基于决策支持度的决策树生成算法

关晓蔷[1],梁吉业[1],钱宇华[1],刘煜伟[2]
GUAN Xiao-qiang[1],LIANG Ji-ye[1],QIAN Yu-hua[1],LIU Yu-wei[2]

1.山西大学 计算机与信息技术学院,太原 030006
2.山西省军区司令部,太原 030013
1.School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2.The Shanxi Provincial Military District Headquarters,Taiyuan 030013,China
E-mail:gxq0079@163.com

GUAN Xiao-qiang,LIANG Ji-ye,QIAN Yu-hua,et al.Decision trees generation algorithm based on decision support degree.Computer Engineering and Applications,2008,44(27):148-150.

Abstract:Based on the viewpoint that conditions attributes have different decision support ability,the concept of decision support degree is introduced,and a novel algorithm for building decision tree is proposed,in which the decision support degree is regarded as heuristic information.The experimental analyzes on several public date sets show that unlike the decision tree built by traditional algorithms,the decision tree built by this algorithm has much better tree structure and classification precision.
Key words:decision tree;information system;decision support degree

摘 要:从条件属性对决策支持程度不同的角度出发,引入了决策支持度的概念,提出了一种以其为启发式信息的决策树生成算法。实验分析表明,相对于传统的决策树生成算法,此算法改善了决策树的结构,有效提高了决策分类的精度。
关键词:决策树;信息系统;决策支持度
DOI:10.3778/j.issn.1002-8331.2008.27.047 **文章编号**:1002-8331(2008)27-0148-03 **文献标识码**:A **中图分类号**:TP18

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,通常用来形成分类器和预测模型,可以对未知数据进行分类或预测。在各种决策树算法中最有影响的是Quinlan于1986年提出的ID3算法,但这种算法不是最优的,为此,出现了许多决策树的优化算法,如C4.5,Gehrke J提出的RainForest分类方法[1],Ruggieri R提出的EC4.5[2],洪家荣的决策树归纳学习算法[3],刘小虎的优化算法[4],Chou S等人提出的MMDT[5]等。

一般情况下,在众多能够拟合给定训练例子的决策树中,树越小则其预测能力越强。要构造尽可能小的决策树,关键在于选择恰当的属性,由于构造最小树是个NP完全问题,因此大量的研究只能采用启发式策略选择好的属性。在选择属性时,现在广泛采用的是Quinlan J R的算法体系,即ID3和C4.5算法,其中测试属性的次序选择是基于一个称作信息增益的统计属性。信息增益计算实际上是基于概率论和信息论中熵的概念,它反映了属性值出现的概率及其对类型值的"有效性"。

基于选择可区分元素对数最大的属性来作为启发式信息,本文提出了决策支持度的概念,并用来评测各个候选属性所导致的划分之间能将决策类进行正确区分的能力。以该度量为选择属性的标准,提出了一种基于决策支持度的决策树生成算法(Decision Trees Based on Decision Support degree,DTBDS)。仿真实验表明利用该算法得到的决策树较经典的ID3、C4.5等决策树算法通常有较小的规模和较高的分类精度。

1 信息系统

定义1 信息系统[6]

信息系统(Information system)形式上是一个四元组$S=(U,A,V,f)$。其中:U为对象的非空有限集合,称为论域;A为属性的非空有限集合;$V=\bigcup V_a$,V_a是属性a的值域;$f:U\times A\to V$是一个信息函数,它为每个对象的每个属性赋予一个信息值。通常$S=(U,A,V,f)$也简记为$S=(U,A)$。如果$A=C\cup D$,$C\cap D=\varnothing$,C称为条件属性集,D称为决策属性集。则具有条件属性和决策属性的信息系统$S=(U,C\cup D)$称为决策表。

定义2 不可区分关系[6]

令$S=(U,A,V,f)$表示信息系统,那么任意属性子集$B\subseteq A$所对应的不分明关系$INI(B)$可定义为:

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70471003);高等院校博士学科点专项科研基金(the China Specialized Research Fund for the Doctoral Program of Higher Education under Grant No.20050108004);教育部科学技术研究重点项目(No.206017);山西省青年科技研究基金(No.2006021019)。
作者简介:关晓蔷(1979-),女,讲师,主要研究方向:数据挖掘;梁吉业(1962-),男,教授,博士生导师,博士后,主要研究方向:粗糙集理论、数据挖掘、人工智能;钱宇华(1976-),男,博士研究生,主要研究方向:粒度计算;刘煜伟(1978-),男,硕士,主要研究方向:作战模拟。
收稿日期:2007-11-13 修回日期:2008-02-18

$IND(B)=\{(x,y)\in U\times U\mid\forall a\in B, f(x,a)=f(y,a)\}$

不分明关系也称作不可辨识关系。根据不分明关系,论域被划分成一个类族,每个类内部的对象都是不可区分的。不分明关系 $IND(B)$ 经常简记为 B。根据不分明关系 $IND(B)$ 可导出一个等价划分 $U/IND(B)$,可简记为 U/B。等价划分 U/B 中包含对象 x 的等价类一般记做 $[x]_B$。

令 $P,Q\subseteq A$,$P\preceq Q$ 或 $U/IND(P)\preceq U/IND(Q)$ 表示对任意 $X\in U/IND(P)$,存在 $Y\in U/IND(Q)$,使得 $X\subseteq Y$。这意味着 P 的划分比 Q 更精细,或者 Q 的划分比 P 更粗糙。$P\prec Q$ 或 $U/IND(P)\preceq U/IND(Q)$ 表示对于任意 $X\in U/IND(P)$,存在 $Y\in U/IND(Q)$,使得 $X\subseteq Y$,且存在 $X_0\in IND(P)$,$Y_0\in U/IND(Q)$,使得 $X_0\subset Y_0$。

2 基于决策支持度的决策树算法

2.1 属性选择原理

在各种决策树算法中,选择测试属性的方法各有不同,但是总的依据是选择进行分类时不确定性最小的属性作为测试属性,即测试属性能最大程度的做出正确的决策。

定义 3 组合熵[7]

令 $K=(U,R)$ 是一个近似空间,R 是 U 上的一个划分。R 的组合熵定义为:

$$CE(R)=\sum_{i=1}^{m}\frac{|R_i|}{|U|}\cdot\frac{C_{|U|}^2-C_{|R_i|}^2}{C_{|U|}^2}=\sum_{i=1}^{m}\frac{|R_i|}{|U|}\left(1-\frac{C_{|R_i|}^2}{C_{|U|}^2}\right)$$

这里 $C_{|U|}^2=\frac{|U|\times(|U|-1)}{2}$,$\frac{|R_i|}{|U|}$ 表示等价类 R_i 在论域 U 上的概率;$\frac{C_{|U|}^2-C_{|R_i|}^2}{C_{|U|}^2}$ 表示论域上互相可以区分的元素的对数在论域 U 上总的元素对数中所占的比率。

这个熵可以用相互区分的元素对数来刻画信息系统的知识含量,对进行精确分类有着更加清晰的刻画。

利用条件属性进行的划分能将类进行正确区分的能力越强,则相对决策属性利用条件属性不能区分的元素对数越少,可以区分的元素对数越多。基于这一思想,根据组合熵的定义可给出决策支持度的度量。

定义 4 决策支持度

令 $S=(U,C\cup D)$ 是一个决策系统,$C\cap D=\varnothing$,C 称为条件属性集,D 称为决策属性集,$R\subseteq C$,$U/R=\{R_1,R_2,\cdots,R_m\}$,$U/D=\{D_1,D_2,\cdots,D_n\}$。定义条件属性子集 R 对决策属性集 D 的决策支持度为:$\gamma(R,D)=1-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|R_i\cap D_j|\times|R_i-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}$,其中 $C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2$ 表示论域 U 上需要区分的元素对数,$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}|R_i\cap D_j|\times|R_i-D_j|$ 表示相对于 D 利用条件属性 R 不能区分的元素对数。

决策支持度 $\gamma(R,D)$ 表示划分 U/R 对划分 U/D 的支持程度,$\gamma(R,D)$ 的值越大,U/R 越接近 U/D,表明子集 R 对分类的贡献越大,那么选择属性集 R 进行分类的确定性越大。

决策支持度 $\gamma(R,D)$ 具有以下性质:

性质 1 $0\leq\gamma(R,D)\leq 1$。

性质 2 当 $U/R=U/D$ 时,$\gamma(R,D)=1$。

性质 3 当 $U/R=\delta$ 且 $U/D\neq\delta$ 时,$\gamma(R,D)=0$(其中 $\delta=\{U\}$)。

定理 1 令 $S=(U,C\cup D)$ 是一个决策表,$C\cap D=\varnothing$,C 称为条件属性集,D 称为决策属性集 $R\subseteq C$,$U/R=\{R_1,R_2,\cdots,R_m\}$,$U/D=\{D_1,D_2,\cdots,D_n\}$。则 $R\preceq D$ 当且仅当 $\gamma(R,D)=1$。

证明 (1) 假设 $R\preceq D$。由 $R\preceq D$ 知,对于任何 $R_i\in R$ 和任何 $D_j\in D$,均有 $R_i\cap D_j=\varnothing$ 或 $R_i\subseteq D_j$,因此,对于任何 $R_i\in R$ 和任何 $D_j\in D$,有 $|R_i\cap D_j|\times|R_i-D_j|=0$。从而有:

$$\gamma(R,D)=1-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|R_i\cap D_j|\times|R_i-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}=1$$

(2) 假设 $\gamma(R,D)=1$。如果 $R\preceq D$ 不成立,则存在一个 $R_k\in R$ 使得 $R_k\subseteq D_j$ 不成立($\forall D_j\in D$)。则 $|D_j\cap R_k|>0$ 且 $|R_k-D_j|>0$,$j=1,2,\cdots,n$。因此:

$$\gamma(R,D)=1-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|R_i\cap D_j|\times|R_i-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}\leq 1-\sum_{j=1}^{n}\frac{|R_k\cap D_j|\times|R_k-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}<1$$

这产生了一个矛盾,故 $R\preceq D$。证毕。

定理 2 令 $S=(U,C\cup D)$ 是一个决策表,$C\cap D=\varnothing$,C 称为条件属性集,D 称为决策属性集,$P\subseteq C$,$Q\subseteq C$,$U/D=\{D_1,D_2,\cdots,D_n\}$。如果 $P\prec Q$,则 $\gamma(P,D)\geq\gamma(Q,D)$。

证明 设 $U/P=\{P_1,P_2,\cdots,P_m\}$,$U/Q=\{Q_1,Q_2,\cdots,Q_n\}$,因为 $P\prec Q$,所以 $m>n$,且存在 $\{1,2,\cdots,m\}$ 的一个划分 $C=\{C_1,C_2,\cdots,C_n\}$ 使得 $Q_j=\bigcup_{k\in C_j}P_k$,$j=1,2,\cdots,n$。因此

$$\gamma(Q,D)=1-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r}\frac{|Q_i\cap D_j|\times|Q_i-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}=1-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r}\frac{|\bigcup_{k\in C_i}P_k\cap D_j|\times|\bigcup_{k\in C_i}P_k-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}\leq 1-\frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{r}\frac{|P_k\cap D_j|\times|P_k-D_j|}{C_{|U|}^2-\sum_{l=1}^{n}C_{|D_l|}^2}=\gamma(P,D)$$

证毕。

定理 2 的逆关系一般是不成立的。

2.2 算法描述

依据上述原理,以 $\gamma(R,D)$ 为选择属性的标准,设计出基于决策支持度的决策树生成算法 DTBDS,其中用 T 代表当前样本集,当前的候选属性集用 T_attributelist 表示,|T_attributelist| 表示候选属性集 T_attributelist 中的属性个数。候选属性集中的所有属性皆为离散型,连续值属性事先经过已有方法进行离散化。

算法 DTBDS(T, T_attributelist)

输入:决策表 $S=(U,C\cup D)$;

输出:一棵决策树。

算法步骤:

(1) 创建根结点 N,包含的样本集为 T,T_attributelist=C;

(2) IF T 都属于同一类 C,则返回 N 为叶结点,标记为类 C;

(3) IF T_attributelist 为空

则返回 N 为叶结点,标记 N 为 T 中出现最多的类;

(4) FOR EACH T_attributelist 中的条件属性 R_i,计算其对决策属性 D 的决策支持度:$\gamma(R_i,D)$;其中($1\leq i\leq$|T_attributelist|);

(5) N的测试属性Test_attributelist=T_attributelist中具有最高决策支持度的属性R;

(6) 将属性R从T_attributelist中去掉,形成新的测试属性列表T'_attributelist

(7) FOR EACH Test_attributelist=R的取值

{ 由N结点长出一个新子结点,

IF 新叶结点对应的样本子集T为空

则删除此结点;

ELSE

在该结点上执行

DTBDS(T,T'_attributelist)。

}

3 实验仿真与分析

决策系统S=(U,C∪D)如表1所示,C={a_1,a_2,a_3,a_4},D={d}。

表1 决策系统

U	a_1	a_2	a_3	a_4	d
1	1	2	2	1	1
2	1	3	3	2	1
3	1	2	2	2	1
4	2	2	2	1	1
5	2	3	2	3	2
6	1	3	2	1	1
7	1	2	3	1	2
8	2	3	1	3	2
9	1	2	2	3	1
10	1	1	3	3	1
11	2	1	2	3	2
12	1	1	2	2	1

利用定义4中决策支持度的计算公式分别计算每个条件属性对决策属性的决策支持度为:γ(U,a_1)=0.687 5;γ(U,a_2)=0.687 5;γ(U,a_3)=0.562 5;γ(U,a_4)=0.718 8;显然条件属性a_4的决策支持度最大,所以将a_4作为根,根据a_4的属性值把T中的对象分成3个子集,构造出3个分支;把每个子集中的对象作为新的论域,依次对每个分支中的对象按算法重复上述操作,得到图1所示的决策树。

对表1所示的决策系统,用ID3算法得到图2所示的决策树。

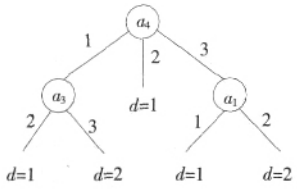

图1 基于决策支持度的决策树

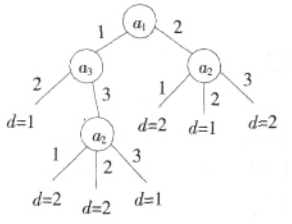

图2 基于ID3算法的决策树

从图2可以看出,基于ID3算法的决策树中总结点数是11个,叶子结点的个数为7个,生成的规则是7个,而图1利用决策支持度得到的决策树中的总结点数是8个,叶子结点个数为5个,生成的规则是5个。通过这个例子说明,本文提出的DTBDS算法可以降低生成决策树的复杂度,生成较少的规则数,有效的提高了分类效果。

实验采用Microsoft SQL Server 2000中的T-SQL语言编写存储过程,实现了ID3算法、C4.5算法与DTBDS算法,软件界面基于Microsoft Visual Studio.NET 2003平台,使用Visual Basic.NET语言开发。使用UCI(http://ftp.ics.uci.edu/pub/machine-learning-database)中的部分数据集作为实验数据(如表2所示),采用10次交叉验证对DTBDS算法、ID3算法和C4.5算法作进一步的性能测试比较,实验结果如表3所示。

表2 UCI实验数据集

数据集	训练集	测试集	属性数	类别数
Monk	124	432	7	2
Breast cancer	600	99	10	2
Balance scale	500	125	5	3

表3 算法性能比较

数据集	结点数			规则数			分类精度		
	ID3	C4.5	DTBDS	ID3	C4.5	DTBDS	ID3	C4.5	DTBDS
Monk	40.0	41.8	38.4	25.4	25.6	25.0	0.630	0.590	0.770
Breast cancer	112.2	111.0	114.2	92.2	86.5	92.7	0.900	0.900	0.910
Balance scale	448.5	447.1	446.2	352.9	352.1	350.7	0.351	0.357	0.353

从表3的数据可以看出,DTBDS算法能对实验数据中的训练事例正确分类,与ID3算法、C4.5算法进行比较,DTBDS算法多数情况下计算简单、计算量小,并且DTBDS算法比ID3算法、C4.5算法生成的决策树规模更小,得到的规则数更少,分类精度更高。

4 结语

本文提出的基于决策支持度的决策树生成算法DTBDS,利用不可区分关系定义了条件属性对于正确决策的支持度,并通过决策支持度找到对正确决策贡献最大的属性作为测试属性。由于决策支持度可以更好地刻画条件属性对数据进行正确区分的能力,所以根据决策支持度选择的属性对分类的贡献更大,实验证明,与传统决策树算法比较,本文提出的DTBDS算法进一步改善了决策树模型的复杂程度,并提高了决策树的分类精度。

参考文献:

[1] Gehrke J,Ramakrishnan R,Ganti V.RainForest:a framework for fast decision tree construction of large database[C]//Proceedings of 24th,International Conference on Very Large Data Bases,New York,USA,1998.

[2] Ruggieri S.Efficient C4.5[J].IEEE Transactions on Knowledge and Data Engineering,2002,14(2):438-444.

[3] 洪家荣,丁明峰,李星原,等.一种新的决策树归纳学习算法[J].计算机学报,1995,18(6):470-474.

[4] 刘小虎,李生.决策树的优化算法[J].软件学报,1998,9(10):797-800.

[5] Chou S,Hsu C L.MMDT:a multi-valued and multi-labeled decision tree classifier for data mining[J].Expert Systems with Application,2005,28:799-812.

[6] 梁吉业,李德玉.信息系统中的不确定性与知识获取[M].北京:科学出版社,2005.

[7] Qian Y H,Liang J Y.Combination entropy and combination granulation in incomplete information system[J].Lecture Note in Artificial Intelligence,2006,4062:184-190.