基于属性重要度的决策树算法

王 蓉[1] 刘遵仁[2] 纪 俊[2]

(青岛大学数据科学与软件工程学院 青岛 266071)[1] (青岛大学计算机科学技术学院 青岛 266071)[2]

摘 要 传统的ID3决策树算法存在属性选择困难、分类效率不高、抗噪性能不强、难以适应大规模数据集等问题。针对该情况,提出一种基于属性重要度及变精度粗糙集的决策树算法,在去除噪声数据的同时保证了决策树的规模不会太庞大。利用多个UCI标准数据集对该算法进行了验证,实验结果表明该算法在所得决策树的规模和分类精度上均优于ID3算法。

关键词 决策树,属性重要度,变精度粗糙集,属性约简,数据挖掘

中图法分类号 TP18 **文献标识码** A

Decision Tree Algorithm Based on Attribute Significance

WANG Rong[1] LIU Zun-ren[2] JI Jun[2]

(Department of Data Science and Software Engineering, Qingdao University, Qingdao 266071, China)[1]

(Department of Computer Science and Technology, Qingdao University, Qingdao 266071, China)[2]

Abstract The traditional ID3 decision tree algorithm is difficult in selecting attribute, its classification efficiency is not high, and anti-noise performance is not strong, so it is difficult to adapt to large-scale data set and other issues. Aiming at this situation, a decision tree algorithm based on attribute significance and variable precision rough set was proposed to ensure that the tree size is not too large while removing the noise data. The algorithm was validated by using multiple UCI standard data sets. The experimental results show that the algorithm is superior to the ID3 algorithm in the scale and classification accuracy of the decision tree.

Keywords Decision tree, Attribute significance, Variable precision rough set, Attribute reduction, Data mining

1 引言

分类是一种重要的数据分析形式,是数据挖掘中一项非常重要的任务。在众多的分类算法中,决策树因分类速度快、精度高、模式易于理解而受到了广泛的关注[1-2]。

ID3算法[3]是一种被广泛使用的决策树算法。该算法高将信息熵作为节点选择属性的标准,执行过程简单且高效,但其也存在一些不足:对于一个存在大量冗余属性的数据集而言,ID3算法形成的决策树过于庞大,且没有很好的抗噪性[4-6]。

Wang[7]首次将变精度粗糙集引入到决策树的构建中,其将最大的变精度明确区作为节点选择属性的标准。随后,研究人员在此基础上进行了进一步的改进和应用[8-17]。但以上研究未考虑冗余属性对算法的影响以及条件属性对决策属性的影响。

为此,本文提出了一种基于属性重要度及变精度粗糙集的决策树算法。该算法对数据集进行了属性约简,在去除噪声数据的同时保证了决策树的规模不会太庞大;考虑到条件属性对决策属性的影响,将属性重要度与β-正域相结合作为节点选择属性的标准;针对过拟合现象,通过引入置信度*delta*来控制决策树的生长。与ID3算法相比,本文算法具有有效性。

2 相关概念与原理

2.1 知识约简

定义1 给定一个知识库$K=(U,S)$和知识库中的一个等价关系簇$P\subseteq S$,$\forall R\in P$,若$IND(P)=IND(P-\{R\})$成立,则称知识R为P中不必要的,否则称R为P中必要的。

定义2(知识约简) 给定一个知识库$K=(U,S)$和知识库上的一簇等价关系$P\subseteq S$,对任意$G\subseteq P$,若G满足以下个两条件:

(1)G是独立的;

(2)$IND(G)=IND(P)$。

则称G是P的一个约简,记为$G\in RED(P)$,其中,$RED(P)$表示P的全体约简组成的集合。

定义3(知识的核) 给定一个知识库$K=(U,S)$和知识库上的一簇等价关系$P\subseteq S$,对任意的$R\in P$,若R满足$IND(P-\{R\})\neq IND(P)$,则称R为P中必要的,P中所有必要的知识组成的集合称为P的核,记为$CORE(P)$。

2.2 属性重要度

在决策表中,不同的条件属性相对于决策属性有不同的重要性。

本文受国家自然科学基金项目(61503208)资助。

王 蓉(1989-),女,硕士生,主要研究方向为粗糙集理论、数据挖掘,E-mail:475985222@qq.com;刘遵仁(1963-),男,博士,硕士生导师,主要研究方向为粗糙集理论、智能计算、数据挖掘等,E-mail:liuzunren@126.com(通信作者);纪 俊(1982-),男,博士,主要研究方向为数据挖掘、大数据应用、转化医学等,E-mail:1120108823@qq.com。

定义 4(决策表中属性的重要度) 给定一个决策表 $DT=(U,C\cup D,V,f)$,$\forall B\subseteq C$,$\forall \beta\in C$ 以及 $\forall \alpha\in C-B$,定义:

$$sig(\beta,C;D)=\frac{|pos_C(D)|-|pos_{C-\{\beta\}}(D)|}{|U|} \quad (1)$$

为条件属性 β 对条件属性全集 C 相对于决策属性 D 的重要度。

2.3 可变精度粗糙集[14]

定义 5 给定知识库 $K=(U,S)$,其中,U 为论域,S 表示论域 U 上的等价关系簇,则 $\forall X\subseteq U$ 和论域 U 上的一个等价关系 $R\in IND(K)$,定义子集 X 关于知识 R 的下近似和上近似分别为:

$$\underline{R}(X)=\{x|(\forall x\in U)\wedge([x]_R\subseteq X)\}$$

$$\overline{R}(X)=\{x|(\forall x\in U)\wedge([x]_R\cap X\neq\varnothing)\}$$

其中,$pos_R(X)=\underline{R}(X)$ 为 X 的 R 正域。

定义 6(β-正域) 给定一个信息系统 $S=(U,A,V,f)$,对于每个子集 $X\subseteq U$ 和等价关系 R,$U/R=\{Y_1,Y_2,\cdots,Y_n\}$,则定义 β-正域为:

$$\underline{R}_\beta(X)=\bigcup\{Y_i|Y_i\in U/R:P(Y_i,X)\geqslant\beta\}$$

其中,$P(Y_i,X)=\frac{|Y_i\cap X|}{|Y_i|}$ 为集合 Y_i 到 X 的正确分类率。

由以上定义可知:当 β 取值为 1 时,变精度粗糙集退化为传统的 Pawlak 粗糙集;且随着 β 的减小,正域范围越大,这说明变精度粗糙集模型对噪声数据有一定的容忍度。

3 基于属性重要度的决策树算法

本文首先采用 Pawlak 粗糙集中的约简算法[15]来对数据集进行约简,然后将约简后的数据集用于决策树的构造。构造决策树的过程主要涉及两个方面:1)分类属性的选择;2)决策树的规模。叶节点的选择方式对其具有决定性的作用。

3.1 分类属性的选择标准

基于 2.3 节中的可变精度粗糙集,Wang[7]将 β-正域最大的属性作为节点分类的属性;Hong[8]将 β-边界域最小的属性作为节点分类的属性。这些方法虽然可以选择划分效果较好的属性作为节点,但是未考虑条件属性对决策属性的影响。

为此,本文提出了一种将属性重要度与 β-正域相结合的方式,按照 k 权重进行组合,将属性重要度与 β-正域相结合作为节点选择属性的标准。

定义 7 对于一棵决策树,在节点属性的选择过程中定义判断标准的启发式函数为:

$$S(C)=k*\mathrm{sig}(\beta,C,D)+(1-k)*\gamma(C) \quad (2)$$

其中,k 是权重系数,$\gamma(C)=\frac{|\underline{R_\beta(C)}|}{|U|}$。

由定义 7 可知:当 k 的取值为 0 时,启发式函数退化为 Wang 的选择标准,即以 β-正域率作为节点属性选择的标准,这种单一的标准会使所得到的决策树的分类精度不高。因此,鉴于条件属性对决策属性的影响,本文增加属性重要度作为分类属性的选择标准。

3.2 叶节点的选择方式

在决策树的构建过程中,为了避免出现过拟合的现象,引入置信度 $delta(0.8\leqslant delta<1)$。

定理 1 决策表 $DT=(U,C\cup D,V,f)$,$\forall P\subseteq C$,满足 $IND(P)=\{X_1,X_2,\cdots,X_p\}$,$IND(D)=\{Z_1,Z_2,\cdots,Z_q\}$。若存在 $P(X_i,Z_i)\geqslant delta$,则将此节点作为叶子节点,不再进行划分。

证明:在 $0.8\leqslant delta<1$ 的情况下,对于任意 X_i,若满足 $P(X_i,Z_i)\geqslant delta$,则 Z_i 一定是唯一的,因此其叶节点的分类也一定是唯一的。

由定理 1 可知,$delta$ 值的设定将影响叶节点的选择。当 $delta$ 值设定得过小(即置信度较低)时,将使得原本的分支节点被判断为叶子节点,从而出现欠拟合现象,进而导致决策树的分类能力降低。当 $delta$ 值设定过大,即置信度较高时,将使得原本的叶子节点被判断为分支节点,从而出现过拟合现象,同样导致决策树的分类能力降低。本文在实验部分对 $delta$ 的取值进行了讨论。

3.3 算法描述

根据上述分析,提出一种基于属性重要度的决策树算法(Decision Tree Algorithm Based on Attribute Significance,DTAAS)。该算法的主体部分采用了递归思想,其具体步骤为:1)根据 Pawlak 粗糙集中的属性约简算法对数据集进行约简,在去除冗余属性的前提下开始构建决策树;2)根据式(2)计算 max($S(\alpha)$)的属性 α,并将其作为分支节点。在属性 α 各分支下的节点中,统计不同决策属性值的样本个数,然后分别求得不同决策属性值的样本占所有样本的比值。若存在某一比值大于设定的 $delta$,则将此节点作为叶子节点,且其取值为该比值对应的决策属性值;若不存在这样的属性值,则将此节点作为分支节点,按照第二步继续分析此节点。

DTAAS 算法的具体步骤如下:

```
Input:S=(U,C,D,V,f),β,delta
Output:决策树 T
Step1  初始化数据集 data,RED=∅;
Step2  计算 CORE(C),RED=CORE(C),C=C-CORE(C);
Step3  α=max(sig(β,C,D));
       if POS_C(D)≠POS_{CORE(C)+α}(D)
         RED=RED+α;
         C=C-α;
         go to Step3;
       end
Step4  Tree=∅;
Step5  Node=max(k*sig(α,C,D)+(1-k)*γ(α));
       if P(X_α,D_i)≥delta
         Node.label=D;
         Tree=Tree+Node;
       else
         根据节点属性值分裂为 N_1,N_2,…,N_i;
         Node.value=α.value;
         Tree=Tree+Node;
         go to Step5;
       end
Step6  return Tree;
```

4 实验分析

由以上分析可知,k 的取值影响着 DTAAS 算法的分类精度,$delta$ 的取值影响着决策树的规模。因为 k 和 $delta$ 是两个独立的变量,所以本实验采取以下步骤:首先,在确定 $delta$ 的情况下分析 k 的取值对算法精度的影响,从而确定 k 的较好取值区间;然后,在确定 k 的情况下分析 $delta$ 的取值对决策树规模的影响,从而确定 $delta$ 的较好取值区间;最后,在确定 k 和 $delta$ 的情况下,比较 DTAAS 算法和 ID3 算法在分类精度和决策树规模上的差别,从而验证 DTAAS 算法的有效性。

4.1　实验环境

UCI(University of California Irvine)提供了一系列用于测试的标准数据集。本文从 UCI 中选取了 4 个离散型数据集，如表 1 所列。

表 1　数据集描述

编号	数据集	样本数	属性数	类别数
1	Zoo	101	17	7
2	Tic	958	9	2
3	Car	1728	6	4
4	Nursery	12960	8	5

本实验在一台 CPU 为 G645 和内存为 4GB 的 PC 机上采用 Windows 7 环境下的 MATLAB R2014a 进行算法仿真。

实验方案如下：对于一个数据集，算法每次随机选取 2/3 的样本作为训练样本，将剩余的样本作为测试样本。算法共执行 20 次，算法的分类精度和决策树的节点数取均值。

4.2　k 和 *delta* 的取值区间

4.2.1　k 权值的确定

由式(2)可知，当 k 取不同值时，DTAAS 算法在构建树的过程中所选择的分支节点不同，进而造成分类精度不同。在区间[0.1，0.9]上，以 0.1 为递增值，取 9 个不同的 k 值。取 $delta=0.92$，在不同的 k 值下，记录在 Zoo，Tic 和 Car 这 3 个数据集下 DTAAS 算法的分类精度，实验结果如图 1 所示。

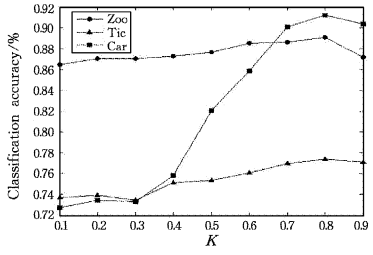


图 1　k 与分类精度的关系

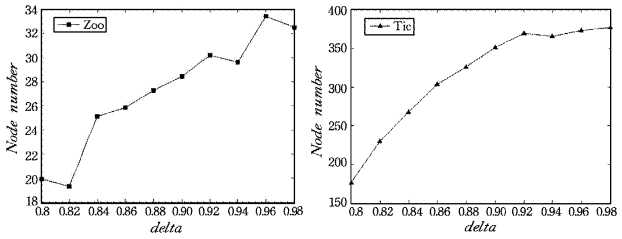
图 1 中，随着 k 值的增大，3 条折线逐渐上升，直至达到某个顶点后再开始下降，称此顶点对应的 k 值为饱和点，这说明算法的分类精度受属性重要度和 β-正域率共同影响。3 条折线的饱和点均不小于 0.8，且在 k 值从 0.1 增至饱和点的过程中，属性重要度对分类属性选择标准的贡献率逐渐增大，算法的分类精度逐渐提高，这说明相较于 β-正域率，属性重要度对算法的影响更大。在 k 值从饱和点增至 0.9 的过程中，属性重要度逐渐等价于属性选择标准，算法的分类精度逐渐降低，这说明 β-正域率在分类属性的选择标准中是必不可少的。通过上述分析证明了将属性重要度与 β-正域相结合作为选择分类属性标准的合理性。其次，在 k 值从 0.1 增至饱和点的过程中，对规模较小的 Zoo 和 Tic 数据集而言，其折线上升较为缓慢；对规模较大的 Car 数据集而言，其折线上升较为快速。在折线的下降过程中也存在该情况，这说明规模越大的数据集对 k 的取值越敏感。

基于上述实验结果，当本文实验中的 k 取值为 0.8 时，DTAAS 算法的分类精度最高，效果最为理想。

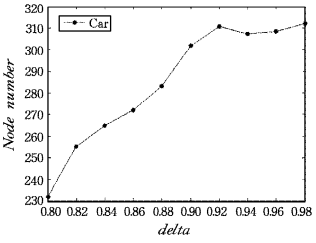
4.2.2　*delta* 值的取值区间

由定理 1 可知，当 *delta* 的取值不同时，DTAAS 算法在构建树的过程中所选择的叶子节点不同，从而造成所得的树的规模不同，即节点个数不同。在区间[0.8，0.98]上，以0.02 为递增值，取得 10 个不同的 *delta* 值。取 $k=0.8$，在不同的 *delta* 值下，记录在 Zoo，Tic 和 Car 这 3 个数据集下 DTAAS 算法的节点个数，实验结果如图 2 所示。

在图 2 中，随着 *delta* 的增大，3 条折线逐渐上升，直至达到某个顶点后趋于平缓，称此顶点对应的 *delta* 值为饱和点。这说明 *delta* 取值直接影响着算法得到的树的节点个数。在 *delta* 值从 0.8 增至饱和点的过程中，算法得到的树的节点个数逐渐增加，这说明算法对数据集的拟合程度增加，即算法的可靠性逐渐增加。在 *delta* 值从饱和点增至 0.98 的过程中，算法得到的树的节点个数逐渐稳定，这说明算法对数据集的拟合程度已经饱和，即算法的可靠性不再增加。因此为了避免因 *delta* 取值过小造成的算法欠拟合问题以及 *delta* 取值过大造成的算法过拟合问题，在保证算法可靠性的前提下，在本文实验中 0.92 是较好的 *delta* 取值。



(a)在 Zoo 数据集上 *delta* 与节点个数的关系　(b)在 Tic 数据集上在 *delta* 与节点个数的关系



(c)在 Car 数据集上 *delta* 与节点个数的关系

图 2　*delta* 与节点个数的关系

4.3　DTAAS 算法和 ID3 算法的对比

根据上述分析，DTAAS 算法中 $k=0.8$，$delta=0.92$。分别在表 1 中的 4 个数据集上运行 DTAAS 算法和 ID3 算法，其运行结果如表 2 所列。

表 2　实验结果

编号	数据集	分类精度/%		节点个数	
		ID3	DTAAS	ID3	DTAAS
1	Zoo	100	90.8	23	26
2	Tic	77.7	77.3	419	366
3	Car	85.8	90.7	559	303
4	Nursery	84.5	97.0	1747	816

相对而言，前两个数据集的规模较小，后两个数据集的规模较大。表 2 中各数据集在两种算法下的分类精度和节点个数的折线图如图 3、图 4 所示。

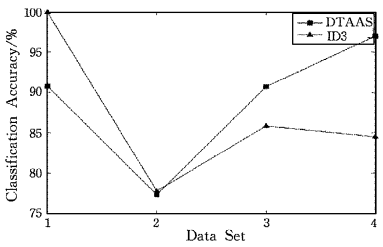


图 3　分类精度折线图

在图 3 中，对于前两个数据集，DTAAS 算法的折线处于

ID3 算法折线的下方，这说明相较于 ID3 算法，DTAAS 算法对规模较小的数据集的分类效果较差；对于后两个数据集，DTAAS 算法的折线处于 ID3 算法折线的上方，这说明相较于 ID3 算法，DTAAS 算法对规模较大的数据集的分类效果较好；且随着数据集规模的增加，DTAAS 算法和 ID3 算法在分类精度上的差值逐渐增大，最高可达到 13 个百分点，这进一步说明在分类精度方面 DTAAS 算法对规模较大的数据集的运行效果更好。

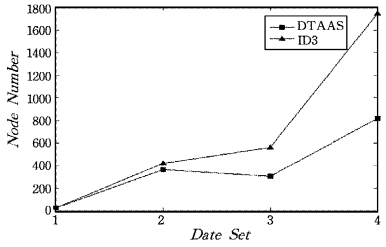


图 4 节点个数折线图

在图 4 中，随着数据集规模的增加，两条折线均呈上升趋势，这说明数据集规模越大，两种算法得到的树的节点数越多。DTAAS 算法的折线处于 ID3 算法折线的下方，这说明相较于 ID3 算法，DTAAS 算法得到的树的节点数更少，这意味着 DTAAS 算法的规则更少，其时间开销更少。且随着数据集规模的增加，DTAAS 算法和 ID3 算法在节点数上的差值逐渐增大，最高可以缩减一半左右，这进一步说明在节点个数方面，对于规模不同的数据集，DTAAS 算法对规模较大的数据集的运行效果更好。

综上，相较于 ID3 算法，DTAAS 算法在对规模较大的数据集进行分类任务时的分类精度更高，得到的树的节点数更少，算法运行得更快。DTAAS 算法适用于规模较大的数据集。

结束语 为了提高传统决策树的分类精度，本文将属性重要度与 β-正域相结合，提出了一种新的分类属性的选择标准，且通过置信度 *delta* 来控制决策树的生长，提出了 DTAAS 算法。实验证明，相较于 ID3 算法，DTAAS 算法在针对数据量较大的数据集进行决策树构造时效果更好，可以在提高分类精度的同时减小决策树的规模。但如何将粗糙集知识与决策树算法进行更好的结合，使得生成的决策树最优是未来的研究方向。

参 考 文 献

[1] 梁凤兰. 优化决策树改进挖掘算法仿真[J]. 计算机仿真，2013，30(11)：264-267.

[2] 张桉，曹健. 面向大数据分析的决策树算法[J]. 计算机科学，2016，43(6A)：374-379.

[3] QUINLAN J R. Induction of Decision Trees[J]. Machine Lerning，1986，1(1)：81-106.

[4] QUINLAN J R. Simplifying Decision Trees[J]. International Journal of Man-machine Studies，1987，27(3)：221-234.

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

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

[7] WANG S Q，WEI J M，YOU J P，et al. A VPRSM based approach for inducing decision trees[C]// RSKT2006. Chongqing，China，2006：421-429.

[8] 洪雪飞，徐维祥. 基于变精度粗糙集的决策树改进方法[J]. 计算机工程与应用，2009，45(13)：163-165.

[9] 丁春荣，李龙澍. 变精度粗糙集模型在决策树构造中的应用[J]. 计算机工程与科学，2010，32(7)：86-88.

[10] 鄂旭，任骏原，毕嘉娜，等. 基于粗糙变精度的食品安全决策树研究[J]. 计算机技术与发展，2014，24(1)：242-245.

[11] BARANAUSKAS J A. The number of classes as a source for instability of decision tree algorithms in high dimensional datasets [J]. Springer，2015，43(2)：301-310.

[12] LIANG C Q，ZHANG Y，SHI P，et al. Learning accurate very fast decision trees from uncertain data streams[J]. Taylor & Francis，2015，46(16)：3032-3050.

[13] 王婧，王兴伟，赵悦. 基于变精度粗糙集决策树垃圾邮件过滤[J]. 系统仿真学报，2016，28(3)：705-710.

[14] APNIK V. The nature of statistical learning theroy[M]. New York：Springer，1995.

[15] PAWLAK Z，SO-WINSKI R. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research，1994，72(3)：443-459.

[16] LIU Y，HUANG W，JIANG Y，et al. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences，2014，271(7)：65-81.

[17] 娄畅，刘遵仁，郭功振. 基于块集的邻域粗糙集的快速约简算法[J]. 计算机科学，2014，41(S2)：337-339.

(上接第 118 页)

[2] 路纲，周明天，唐勇，等. 任意图支配集精确算法回顾[J]. 计算机学报，2010，33(6)：1073-1087.

[3] 马晨明，王万良，洪榛. 无线传感器网络(k，m)-容错连通支配集的分布式构建[J]. 计算机科学，2016，43(1)：128-133.

[4] WANG D，ZHANG Q，LIU J. The self-pro-tection problem in wireless sensor networks[J]. ACM Transactions on Sensor Networks (TOSN)，2007，3：20.

[5] WANG Y，LI X，ZHANG Q. Efficient algorithms for p-self-protection problem in static wireless sensor networks[J]. IEEE Transactions on Parallel and Distributed Systems，2008，19：1426-1438.

[6] PFAFF J，LASKAR R C，HEDETNIEMI S T. NP-completeness of total and connected domination and irredundance for bipartite graphs[R]. Clemson University，1983.

[7] HE J，LIANG H. Complexity of total {k}-domination and related problems [C]// Proceedings of Frontiers in Algorithmics and Algorithmic Aspects in Information and Management(FAW-AAIM). Berlin：Springer Berlin Heidelberg，2011：147-155.

[8] ALBER J，NIEDERMEIER R. Improved tree decomposition based algorithms for domination-like problems[C] // Proceedings of the 5th Latin American Symposium on Theoretical Informatics (LATIN'02). Berlin：Springer Berlin Heidelberg，2002：613-627.

[9] ARCHDEACON D，ELLIS-MONAGHAN J，FISCHER D，et al. Some remarks on domination[J]. Journal of Graph Theory，2004，46：207-210.

[10] ZHAO W，WANG H，XU G. Total k- domination number in graphs[J]. International Journal of Pure and Applied Mathematics，2007，35(2)：235-242.

[11] 骆伟忠，冯启龙，王建新，等. 完全 p-支配集的参数算法[J]. 计算机学报，2013，36(9)：1868-1879.

[12] GANIAN R，SLIVOVSKY F，SZEIDER S，et al. Meta-kernelization with structural parameters[J]. Journal of Computer and System Sciences，2016，82(2)：333-346.