR. C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA, 1993.

3 王金凤,王熙照.敏感属性与不敏感属性对决策树的影响.计算机工程与应用, 2003,26:78—80.

4 毛国君,段立娟,王实等.数据挖掘原理与算法.北京:清华大学出版社, 2005.

5 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.辽宁师专学报(自然科学版), 2005,7(4):23—24.

6 韩松来,张辉,周华平.基于关联度函数的决策树分类算法.计算机应用, 2005,11:2655—2657.

7 倪春鹏,王正欧.一种新型决策树属性选择标准.武汉科技大学学报(自然科学版), 2004,1:437—440.

8 韩家新,王家华.一种以相关性确定条件属性的决策树.微机发展, 2003,5:38—39

9 郭玉滨.一种改进的 ID3 算法.肇庆学院学报, 2005,26(5):14—17.