决策树的优化算法*

刘小虎 李 生

(哈尔滨工业大学计算机科学系 哈尔滨 150001)

摘要 决策树的优化是决策树学习算法中十分重要的分支.以 ID3 为基础,提出了改进的优化算法.每当选择一个新的属性时,算法不是仅仅考虑该属性带来的信息增益,而是考虑到选择该属性后继续选择的属性带来的信息增益,即同时考虑树的两层结点.提出的改进算法的时间复杂性与 ID3 相同,对于逻辑表达式的归纳,改进算法明显优于 ID3.

关键词 机器学习,决策树,分类,信息增益,熵.

中图法分类号 TP18

自从 Quinlan[1]介绍了 ID3 算法以来,学者围绕该算法进行了十分广泛的研究.ID3 是基于信息熵的决策树分类算法,根据属性集的取值选择实例的类别.ID3 的算法核心是在决策树中各级结点上选择属性,用信息增益率作为属性选择标准,使得在每一非叶结点进行测试时,能获得关于被测试例子最大的类别信息,使用该属性将例子集分成子集后,系统的熵值最小.期望该非叶结点到达各后代叶结点的平均路径最短.使生成的决策树平均深度较小,提高分类速度和准确率.

ID3 的基本原理如下:设 $E=F_1\times F_2\times\ldots\times F_n$ 是 n 维有穷向量空间,其中 F_j 是有穷离散符号集,E 中的元素 $e=(v_1,v_2,\ldots,v_n)$,叫作例子,其中 $v_j\in F_j, j=1,2,\ldots,n$.设 PE 和 NE 是 E 的两个例子集,分别叫作正例集和反例集.

假设向量空间 E 中的正例集 PE 和反例集 NE 的大小分别为 p 和 n,ID3 基于下列两个假设:(1) 在向量空间 E 上的一棵正确决策树对任意例子的分类概率同 E 中正反例的概率一致;(2) 一棵决策树能对一例子作出正确类别判断所需的信息量为

$$I(p,n)=-\frac{p}{p+n}\log_2\frac{p}{p+n}-\log_2\frac{n}{p+n}. \tag{1}$$

如果以属性 A 作为决策树的根,A 具有 v 个值$\{v_1,v_2,\ldots,v_v\}$,它将 E 分为 v 个子集$\{E_1,E_2,\ldots,E_v\}$,假设 E_i 中含有 p_i 个正例和 n_i 个反例,子集 E_i 的信息熵为 $I(p_i,n_i)$,以属性 A 为根分类后的信息熵为 $E(A)=\sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i+n_i)$.因此,以 A 为根的信息增益是 $\mathrm{gain}(A)=I(p,n)-E(A)$.ID3 选择使 $\mathrm{gain}(A)$最大(即 $E(A)$最小)的属性 A^* 作为根结点.对 A^* 的不同取值对应的 E 的 v 个子集 E_i 递归调用上述过程,生成 A^* 的子结点 $B_1,B_2,\ldots,B_v$.

ID3 的基本原理是基于两类分类问题,但很容易将其扩展到多类,设样本集 S 共有 C 类样本,每类样本数为 $P_i, i=1,2,\ldots,C$.如果以属性 A 作为决策树的根,A 具有 v 个值 $v_1,v_2,\ldots,v_v$,它将 E 分为 v 个子集$\{E_1,E_2,\ldots,E_v\}$,假设 E_i 中含有第 j 类样本的个数为 $P_{ij}, j=1,2,\ldots,C$,那么,子集 E_i 的信息量是 $I(E_i)$.

$$I(E_i)=\sum_{j=1}^{C}-\frac{P_{ij}}{|E_i|}*\log\frac{P_{ij}}{|E_i|}, \tag{2}$$

以 A 为根分类后的信息熵为

$$E(A)=\sum_{i=1}^{v}\frac{|E_i|}{E|}*I(E_i). \tag{3}$$

选择属性 A^* 使 $E(A)$最小,信息增益也将最大.

理想的决策树分为 3 种:①叶结点数最少;②叶子结点深度最小;③叶结点数最少且叶子结点深度最小.决策树的

* 本文研究得到国家 863 高科技项目基金资助.作者刘小虎,1970 年生,博士生,主要研究领域为机器翻译,机器学习.李生,1943 年生,教授,博导,主要研究领域为机器翻译,人工智能.

本文通讯联系人,刘小虎,哈尔滨 150001,哈尔滨工业大学计算机科学系

本文 1997-05-26 收到原稿,1997-09-15 收到修改稿

好坏，不仅影响了分类的效率，而且影响分类的准确率. 因此，许多学者致力于寻找更优的启发式函数和评价函数. 洪家荣[2]、Pei-Lei Tu[3]等人分别证明了要找到这种最优的决策树是NP难题. 因此，人们为寻找较优的解，不得不寻求各种启发式方法. 文献[4,5]对属性集的选择进行了细致的研究，文献[3]采用了基于属性相关性的启发式函数，文献[6]对生成的决策树进行剪枝处理，文献[7]扩充了决策树，形成决策图.

文献[8]采用的优化算法简单，其基本思想是：首先用ID3选择属性F_1，建立树T_1，左右子树的属性分别为F_2，F_3；再以F_2，F_3为根，重建树T_2，T_3；比较T_1，T_2，T_3的结点个数，选择结点最少的树. 对于选定树的儿子结点采用同样的方法，递归建树. 尽管作者用一个实验证明能够建立理想的决策树，但算法有较大的弱点：时间开销太大，因为每选择一个新的属性，算法需要建立3棵决策树，从中选优.

本文第1节介绍改进的优化算法MID3，并分析时间复杂性. 第2节用实验比较改进算法与ID3. 最后总结全文.

1 优化算法

ID3选择属性A作为新的属性的原则是，A使得$E(A)$最大. 这种启发式方法存在一个弊端，即算法往往偏向于选择属性取值较多的属性，而属性值较多的属性却不总是最优的属性.[1]

受到更多攻击的是ID3学习简单的逻辑表达式能力较差.[3]本文针对这一问题提出了如下的改进方案. 设A为候选的属性，A有v个属性值，对应的概率分别为$P_1,P_2,\ldots,P_v$，按照最小信息熵原理对属性A扩展，$\{B_1,B_2,\ldots,B_v\}$为v个子结点选择的属性，分别对应的信息熵为$E(B_1),E(B_2),\ldots,E(B_v)$，则

$$E'(A)=\sum_{i}^{v}P_i * E(B_i). \tag{4}$$

算法选择属性A^*的标准是，A^*使得$E'(A)$最小.

算法的详细步骤如下：

(1) 对任意未选择的属性A，假设A有v个属性值，对应的概率分别为$P_1,P_2,\ldots,P_v$，以属性A扩展，生成v个子结点$\{B_1,B_2,\ldots,B_v\}$，B_i是属性A取第i个值时，按照最小信息熵原理选择的A的后继属性，分别对应的信息熵为$E(B_1),E(B_2),\ldots,E(B_v)$.

(2) 根据公式(4)，计算$E'(A)$.

(3) 选择A^*使得$E'(A^*)$最小，将A^*作为新选的属性.

(4) 利用步骤(1)的计算结果，建立结点A^*的后继结点$\{B_1,B_2,\ldots,B_v\}$.

(5) 对所有的B_i，若为叶结点，则停止扩展此结点，否则递归执行(1)～(5)的过程.

该算法(1)～(4)是选定A作为新的属性，与ID3相比，不是计算选择A后带来的信息增益，而是继续选择A的后继属性，计算系统的熵值. 也就是说，该算法改进了选择新属性的启发式函数，以达到更好的分类效果.

MID3的时间复杂性与ID3相同，两者的差别只在于属性选择的计算上，即ID3采用公式(3)选择属性，而MID3利用公式(4). 假设类别个数为m，属性个数为n，属性值的平均个数为v，$v\ll n$. 公式(2)的时间复杂性为$O(m)$，则公式(3)计算每个属性A的时间复杂性为$O(m*v)$，所以计算所有的属性的时间为$O(n*m*v)$. 这就是ID3选择一个结点的属性的时间开销. 比较公式(3)和(4)，很容易看出MID3选择一个结点的属性的时间复杂性为$O(v*n*m*v)$，由于$v\ll n$，$O(v*n*m*v)=O(n*m*v)=O(m*n)$. 即ID3与MID3在属性选择上时间复杂性相同.

2 实 验

我们使用F-family作为实例集比较算法学习的性能. F-family是测试集FAMn的家族，其中$n=k+2^k$(k为正整数). 测试集FAMn有n个特征，前k个特征相当于地址位，后面的2^k个特征相当于数据位. 由地址位决定例子的类别与某一个数据位相等. 所有属性和类取值为0和1. 例如，FAM6有6个特征$F1,F2,F3,F4,F5,F6$，类为C. 实例的个数为64. 类C与属性的关系为$C=(!\,F1\&!\,F2\&F3)|(!\,F1\&F2\&F4)|(F1\&!\,F2\&F5)|(F1\&F2\&F6)$. $F1$与$F2$相当于地址选择器，由$F1$与$F2$的4种组合决定C等于$F3,F4,F5$或$F6$.

对例子集FAM6，ID3学习到的决策树如图1所示，叶结点18个. 然而，最理想的决策树如图2所示，叶结点8个. 很明显，ID3学习到的决策树与最优树相距甚远. 利用改进的MID3算法学习的结果如图3所示，叶结点12个.

图中，用圆表示的结点为叶结点，其他的为中间结点. 每个中间结点的左儿子是属性值为0的分支，右儿子是属性值为1的分支.

由于$P(C=1|F1=1)=0.5,P(C=1|F2=1)=0.5,P(C=1|F3=1)=P(F1=0\&\&F2=0)+P(F1=1\|F2=1)$

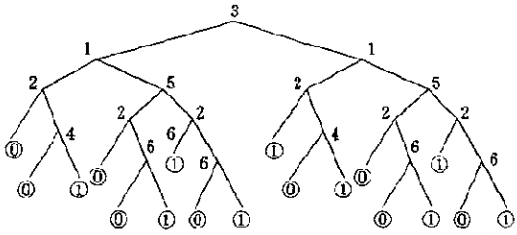

图1　ID3从实例集F6学习到的决策树

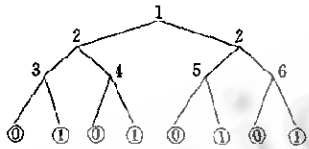

图2　实例集F6的理想决策树

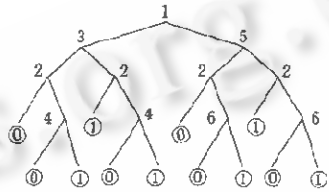

图3　MID3从实例集F6学习到的决策树

$*0.5=0.25+0.75*0.5>0.5$,同理,$P(C=1|F4=1)>0.5$,$P(C=1|F5=1)>0.5$,$P(C=1|F6=1)>0.5$,所以,$E(F3),E(F4),E(F5),E(F6)<E(F1),E(F1)$,ID3只会选择$F3$或$F4$,$F5$,$F6$作根结点,而不是最优的$F1$或$F2$.其中$P(X|Y)$表示条件概率,$P(X\&\&Y)$表示$X$与$Y$都真的概率,$P(X\|Y)$表示$X$或$Y$为真的概率.

改变例子空间,使用FAM6的变化形式FAM6A,FAM6B.在FAM6A中,$C=(F1\&F2\&F3\&F4)|(F1\&!F2\&F4\&F5)|(!F1\&F2\&F5\&F6)|(!F1\&!F2\&F6\&F3)$;在FAM6B中,$C=(F1\&F2\&(F3|F4))|(F1\&!F2\&(F4|F5))|(!F1\&F2\&(F5|F6))|(!F1\&!F2\&(F6|F3))$.FAM11是$k=3$的F-family成员,由属性$F1,F2,F3$选择$C$等于$F4,F5,\ldots,F11$.分别用ID3与MID3学习,比较结果如表1所示.可以看出,在这4个实验中,MID3虽然没有生成理想的决策树,但优于ID3.

表1　ID3与MID3的比较实验

(叶结点数,树高)	FAM6	FAM6A	FAM6B	FAM11
ID3	(18,5)	(22,6)	(22,6)	(54,6)
MID3	(12,4)	(12,5)	(12,5)	(40,5)
理想的决策树	(8,3)	(12,4)	(12,4)	(16,4)

作者在应用实际数据German以及Zoo(Hans Hofmann博士提供)时,发现MID3与ID3相比,性能也有提高,但不如针对逻辑表达式学习的优势明显.其主要原因在于,在逻辑表达式学习中,例子集合包括了每个属性的各种取值,并且每种组合只有1个例子,不呈现统计特征.而在实际例子空间中,特定属性值的组合可能出现多次或1次,甚至没有出现,体现了实际的概率,因此,ID3已经能够较好地分类.

3　结束语

本文以ID3为基础,提出了改进的优化算法.每当选择一个新的属性时,算法不是仅仅考虑该属性带来的信息增益,而是考虑到选择该属性后继续选择的属性带来的信息增益,即同时考虑树的两层结点.

通过分析算法的时间复杂性与ID3一致,对于逻辑表达式的学习,该算法弥补了ID3的不足.改进算法MID3目前只能针对特征值离散的情况,但特征可以是任意多值的,类别数也没有限制.

参考文献

1　Quinlan J R. Induction of decision trees. Machine Learning, 1986,(1):81～106

2　Hong J R. AE1: an extension approximate method for general covering problem. International Journal of Computer and Information Science, 1985,14(6):421～437

3　Tu Pei-lei, Chung Jen-yao. A new decision-tree classification algorithm for machine learning. In: Proceedings of the 1992 IEEE International Conference on Tools for Artificial Intelligence. Arlington, VA, 1992

4 Aha D W. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithm. International Journal of Man-Machine Studies, 1992,(6):267～287

5 Kira K, Rendell L. The feature selection problem: traditional methods and a new algorithm. In: AAAI-92 Proceedings of the 9th National Conference on Artificial Intelligence. 1992. 129～134

6 Jensen D. Adjusting for multiple testing in decision tree pruning. In: Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics. January 1997. 295～302

7 Oliver J, Dowe A D L, Wallace C S. Inferring decision graphs using the minimum message length principle. In: Proceedings of the 1992 Australia Joint Conference on Artificial Intelligence. Hobart, Tasmania, 1992. 361～367

8 Won Chan Jung, Bush Johes J, Chen Jian-hua. Optimization of decision tree. In: Proceedings of the 1991 IEEE International Conference on Tools for Artificial Intelligence. San Jose, CA-Nov, 1991

An Optimized Algorithm of Decision Tree

LIU Xiao-hu　LI Sheng

(Department of Computer Science　Harbin Institute of Technology　Harbin　150001)

Abstract　Optimization of decision-tree is a significant branch in decision tree learning algorithm. An optimized learning algorithm of ID3, a typical decision-tree learning algorithm is presented in this paper. When the algorithm selects a new attribute, not only the information gain of the current attribute, but also the information gain of succeeding attributes of this attribute is taken into consideration. In other words, the information gain of attributes in two levels of the decision tree is used. The computational complexity of the modified ID3 (MID3) is the same as that of the ID3. When the two algorithms are applied to learning logic expressions, the performance of MID3 is better than that of ID3.

Key words　Machine learning, decision-tree, classification, information gain, entropy.