维普资讯 http://www.cqvip.com



# 一种改进的基于正区域的决策树算法*)

高 静[1] 杨炳儒[1] 徐章艳[1,2] 宋 威[1]
(北京科技大学信息工程学院 北京 100083)[1] (广西师范大学计算机系 桂林 541004)[2]

**摘 要** 首先分析了基于正区域、基于粗糙边界和基于依赖度的属性选择标准的关系,并证明了这三种属性选择标准彼此等价。然后以正区域的属性选择标准为代表,分析了基于正区域的决策树生成算法的优点和不足。针对这些不足,给出了一种新的属性选择标准,即基于伴随正区域的属性选择标准。用新的属性选择标准生成的决策树一般具有叶子数目较少,叶子的平均深度也较小,且叶子具有较强的泛化能力。最后,用一实例说明了新的属性选择标准的优越性。

**关键词** 决策树,粗糙集,正区域,粗糙边界,依赖度,伴随正区域

## Improved Algorithm for Decision Tree Based on Positive Region

GAO Jing[1] YANG Bing-ru[1] XU Zhang-yan[1,2] SONG Wei[1]
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083,China)[1]
(Department of Computer, Guanxi Normal University, Guangxi 541004 ,China)[2]

**Abstract** In this paper, the relationship of selected attribute standards between based on positive region, based on rough bound and based on attribute dependency is firstly analyzed. At the same time, it is proved that the three kinds of selected attribute standards are equivalent to each other. And then advantages and disadvantages of algorithm for constructing decision tree based on positive region are analyzed. Meanwhile, aiming at these disadvantages, a new selected attribute standard based on adjoint positive region is proposed. The decision tree constructed with the new standard of attribute selection has the following characteristics: fewer leaf nodes, fewer levels of average depth, better generalization of leaf nodes. Finally an example is used to illustrate the advantages of this new selected attribute standard.

**Keywords** Decision tree,Rough set,Positive region,Rough bound,Attribute dependency,Adjoint positive region

## 1 引言

自从 Quinlan[1,2]介绍了 $ID_3$ 算法以来,学者围绕该算法进行了十分广泛的研究。$ID_3$ 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。$ID_3$ 算法的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,使得在对每一个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。使用该属性将实例集分成子集后,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。但是 $ID_3$ 算法本身也存在很多不足。针对这些不足,文献[2-4]提出了各种改进算法,使得 $ID_3$ 算法更加完善。

由波兰学者 Pawlak[5]教授提出的粗糙集模型是分析不完整、不精确信息系统的有力工具。近年来,该模型在机器学习、数据挖掘、人工神经网络等多个领域得到了广泛的应用。一些学者将粗糙集理论[6,7]引入决策树,提出了基于粗糙集的决策树生成算法[8-14]。文献[9,10]是用粗糙边界的大小作为属性选择标准,但在实际计算时,粗糙边界最小的属性会有多个,即用粗糙边界作为选择标准不够细腻,同时得到的决策树往往因为分枝过细,使得得到的决策树的泛化能力不好[10];针对这些问题,文献[10]给出了一个新的剪枝算法,决策树通过剪枝后能得到较合理的决策树,不足之处是决策树的计算量较大。文献[11]是以正区域作为属性的选择标准,它的不足同文献[9]一样,为了改进正区域这种选择标准不够细腻,文献[11]给出参数依赖度的策略,能得到较好的决策树。文献[12]中给出的属性选择标准是属性的依赖度,它实际上与正区域等价,因而它的不足与文献[11]相同。文献[13,14]是以属性重要度作为属性的选择标准,它虽然考虑了所有属性之间的关系,但也存在标准不够细腻的不足,另外属性重要度的计算量要比其它的选择标准大得多。

本文首先分析了基于正区域、基于粗糙边界和基于依赖度的属性选择标准,并证明了这三种属性选择标准彼此等价。然后以正区域的属性选择标准为代表,分析了基于正区域的决策树生成算法的优点和不足。其优点为:得到的决策树通常叶子结点深度较小。其不足为:(1)在属性选择时,通常正区域相同的属性较多,不易选择属性,若任意选择,则生成的决策树差别可能较大,这说明该属性选择标准太粗略;(2)属性选择时,有可能选择的属性会产生很多叶子结点,这样,有些叶子结点上包含的实例数太少,以至这些叶子结点的泛化能力太小而没有意义,同时在预测时是比较费时。针对这些不足,本文给出了一种新的属性选择标准,即基于伴随正区域的属性选择标准。新的属性选择标准具有如下优点:(1)继承了基于正区域的属性选择标准的优点,即优先考虑一个属性的正区域大小,这样能保证生成的决策树中,叶子结点深度较

*)基金项目:国家自然科学基金重点项目(69835001)资助;教育部科技重点项目([2000]175)资助;北京市自然科学基金项目(4022008)资助。
**高 静** 讲师,博士研究生,研究方向为决策树,数据挖掘;**杨炳儒** 教授,博士生导师,主要研究方向为知识发现与智能系统、柔性建模与集成技术;**徐章艳** 副教授,博士研究生,研究方向为粗糙集的理论及其应用,数据挖掘,近似推理;**宋 威** 博士研究生,主要研究方向为数据挖掘。

维普资讯 http://www.cqvip.com

小。(2)在正区域大小相同的情况下,再考虑属性分出来的叶子数目(注意所选属性分出的叶子的决策属性值若相同则合并叶子)和叶子所覆盖的实例数,这样能保证生成的决策树的叶子数目较小,且生成的叶子具有较大的泛化能力。因而用新的属性选择标准生成的决策树一般具有叶子数目较少,叶子的平均深度也较小,且叶子具有较强的泛化能力。最后,用一实例说明了新的属性选择标准的优越性。

## 2 相关工作的分析

为了分析相关工作,我们先引入如下的基本概念。

**定义 1**[8] 一个信息系统(决策表)可以定义为:$S=(U,C,D,V,f)$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 是论域;$C$ 为条件属性集;$D$ 为决策属性集;$f:U\times(C\cup D)\to V$ 是信息函数,其中 $F=C\cup D$,$V=\bigcup V_a$,$a\in F$,$V_a$ 表示属性 $a$ 的值域。

**定义 2**[9] 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$\forall X\subseteq U$,记 $A_*(X)=\bigcup\{A_j\mid A_j\subseteq X\}$,则称 $A_*(X)$ 为 $X$ 在 $U$ 上关于 $A$ 的下近似集;记 $A^*(X)=\bigcup\{A_j\mid A_j\cap X\neq\emptyset\}$,则称 $A^*(X)$ 为 $X$ 在 $U$ 上关于 $A$ 的上近似集;记 $NEG_A(X)=\bigcup\{A_j\mid A_j\cap X=\emptyset\}$,则称 $NEG_A(X)$ 为 $X$ 在 $U$ 上关于 $A$ 的负区域;记 $BND_A(X)=A^*(X)-A_*(X)$,则称 $BND_A(X)$ 为 $X$ 在 $U$ 上关于 $A$ 的边界区域。

**定义 3**[9] 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分,称

$BND_A(D)=\bigcup_{D_i\in U/D}(A^*(D_i)-A_*(D_i))$ 为条件属性集 $A$ 在论域 $U$ 关于决策属性 $D$ 的粗糙边界,简称为粗糙边界。

**定义 4**[11-14] 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分,称 $POS_A(D)=\bigcup_{D_i\in U/D}A_*(D_i)$ 为条件属性集 $A$ 在论域 $U$ 上关于决策属性 $D$ 的正区域,简称为正区域;称 $\gamma_A(D)=|POS_A(D)|/|U|$ 为决策属性 $D$ 在论域 $U$ 上关于条件属性集 $A$ 的依赖度,简称依赖度。

**引理 1** 设 $A,B$ 是任意两个集合,则有 $A-B=A\cap\overline{B}$;$A\cap B=\overline{\overline{A}\cup\overline{B}}$;$\overline{\overline{A}}=A$;$A\cup B=\overline{\overline{A}\cap\overline{B}}$,其中 $\overline{A}$ 表示集合 $A$ 的补集。

**定理 1** 设 $A_1,A_2,B_1,B_2$ 是任意四个集合且 $A_1\cap B_2=\emptyset$,$A_2\cap B_1=\emptyset$,则有

$(A_1-B_1)\cup(A_2-B_2)=(A_1\cup A_2)-(B_1\cup B_2)$

证明:$(A_1-B_1)\cup(A_2-B_2)=(A_1\cap\overline{B_1})\cup(A_2\cap\overline{B_2})$

$=(A_1\cup A_2)\cap(A_1\cup\overline{B_2})\cap(\overline{B_1}\cup A_2)\cap(\overline{B_1}\cup\overline{B_2})(A_1\cap B_2$

$=\emptyset,A_2\cap B_1=\emptyset\Rightarrow A_1\cup\overline{B_2}=\overline{B_2},A_2\cup\overline{B_1}=\overline{B_1})$

$=(A_1\cup A_2)\cap\overline{B_2}\cap\overline{B_1}\cap(\overline{B_1}\cup\overline{B_2})$

$=(A_1\cup A_2)\cap(\overline{B_2\cup B_1})\cap(\overline{B_1\cap B_2})$

$=(A_1\cup A_2)\cap(\overline{(B_2\cup B_1)\cup(B_1\cap B_2)})$

$=(A_1\cup A_2)\cap(\overline{B_2\cup B_1})$

$=(A_1\cup A_2)-(B_2\cup B_1)$

**定理 2** 设 $A_1,\cdots,A_n,B_1,\cdots,B_n$ 是任意 $2n$ 个集合且 $A_i\cap B_j=\emptyset(i\neq j,i,j=1,2,\cdots,n)$,则有 $\bigcup_{i=1}^{n}(A_i-B_i)=\bigcup_{i=1}^{n}A_i-\bigcup_{i=1}^{n}B_i$。

证明:仿定理 1 即可证明。

**定理 3** 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分,则有

$A^*(D_i)\cap A_*(D_j)=\emptyset\quad(i\neq j,i,j=1,2,\cdots,h)$。

证明:

$A_*(D_j)=\bigcup(X\mid X\in U/A\wedge X\subseteq D_j\}$,$A^*(D_i)=\bigcup(X\mid X\in U/A\wedge X\cap D_i\neq\emptyset\}$;假设 $A^*(D_i)\cap A_*(D_j)\neq\emptyset(i\neq j)$,则有

$X_0\in U/A\wedge X_0\subseteq A^*(D_i)\wedge X_0\subseteq A_*(D_j)$(设 $D_k=\{x\mid x\in U,f(x,D)=d_k\}(k=1,2,\cdots,h)$),即一定存在 $x\in X_0$ 使得 $f(x,D)=d_i$。另一方面,由于 $X_0\subseteq A_*(D_j)$,故有 $f(x,D)=d_j$,而 $i\neq j$,从而有 $d_i\neq d_j$,即 $f(x,D)=d_i\neq d_j=f(x,D)$,这就产生矛盾。从而假设不成立。由 $i$ 和 $j$ 的任意性,知命题成立。

**定理 4** 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分,则有:

$$BND_A(D)=\bigcup_{D_i\in U/D}(A^*(D_i)-A_*(D_i))$$
$$=\bigcup_{D_i\in U/D}A^*(D_i)-\bigcup_{D_i\in U/D}A_*(D_i)$$

证明:由定理 2 和定理 3 即得。

**定理 5** 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分,则有:$\bigcup_{D_i\in U/D}A^*(D_i)=U$。

证明:显然有 $\bigcup_{D_i\in U/D}A^*(D_i)\subseteq U$。设 $\forall x\in U,f(x,D)=d_j$,则有 $[x]_A\subseteq A^*(D_j)$,从而 $x\in\bigcup_{D_i\in U/D}A^*(D_i)$,由 $x$ 的任意性知,$U\subseteq\bigcup_{D_i\in U/D}A^*(D_i)$,从而有 $\bigcup_{D_i\in U/D}A^*(D_i)=U$。

**定理 6** 在信息系统 $S=(U,C,D,V,f)$ 中,对 $\forall A\subseteq C$,设 $U/A=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性集 $A$ 对论域 $U$ 的划分,$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分,则有:$BND_A(D)=U-POS_A(D)$。

证明:由定理 4 和 5 即得。

**定理 7** 基于正区域的属性选择标准、基于依赖度的属性选择标准与基于粗糙边界的属性选择标准彼此等价。

证明:$|POS_{\{b\}}(D)|=\max_{a\in C}\{|POS_{\{a\}}(D)|\}$

$\Leftrightarrow\gamma_{\{b\}}(D)=|POS_{\{b\}}(D)|/|U|\geqslant$
$\quad|POS_{\{a\}}(D)|/|U|(\forall a\in C)$

$\Leftrightarrow\gamma_{\{b\}}(D)=\max_{a\in C}\{\gamma_{\{a\}}(D)\}$

$\Leftrightarrow|BND_{\{b\}}(D)|=|U-POS_{\{b\}}(D)|$
$\quad\leqslant|U-POS_{\{a\}}(D)|(\forall a\in C)$

$\Leftrightarrow|BND_{\{b\}}(D)|=|U-POS_{\{b\}}(D)|$
$\quad=\min_{a\in C}|U-POS_{\{a\}}(D)|$

在基于正区域的属性选择标准中,是以选择正区域最大的属性作为分割属性;而在基于依赖度的属性选择标准中是以选择属性依赖度最大的属性作为分割属性;在基于粗糙边界的属性选择标准中是以选择粗糙边界最小的属性作为分割标准。从上面的分析知,三种属性选择标准是彼此等价的。

基于属性重要度的选择标准,由其计算公式易知,计算一个属性的重要度的时间复杂度是计算一个属性的正区域的时间复杂度的 $|C|$ 倍。由于计算一个属性的重要度太费时,因

维普资讯 http://www.cqvip.com

此本文就不讨论这种属性选择标准。由定理 7 知三种属性选择标准彼此等价，因此我们就只拿出基于正区域的属性选择标准来分析这些标准的优缺点。

在文献[11]中，研究者曾指出基于正区域的属性选择标准的不足，并给出了一些改进方法。我们从正区域的本质入手，先分析基于正区域的属性选择标准的优点。

理想的决策树分为三种：(1)叶子结点最少；(2)叶子结点深度最小；(3)叶子结点最少且叶子结点深度最小。实际的决策树很难满足这些标准。特别是用到的属性选择标准通常是尽可能地满足这些标准。在 $ID_3$ 算法中，是以属性的信息增益作为属性的选择标准，从属性的信息增益的计算公式来看，属性的信息增益越大，由该属性分割出的决策树的平均深度较小，因此 $ID_3$ 算法是一种较好的决策树生成算法。但由于它在选择属性时偏向于属性取值较多的属性，这样，得到的决策树就不大理想。针对这些不足，许多学者给出了改进的 $ID_3$ 算法。

由正区域的定义易知，一个属性的正区域越大，就说明该属性一次能划分出更多的具有决策结果的实例，用这一标准得出的决策树通常是叶子结点深度较小，也算是一种较好的属性选择标准。但是这种方法具有如下不足：(1)在属性选择时，通常正区域相同的属性较多，不易选择属性，若任意选择，则生成的决策树差别可能较大。这说明该属性选择标准太粗略[11]。(2)属性选择时，有可能选择的属性会产生很多叶子结点，这样，有些叶子结点上包含的实例数太少，以至这些叶子结点的泛化能力太小而没有意义，同时在预测时是比较费时[10]。针对这些不足，文献[10]给出了剪枝算法，尽可能地去掉那些泛化能力很小的叶子，从而得到一棵较小，较理想的决策树。这种方法得到的决策树虽然较理想，但其计算量要比一般的决策树生成算法大得多，不是特别理想的决策树生成算法。文献[11]中给出一种新的属性选择标准——参数依赖度，用该方法每次选择参数依赖度最大的属性，虽然增强了决策树的叶子结点的泛化能力，但是每次选择的属性的正区域就不一定是最大的，也就是说得到的决策树的叶子深度不是最小的。为了更好地说明这一问题，我们用下面的例子说明。

例 1：$U=\{a1,a2,a3,a4,a5,a6,a7,a8,a9\}$，$P,Q,R$ 是条件属性，$Y$ 是决策属性。$U/Y=\{\{a1,a2,a3,a4,a5,a6\},\{a7,a8,a9\}\}$；$U/P=\{a1,a2\},\{a3,a4,a5\},\{a7,a8\},\{a6,a9\}\}$；$U/Q=\{\{a1,a2,a3,a4\},\{a5,a7,a8\},\{a6,a9\}\}$；$U/R=\{\{a1\},\{a2\},\{a3\},\{a4\},\{a5\},\{a7,a8\},\{a6,a9\}\}$。

易算出：$POS_{\{P\}}(Y)=\{a1,a2,a3,a4,a5\}$；$POS_{\{Q\}}(Y)=\{a1,a2,a3,a4\}$；$POS_{\{R\}}(Y)=\{a1,a2,a3,a4,a5\}$；

参数依赖度为：

$$K_P=\max_{X\in U/P\wedge X\subseteq Y_i\wedge Y_i\in U/Y}\frac{|X|}{|Y_i|}$$

$$=\max\{|\{a1,a2\}|/|\{a1,a2,a3,a4,a5,a6\}|,$$

$$|\{a3,a4,a5\}/\{a1,a2,a3,a4,a5,a6\}\}$$

$$=\max\{2/6,3/6\}=1/2$$

同理可求出：$K_Q=2/3$；$K_R=1/6$。

按基于正区域的属性选择标准应该选择属性 $P$ 或者 $R$；按基于参数依赖度的选择标准应该选择属性 $Q$。显然最合理的选择属性应该是 $P$，因为就已知的信息来看，由 $P$ 分割出的决策树的叶子结点最少，且得到的决策树的叶子深度最小。因为由 $P$ 分出两个叶子另还剩下 4 个实例；由 $Q$ 虽然只分出一个叶子，但另还有 5 个实例，实例 5 还得另外产生一个叶子，且深度大于 1。因此，属性 $Q$ 不一定有属性 $P$ 好。$R$ 自然不是最好的，因为它一次就产生 5 个叶子，每个叶子就一个实例，泛化能力也不好。若将同一属性分出来的叶子结点(决策属性值相同)合并，则属性 $P$ 和 $R$ 则只产生一个叶子结点，且分出的叶子结点数是最少，深度也最小。本文给出的属性选择标准将考虑这种情况。

为了解决上述问题，本文在第 3 节中给出一种新的属性选择标准。

## 3 改进的算法

为给出改进的算法，我们先给出如下定义。

**定义 5** 在信息系统 $S=(U,C,D,V,f)$ 中，对 $\forall a\in C$，设 $U/\{a\}=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性 $a$ 对论域 $U$ 的划分，$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分，定义：$|POS_{\{a\}}(D)|+\sum_{i=1}^{h}\frac{|\{a\}_*(D_i)|}{|U|}\frac{|\{a\}_*(D_i)|}{|D_i|}$ 为属性 $a$ 的伴随正区域，记为 $BPOS_{\{a\}}(D)$。

例 2：$U=\{a1,a2,a3,a4,a5,a6,a7,a8,a9\}$，$P,Q,R$ 是条件属性，$Y$ 是决策属性。$U/Y=\{\{a1,a2,a3,a4,a5,a6\},\{a7,a8\},\{a9\}\}$；$U/P=\{a1,a2\},\{a3,a4,a5\},\{a6,a7,a8,a9\}\}$；$U/Q=\{\{a1,a2,a3\},\{a7,a8\},\{a4,a5,a6,a9\}\}$；$U/R=\{\{a1\},\{a2\},\{a3\},\{a4\},\{a5,a6,a7,a8\},\{a9\}\}$。由伴随正区域的定义可得：

$$BPOS_{\{P\}}(Y)$$

$$=|POS_{\{P\}}(Y)|+\sum_{i=1}^{h}\frac{|\{P\}_*(Y_i)|}{|U|}\frac{|\{P\}_*(Y_i)|}{|Y_i|}$$

$$=|\{a1,a2,a3,a4,a5\}|$$

$$+\frac{|\{a1,a2,a3,a4,a5\}|}{|\{a1,a2,a3,a4,a5,a6,a7,a8,a9\}|}$$

$$\frac{|\{a1,a2,a3,a4,a5\}|}{|\{a1,a2,a3,a4,a5,a6\}|}=5+(5/9)(5/6)=5.46$$

同理可求得：$BPOS_{\{Q\}}(Y)=5.39$；$BPOS_{\{R\}}(Y)=5.41$。

易得：$POS_{\{P\}}(Y)=\{a1,a2,a3,a4,a5\}$；

$POS_{\{Q\}}(Y)=\{a1,a2,a3,a7,a8\}$；

$POS_{\{R\}}(Y)=\{a1,a2,a3,a4,a9\}$，

故 $|POS_{\{P\}}(Y)|=|POS_{\{Q\}}(Y)|=|POS_{\{R\}}(Y)|=5$。

$K_P=5/6$；$K_Q=1$；$K_R=1$。

若用正区域的属性选择标准，则三个属性的正区域一样大，是没法选择的；若用参数依赖度的属性选择标准，则应该选择属性 $Q$ 或者 $R$，因为它们的参数依赖度均为 1。但属性 $R$ 的参数依赖度并不能很好地反映它的优越性，因为它是由第三类决策值得到的，即它是由一个实例 9 完全属于决策属性类 $Y_3=\{a9\}$ 而得到的。由我们的属性选择标准知，应选择属性 $P$，这是比较合理的。

下面我们先分析一下新的属性选择标准的性质。

**定理 8** 在信息系统 $S=(U,C,D,V,f)$ 中，对 $\forall a\in C$，设 $U/\{a\}=\{A_1,A_2,\cdots,A_m\}$ 表示由条件属性 $a$ 对论域 $U$ 的划分，$U/D=\{D_1,D_2,\cdots,D_h\}$ 表示决策属性集 $D$ 对论域 $U$ 的划分，则有：$\sum_{i=1}^{h}\frac{|\{a\}_*(D_i)|}{|U|}\frac{|\{a\}_*(D_i)|}{|D_i|}\leqslant 1$。

证明：由下近似的定义知：$\{a\}_*(D_i)\subseteq D_i$，另一方面有：$\bigcup_{i=1}^{h}D_i=U\wedge D_i\cap D_j=\emptyset(i\neq j)$，故有：$\sum_{i=1}^{h}\frac{|\{a\}_*(D_i)|}{|U|}\frac{|\{a\}_*(D_i)|}{|D_i|}\leqslant\sum_{i=1}^{h}\frac{|\{a\}_*(D_i)|}{|U|}\leqslant\sum_{i=1}^{h}\frac{|D_i|}{|U|}\leqslant\frac{|U|}{|U|}=1$。

由定理 8 知，基于伴随正区域的属性选择标准是优先考

维普资讯 http://www.cqvip.com

虑一个属性的正区域大小，因为两个属性的正区域大小若不相等，则至少相差 1。在正区域大小相同的情况下，再考虑属性分出来的叶子数目（注意所选属性分出的叶子的决策属性值若相同则合并叶子）和叶子所覆盖的实例数，即所产生的叶子应具有较大的泛化能力。在例 2 中，三种条件属性的正区域大小相同，但 $P$ 只产生一个叶子结点且该叶子结点覆盖 5 个实例；属性 $R$ 产生两个叶子结点，一个叶子结点覆盖 4 个，另一个叶子覆盖 1 个；属性 $Q$ 产生两个叶子结点，一个叶子结点覆盖 3 个，另一个叶子覆盖 2 个。根据新标准，我们选择属性 $P$ 是比较合理的。

下面给出基于伴随正区域的决策树生成算法。

改进的基于正区域的决策树生成算法：

BPOSDT($U,C,D$)

输入：训练实例集 $U$，条件属性集 $C$，决策属性集 $D$

输出：一棵决策树

(1) 生成一个结点 $N$；

(2) 若 $U$ 在决策属性上取值完全相同，则用该决策属性值来标记结点 $N$ 并返回；

(3) 若 $C$ 为空，则用占训练实例集中大多数的决策属性取值的值来标记结点 $N$ 并返回；

//有的记录被不正确的分类，所以这里会出现这种情况；

(4) $R$ 为 $\max\limits_{a\in C} BPOS_{\{a\}}(D)$ 所对应的条件属性；记 $U/\{R\}=\{U_1,U_1,\cdots,U_k\}$，令 $U_i(i=1,2,\cdots,k)$ 对应的相应条件属性值为 $u_i$；

(5) 返回一棵树，根节点标记为 $R$，弧标记为 $u1$，$u2$，…，$um$，弧对应的子树分别为 NRDT($U_1$，$C-\{R\}$，$D$，$\lambda$)，NRDT($U_2$，$C-\{R\}$，$D$，$\lambda$)，…，NRDT($U_k$，$C-\{R\}$，$D$，$\lambda$)；

(6) 在同一条件属性分出的叶子结点中，若有决策属性取值相同的叶子，合并这些叶子，并用各弧标记的析取来标记新叶子的弧。

## 4 实例

我们以下面的决策表来说明新决策树生成算法。

表 1 决策表 1

| | a1 | a2 | a3 | a4 | D |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 | 0 |
| 3 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 1 | 0 |
| 6 | 1 | 2 | 1 | 1 | 0 |
| 7 | 1 | 2 | 1 | 0 | 1 |
| 8 | 2 | 2 | 2 | 0 | 1 |
| 9 | 2 | 3 | 2 | 2 | 1 |
| 10 | 3 | 3 | 2 | 2 | 1 |
| 11 | 3 | 3 | 2 | 0 | 2 |
| 12 | 3 | 4 | 2 | 2 | 1 |
| 13 | 3 | 4 | 3 | 2 | 2 |
| 14 | 3 | 4 | 3 | 2 | 2 |
| 15 | 3 | 4 | 3 | 2 | 2 |

由表 1 易得：

$U/\{a1\}=\{\{1,2,3\},\{4,5,6,7\},\{8,9\},\{10,11,12,13,14,15\}\}$；$U/\{D\}=\{\{1,2,3,4,5,6\},\{7,8,9,10,12\},\{11,13,14,15\}\}$。故有：

$$BPOS_{\{a1\}}(D)$$

$$=|POS_{\{a1\}}(D)|+\sum_{i=1}^{k}\frac{|\{a1\}_*(D_i)|}{|U|}\frac{|\{a1\}_*(D_i)|}{|D_i|}$$

$$=|\{1,2,3\}(\{8,9\}|+\frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}|}+\frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|}+\frac{|\{7,8\}|}{|\{1,2,\cdots,15\}|}\frac{|\{7,8\}|}{|\{7,8,9,10,12\}|}$$

$$=5+\frac{3}{15}\times\frac{3}{6}+\frac{2}{15}\times\frac{2}{5}=5+0.15=5.15$$

同理可得：

$BPOS_{\{a2\}}(D)=5.28$；$BPOS_{\{a3\}}(D)=5.19$；$BPOS_{\{a1\}}(D)=3.15$。

故选择属性 $a2$。

$U/\{a2\}=\{\{1,2,3,4,5\},\{6,7,8\},\{9,10,11\},\{12,13,14,15\}\}$。对于数据集 $\{6,7,8\}$，$\{9,10,11\}$ 和 $\{12,13,14,15\}$，依上述计算过程，选择相应的条件属性，最后得到的决策树如图 1 所示。

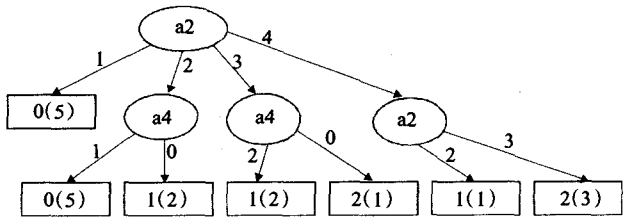


图 1 新算法生成的决策树

若用基于正区域的属性选择标准，则第一次选择的条件属性不一定是 $a2$，这时生成的决策树不仅叶子结点多且叶子的深度也较大。例如若第一次选择条件属性为 $a1$ 或 $a3$，则生成的决策树分别如图 2 和图 3 所示。

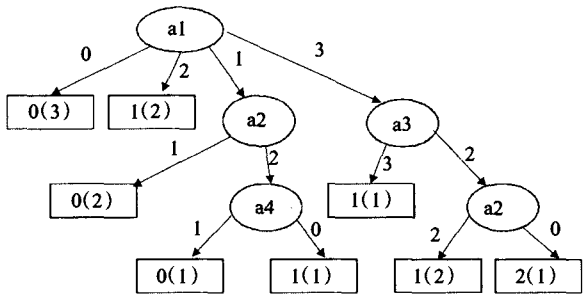


图 2 基于正区域的决策树生成算法（选择属性 $a1$）

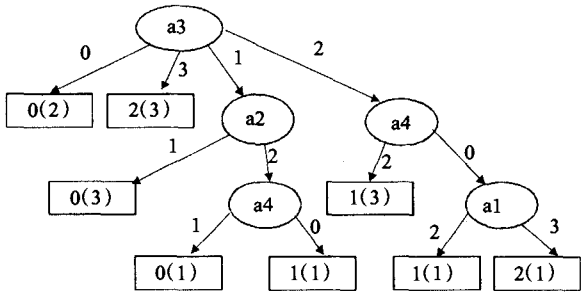


图 3 基于正区域的决策树生成算法（选择属性 $a3$）

**结束语** 本文首先分析了目前基于粗糙集的决策树生成算法中常用的几种属性选择标准的关系，即基于正区域的属性选择标准，基于粗糙边界的属性选择标准和基于依赖度的属性选择标准，并证明了这三种属性选择标准彼此相互等价。然后以正区域的属性选择标准为代表，分析了基于正区域的决策树生成算法的优点和不足。其优点为：得到的决策树通常是叶子结点深度较小，也算是一种较好的属性选择标准。其不足为：(1)在属性选择时，通常正区域相同的属性较多，不易选择属性，若任意选择，则生成的决策树差别可能较大，这

维普资讯 http://www.cqvip.com

说明该属性选择标准太粗略；(2)属性选择时，有可能选择的属性会产生很多叶子结点，这样，有些叶子结点上包含的实例数太少，以至这些叶子结点的泛化能力太小而没有意义，同时在预测时是比较费时。针对这些不足，我们给出了一种新的属性选择标准，即基于伴随正区域的属性选择标准。在新的属性选择标准中，有如下优点：(1) 继承了基于正区域的属性选择标准的优点，即优先考虑一个属性的正区域大小，这样能保证生成的决策树中，叶子结点深度较小。(2)在正区域大小相同的情况下，再考虑属性分出来的叶子数目(注意所选属性分出的叶子的决策属性值若相同则合并叶子)和叶子所覆盖的实例数，这样能保证生成的决策树的叶子数目较小，且生成的叶子具有较大的泛化能力。因而用新的属性选择标准生成的决策树一般具有叶子数目较少，叶子的平均深度也较小，且叶子具有较强的和泛化能力。最后，用一实例说明了新的属性选择标准的优越性。

## 参考文献


[1] Quinlan J R. Induction of decision trees. Machine Learning, 1986,1:81-106

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

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

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

[5] Pawlak Z. Rough sets. International Journal of Computer and Information Sciences, 1982,11:341-356

[6] 王国胤. Rough 集理论与知识获取. 西安：西安交通大学出版社，2001

[7] 张文修，吴伟志，等. 粗糙集理论与方法[M]. 北京：科学出版社，2003

[8] Wang Guoyin. Rough set theory and knowledge acquisition[M]. Xi An: Xi An JiaoTong University Press, 2001

[9] Zhang Wenxiu, WuWeizhi. Rough set theory and method[M]. Beijing: Science press, 2003

[10] 王名扬，卫金茂，伊卫国. 基于粗集理论的新决策树剪枝方法[J]. 东北师范大学(自然科学版)，2005, 37(3): 28-32

[11] 朱红. 基于 Rough Set 的最简决策树确定算法的研究[J]. 计算机工程与应用，2003,41(13):129-131

[12] 赵翔，向一丹，刘同明，祁云嵩. 一种基于粗糙集的决策树生成算法[J]. 华东船舶工业学院学报(自然科学版)，2005,19(4): 73-76

[13] 管红波，田大纲. 基于属性重要性的决策树规则提取算法[J]. 系统工程与电子技术，2004,26(3):334-337

[14] 常志玲，周庆敏，杨清莲. 基于粗糙集理论的决策树构造算法[J]. 南京工业大学学报，2005,27(4):80-83

[15] Jin MaoWei. Rough set based approach to selection of node [J]. International Journal of Computational Cognition, 2003,1(2):25-40

[16] Wang Mingyang, Jin Maowei, Yi Weiguo. A new decision tree pruning method based on RST [J]. Journal of Northeast Normal University (Natural Science Edition), 2005, 37(3): 28-32

[17] Zhu Hong. The Research of the simplest decision tree production algorithm based on rough set [J]. Journal of Computer Engineering and Application, 2003,41(13):129-131

[18] Zhao Xiang, Xiang Yidan, Liu Tongming, et al. An algorithm for decision tree construction based in rough sets [J]. Journal of East China Shipbuilding Institute (Natural Science Edition), 2005,19(4): 73-76

[19] Guan Hongbo, Tian Dagang. Rule abstracting algoritm by decisive tree based on the importance of attribute [J]. Systems Engineering and Electronics, 2004,26(3): 334-337

[20] Chang Zhiling, Zhou Qingmin, Yang Qinglian. A method based on rough set to construct decision tree [J]. Journal of Najing University of Technology, 2005,27(4):80-83


(上接第 107 页)

类型为 CBR，大小为 10b，大探测包为 10 组间隔很小的 CBR，大小为 2000b，每个 CBR 包被分为 10 个 200b 的 UDP 包。背景流量为 4 组指数达到数据流，分别从节点 $S_1$ 到 $D_1$，$D_2$，节点 $S_2$ 到 $D_1$，$D_2$。我们进行两组实验：第一组发送速度为 400kb/s，第二组发送速度为 800kb/s。第一组背景流量较小，通过 Ω 来判断拓扑结构；第二组背景流量较大，采用 Ψ 来判断拓扑结构。

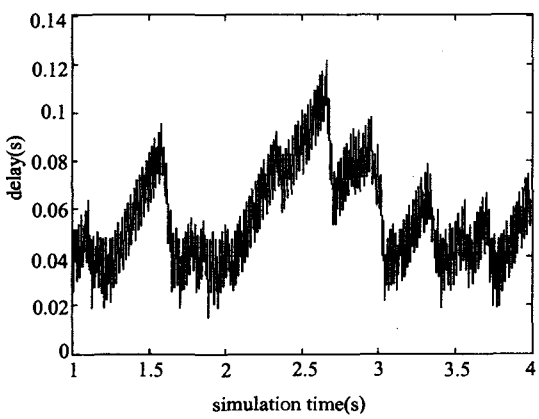


图 5 具有共享链路的端到端延时的相互影响示意图

图 5 是第一组实验中具有共享链路的端到端延时的相互影响示意图。从图中可以看出，两组小数据包的端到端延时在受另一节点发送的大数据包的影响的情况下，具有震荡现象。

实验结果如表 2 所示。在第一组实验中，若两条路径具有共享链路，延时变化大于 20ms，否则小于 2ms，相差大于 10 倍，可选取门限为 10ms。在第二组实验中，若两条路径具有共享链路，丢包率大于 17%，否则小于 3%，相差大于 5 倍，可选取门限为 10%。仿真推断的逻辑拓扑与真实拓扑的设置相符，说明了本文采用探测方法的有效性。

表 2 仿真结果表

| 拓扑 | $P_{S1D1}$ 与 $P_{S2D2}$ | | | | $P_{S1D2}$ 与 $P_{S2D1}$ | | | |
|---|---|---|---|---|---|---|---|---|
| 结构 | 1 | | 2 | | 1 | | 2 | |
| | Ω | Ψ | Ω | Ψ | Ω | Ψ | Ω | Ψ |
| a | 23.2ms | 3.0% | 35.1ms | 22% | 24.54ms | 3.2% | 26.9ms | 27% |
| b | 1.4ms | 0.1% | 5.4ms | 2.7% | 21.8ms | 2.8% | 25.6ms | 17.6% |
| c | 20.2ms | 2.7% | 27.2ms | 18% | 0.8ms | 0% | 2.9ms | 1.7% |
| d | 0.7ms | 0.2% | 1.4ms | 2.3% | 1.3ms | 0.5% | 2.7ms | 3.0% |

**结束语** 多源 NT 技术比单源拓扑推断可以获取更多的拓扑信息。本文通过分析四种 2-by-2 结构特征，发现四种 2-by-2 结构的拓扑推断问题可以转化为判断路径是否有共享链路的问题，而具有共享链路的路径的数据流对彼此的性能具有一定的影响。基于此点，提出一种新的三明治的探测方法，并详细研究了探测包的构造方式。通过仿真验证了探测方法的有效性。本文提出的方法对多源 NT 拓扑推断技术研究具有重要的意义。

## 参考文献


[1] Castro R, Coates M, Liang G, et al. Internet tomography: Recent developments. Statistical Science, 2004,52(3):499-517

[2] Coates M, Nowak R. Sequential monte carlo inference of internal delays in nonstationary communication networks. IEEE Transactions on Signal Processing, 2002,50(2): 366-376

[3] Coates M, Rabbat M, Nowak R. Merging logical topologies using end-to-end measurements// Proc. ACM SIGCOMM Conf. on Internet Measurement. Miami, FL, Oct. 2003

[4] Rabbat M, Nowak R, Coates M. Multiple source, multiple destination network tomography// Proc IEEE Infocom. Hong Kong, Mar. 2004

[5] Rabbat M, Coates M, Nowak R. Multiple Source Internet Tomography. IEEE Journal on Selected Areas in Communications, 2006,24(12):2221-2234