连续属性空间上的规则学习算法*

权光日[1] 刘文远[2] 叶 风[2] 陈晓鹏[1]

[1](哈尔滨工业大学威海分校 威海 264200)

[2](哈尔滨工业大学计算机科学与工程系 哈尔滨 150001)

摘要 文章研究连续属性空间上的规则学习算法.首先简述了研究连续属性空间上的规则学习算法的目的和意义,并将规则学习理论中的一些基本概念推广到连续属性空间.在此基础上,研究了连续属性空间离散化问题,证明了属性空间最小离散化问题是NP困难问题,并将信息熵函数与无穷范数的概念应用到连续属性离散化问题,提出了基于信息熵的属性空间极小化算法.最后,提出了连续属性空间上的规则学习算法,并给出了数值实验结果.

关键词 规则学习算法,连续属性空间,信息熵,无穷范数,NP困难问题.

中图法分类号 TP181

样本空间化简问题是人工智能领域中的重要研究课题.示例学习算法通过样本例子的训练产生识别函数.示例学习系统的学习速度与精度以及识别速度不仅依赖于所采用的学习算法本身,而且与训练样本集合的规模与样本空间的描述密切相关.对实际采样的数据进行筛选和改用适当的描述是提高学习速度与识别精度以及节省存储空间的重要手段.另外,样本例子的筛选与属性空间的优化描述还可以提高识别系统对输入例子噪音的鲁棒性.

1987年,Quinlan在研究连续属性空间上的决策树学习算法时,提出了基于信息熵的属性分割算法[1,2].Fayyad等人在Quinlan工作的基础上提出了能够加快离散化速度的改进算法[2~4].本文在前人工作的基础上,将信息熵函数与无穷范数的概念应用到连续属性离散化问题,提出了基于信息熵的属性空间极小化算法.其目的在于在区间分割过程中防止过分细化,以便提高学习系统的聚类能力以及识别系统对输入例子噪音的鲁棒性.在此基础上,提出了连续属性空间上的规则学习算法.

1 基本概念

设连续闭区间 $D_j=[d_j^0,d_j^1]$,$j=1,2,\ldots,n$,为第 j 个属性 x_j 的值域(取值范围),$E=D_1\times D_2\times\ldots\times D_n$ 是 n 维无穷向量空间,E 中的元素 $e=(v_1,v_2,\ldots,v_n)$ 叫做例子,其中 $v_i\in D_j$.设 PE 和 NE 是 E 的两个子集,为区别起见,分别称为正例集和反例集.

定义1. D_j 的子集 A_j 称为 D_j 的有限个区间的并,当且仅当 A_j 为有限个区间(闭区间、开区间或半开半闭区间)的并,简称区间并.

显然,如果 A_j,B_j 为 D_j 的区间并,则 $A_j\cup B_j$,$A_j\cap B_j$ 和 D_j-A_j 皆是 D_j 的区间并.

定义2. 选择子是形为$\{x_j=A_j\}$或$\{x_j\neq B_j\}$的关系语句,其中 A_j,B_j 为满足 $A_j\subseteq D_j,B_j\subseteq D_j$ 的区间并,并且规定$\{x_j\neq B_j\}=\{x_j=D_j-B_j\}$.公式是1个选择子或几个选择子的合取式,记为 $L=\bigwedge_{j\in J}\{x_j=A_j\}$,其中 $J\subseteq N$,$N=\{1,2,\ldots,n\}$.注意:(1) 对于公式 L 中不出现的属性,规定它取值为该属性的值域,即任何 $j\in N$,如果 $j\notin J$,

---

* 本文研究得到国家863高科技项目基金和煤炭科学基金资助.作者**权光日**,1962年生,博士,副教授,主要研究领域为神经网络,机器学习.**刘文远**,1968年生,博士生,讲师,主要研究领域为数据库理论,机器学习.叶风,1960年生,工程师,主要研究领域为人工智能逻辑基础,机器学习.**陈晓鹏**,1970年生,讲师,主要研究领域为数据库理论.

本文通讯联系人:权光日,威海 264200,山东省威海市哈尔滨工业大学威海分校

本文1998-05-04收到原稿,1998-11-25收到修改稿

则等价于在 L 上逻辑乘选择子$\{x_j=D_j\}$;(2) 例子可以看做是公式的特殊情况统一处理,即 $e=(v_1,v_2,\ldots,v_n)=\bigwedge_{j\in N}\{x_j=\{v_j\}\}$. 选择子 $S=\{x_j=A_j\}$覆盖一个公式 $L=\bigwedge_{j\in J}\{x_j=A'_j\}$当且仅当 $j\in J$ 并且 $A'_j\subseteq A_j$. 已知公式 $L=\bigwedge_{j\in J}\{x_j=A_j\}$及 $L'=\bigwedge_{j\in J'}\{x_j=A'_j\}$,$L$ 覆盖 L'当且仅当 $J\supseteq J'$并且对任何 $j,j\in J,A'_j\subseteq A_j$.

定义 3. 一个公式称为(对已知例子集 $PE\cup NE$)一致的,如果它不覆盖反例集 NE 中的任何反例. 一个规则称为一致的,如果它的每一个公式都是一致的. 一个规则称为完备的,如果任何一个 PE 中的正例子都被它覆盖,即它的某一公式覆盖. 如果一个规则是一致的又是完备的,则简称一致完备的.

定义 4. 已知反例矩阵 NE 和一个公式 $L=\bigwedge_{j\in J}\{x_j=A_j\}$. 对 NE 的每一列 $j\in N$,如果 $j\notin J$,则用死元素"*"对 NE 中第 j 列的所有元素作代换;如果 $j\in J$,则用"*"对 NE 中第 j 列属于 A_j 的所有元素作代换. 这样得到的矩阵叫做公式 L 的扩张矩阵,记为 $EM(L)$. 在扩张矩阵中,分别来自不同行的非死元素组成的集合叫做 $EM(L)$的一条路.

易见,当定义 4 中的公式用正例 e^+代替时,就得到正例 e^+的扩张矩阵 $EM(e^+)$.

定义 5. 设 $EM(L)$是一致公式 L 的扩张矩阵,如果在 $EM(L)$中的某一行中只有一个非死元素,则称该元素为必选元素.

由于 $EM(L)$的必选元素一定属于所有覆盖 L 的公式中的对应选择子,因此有着重要的作用.

定义 6. 已知公式 $L=\bigwedge_{j\in J}\{x_j=A_j\}$及公式 $F=\bigwedge_{j\in J'}\{x_j=B_j\}$,则将 L 和 F 对应的选择子的取值合并得到一个新的公式,称为 L 和 F 的合并,记为 $L\oplus F$. 即 $L\oplus F=\bigwedge_{j\in N}\{x_j=A_j\cup B_j\}=\bigwedge_{j\in J\cap J'}\{x_j=A_j\cup B_j\}$. 注意,对任何 $j,j\in N$,但 $j\notin J\cap J'$,$\{x_j=A_j\cup B_j\}=\{x_j=D_j\}$省略不写.

定理 1. 设 $EM(L)$是一致公式 L 的扩张矩阵,则存在一个从 $EM(L)$中的路到覆盖公式 L 的已知公式的映射. 如果公式 L 覆盖公式 L',则 $EM(L)$中的一条路必定也是 $EM(L')$中的一条路.

定理 2. 公式 L 既覆盖公式 A 又覆盖公式 B,当且仅当它覆盖 $A\oplus B$.

文献[5~9]中对扩张矩阵理论进行了深入的研究,给出了扩张矩阵的重要性质,结果如下.

连续属性空间上的选择子与离散有限属性空间上的选择子的形式描述是一致的,但具体表达式是有所区别的. 例如,$A_1=\{x_1\neq 1,4\}$表示离散空间上的一个选择子,而 $B_1=\{x_1\neq[0,1)\cup(2,3)\cup[6,8]\}$表示连续空间上的一个选择子.

定义 7. 设 $EM(L)$是一致公式 $C=\bigwedge_{j\in J}\{x_j=A_j\}$的扩张矩阵,$path=\bigwedge_{j\in J_P}\{x_j=P_j\}$是 $EM(L)$中的一条路,按下列准则生成的公式 L'叫做路 $path$ 在公式 C 的约束下生成的公式,并记为 $L'=L(path,C)$. 其中 A_i 是区间并,P_j 是离散的有限子集.

(1) 对于任意 $j\in J_P$,在 D_j-A_j 中当且仅当选定含有 P_j 中元素的子区间时,所得区间的并,记做 B_j.

(2) 生成公式 $L'=L(path,C)=\bigwedge_{j\in J_P}\{x_j=B_j\}$.

2 属性空间化简问题的启发式算法

2.1 属性空间最小化问题是 NP 困难问题

设 $E=D_1\times D_2\times\ldots\times D_n$ 是 n 维无穷向量空间,即第 i 个属性 x_i 的值域(取值范围)$D_i=[d_i^0,d_i^1](i=1,2,\ldots,n)$是连续实数闭区间,$E$ 中的元素 $e=(v_1,v_2,\ldots,v_n)$叫做例子,其中 $v_i\in D_i$. 设 PE 和 NE 是 E 的两个子集,为区别起见,分别称为正例集和反例集.

定义 8. PE,NE 分别表示正例集与反例集,

$$d(PE,NE)=\min_{e^+\in PE,e^-\in NE}\{\max_{1\leqslant i\leqslant N}\{|v_i^+-v_i^-|\}\},$$

称 $d(PE,NE)$为正例集与反例集在无穷范数下的距离,简称正例集与反例集的距离. 其中 $e^+=(v_1^+,v_2^+,\ldots,v_n^+)\in PE,e^-=(v_1^-,v_2^-,\ldots,v_n^-)\in NE$ 为任意反例子与正例子.

显然,如果 $d(PE,NE)=0$,则存在一个正例子与反例子 $e_i^+=e_r^-$,即 $PE\cap NE\neq\varnothing$. 当正例集 PE 与反例集 NE 满足 $d(PE,NE)=0$ 时,称正例集与反例集是相交的或不可分离的.

为了讨论方便起见,我们规定子区间是指最小的子区间,即不含分割点的子区间.

定义 9. 子区间的乘积空间称为极小乘积空间,简称极小空间.

定义 10. 当一个属性空间进行划分后,每一个极小乘积空间都不同时含有正例子与反例子,则称此划分为可分离的划分;否则,称为不可分离的划分(存在一个极小乘积空间,它同时包含正例子与反例子).

定义 11. X 是 $E=D_1\times D_2\times\ldots\times D_n$ 上的例子集合,L_i 是属于 D_i 的一个区间,l_i^0,l_i^1 分别表示 L_i 的左右边界,

$$\mathrm{Max}(L_i,X)=\mathrm{Max}\{v_i\mid\forall\ e=(v_1,v_2,\ldots,v_n)\in X,v_i\leqslant l_i^1\},$$

$$\mathrm{Min}(L_i,X)=\mathrm{Min}\{v_i\mid\forall\ e=(v_1,v_2,\ldots,v_n)\in X,v_i\geqslant l_i^0\},$$

分别叫做例子集 X 在属性区间 L_i 上的上确界与下确界.

定义 12. 对于每一个属性值区间 $D_j=[d_j^0,d_j^1](j=1,2,\ldots,n)$,进行 $k_j(j=1,2,\ldots,n)$ 等分 $d_j^0=l_0<l_1<\ldots<l_{k_j}=d_j^1$,$I_j^i$ 表示第 j 个属性的第 i 个区间(即 $I_j^i=[l_{i-1},l_i)$,当 $i=n$ 时,$I_j^i=[l_{n-1},l_n]$),间距 $d_j=(d_j^1-d_j^0)/k_j$,这样得到的属性空间划分称为等距划分,记做 $K(k_1,k_2,\ldots,k_n)$ 或简记为 K.

定理 3. 如果 $d(PE,NE)=d>0$,则等距划分 $K(k_1,k_2,\ldots,k_n)$ 能够分离出正例集与反例集.其中 $k_j=[|D_j|/d]+1(j=1,2,\ldots,n)$.证明略.

子空间数最小的属性区间分割问题:设 $E=D_1\times D_2\times\ldots\times D_n$ 是 n 维无穷向量空间,即 $D_i=[d_i^0,d_i^1](i=1,2,\ldots,n)$ 是连续实数闭区间,PE,NE 分别表示给定的正例集与反例集,$k_i(i=1,2,\ldots,n)$ 表示区间分割后 D_i 区间所包含的子区间的个数,在能够分离正例集与反例集 PE,NE 的属性区间划分中,求出使 $k_1\cdot k_2\cdot\ldots\cdot k_n$ 最小的属性区间划分.

分割点数最小的属性区间分割问题:设 $E=D_1\times D_2\times\ldots\times D_n$ 是 n 维无穷向量空间,即 $D_i=[d_i^0,d_i^1](i=1,2,\ldots,n)$ 是连续实数闭区间,PE,NE 分别表示给定的正例集与反例集,$k_i(i=1,2,\ldots,n)$ 表示区间分割结束后属性区间 D_i 所包含的分割点的个数,在能够分离正例集与反例集 PE,NE 的属性区间划分中,求出使 $k_1+k_2+\ldots+k_n$ 最小的属性区间划分.

定理 4. 子空间数最小的属性区间分割问题和分割点数最小的属性区间分割问题都是 NP 难题.

为了证明上述定理 4,先简单地介绍文献[10,11]中的结果(为了方便起见,在描述方面,对文献[10,11]中的结果进行了适当的改变).

设 $E^*=D_1^*\times D_2^*\times\ldots\times D_n^*$ 是二值离散的向量空间,即 $D_i^*=\{0,1\}(i=1,2,\ldots,n)$,$PE$,$NE$ 分别是任意给定属性的正例集与反例集,并且 $PE\cap NE=\varnothing$.

定义 13. $E_i^*=D_{i_1}^*\times D_{i_2}^*\times\ldots\times D_{i_k}^*$ 是部分属性生成的乘积空间,如果存在正例子 $e^+=(v_1^+,v_2^+,\ldots,v_n^+)\in PE$ 和反例子 $e^-=(v_1^-,v_2^-,\ldots,v_n^-)\in NE$ 满足 $v_{i_1}^+=v_{i_1}^-,v_{i_2}^+=v_{i_2}^-,\ldots,v_{i_k}^+=v_{i_k}^-$,则称正例集 PE 和反例集 NE 在乘积空间 E_i^* 上是不可分离的,否则,称正例集 PE 和反例集 NE 在乘积空间 E_i^* 上是可分离的.

最优属性选择问题:在部分属性生成的乘积空间 $E_i^*=D_{i_1}^*\times D_{i_2}^*\times\ldots\times D_{i_k}^*$ 中,求出能够分离正例集 PE 和反例集 NE 并且维数最小的部分属性生成的乘积空间 E_i^*.

引理. 最优属性选择问题是 NP 难题.引理的证明参看文献[10,11].

定理 4 的证明:设 $E=D_1\times D_2\times\ldots\times D_n$ 是连续的属性空间,其中 $D_i=[0,1](i=1,2,\ldots,n)$;$PE$ 和 NE 是 E 上任意给定的满足下列两个条件的正例集与反例集:(1) $PE\cap NE=\varnothing$;(2) 任意正例子 $e^+=(v_1^+,v_2^+,\ldots,v_n^+)\in PE$ 和反例子 $e^-=(v_1^-,v_2^-,\ldots,v_n^-)\in NE$ 都有 $v_i^+,v_i^-\in\{0,1\}(i=1,2,\ldots,n)$.不难看出,对于上述方式特殊选定的连续属性空间与两个条件约束下的任意正例集与反例集,子空间数最小的属性区间分割问题和分割点数最小的属性区间分割问题都等价于最优属性选择问题,所以它们都不宜解决最优属性选择问题.根据上述引理,最优属性选择问题是 NP 困难问题.因此,子空间数最小的属性区间分割问题和分割点数最小的属性区间分割问题都是 NP 难题.

2.2 基于信息熵的属性空间分割算法

2.2.1 属性区间分割算法的启发式策略准则

(1) 选择信息增益最大的划分点进行分割,加快分割过程,减少分割点的数目.

(2) 如果一个子区间的长度小于正例集与反例集之间的距离 $d(PE,NE)$,则不再进行该区间分割.这一准则能够防止区间分割过分细微,使学习后的识别系统对于输入例子具有良好的鲁棒性.

2.2.2 基于信息熵的属性区间分割算法

第1步. PE,NE 分别表示正例集与反例集,$L_i(i=1,2,\ldots,n)$表示属性区间 $D_i=[d_i^1,d_i^0]$的子区间,$P(L)$,$N(L)$分别表示属于 $L=L_1\times L_2\times\ldots\times L_n$ 的正例子与反例子的集合,$S_i(i=1,2,\ldots,n)$是 D_i 的子区间为元素的集合,l_i^0,l_i^1 分别表示 L_i 的左右边界,l_i^{0*},l_i^{1*} 分别表示例子集 $P(L)\cup N(L)$在 L_i 上的上确界与下确界;初始化 $P(L)=PE,N(L)=NE,L_i=D_i,S_i=\varnothing(i=1,2,\ldots,n)$.

第2步. 如果 $P(L)=\varnothing$或者 $N(L)=\varnothing$或者所有$(l_i^{1*}-l_i^{0*})<d(PE,NE)(i=1,2,\ldots,n)$,则 $PE=PE-P(L)$,$NE=NE-N(L)$,转第8步;否则,对于每个满足$(l_i^{1*}-l_i^{0*})\geqslant d(PE,NE)$的属性区间 $L_j=[l_j^0,l_j^1]$进行划分,$A_j=(l_j^0,(l_j^{1*}-l_j^{0*})/2]$,$B_j=((l_j^{1*}-l_j^{0*})/2,l_j^1]$,$j\in J$(注:$J$ 表示满足$(l_j^{1*}-l_j^{0*})\geqslant d(PE,NE)$的 j 的集合;L_j为左闭区间时 A_j 也是);

第3步. 计算每个划分的信息熵 $E(L_j)$,$j\in J$.

$$I(A_j)=-\left(\frac{p_{A_j}}{p_{A_j}+n_{A_j}}\log\frac{p_{A_j}}{p_{A_j}+n_{A_j}}+\frac{n_{A_j}}{p_{A_j}+n_{A_j}}\log\frac{n_{A_j}}{p_{A_j}+n_{A_j}}\right),$$

$$I(B_j)=-\left(\frac{p_{B_j}}{p_{B_j}+n_{B_j}}\log\frac{p_{B_j}}{p_{B_j}+n_{B_j}}+\frac{n_{B_j}}{p_{B_j}+n_{B_j}}\log\frac{n_{B_j}}{p_{B_j}+n_{B_j}}\right),$$

$$E(L_j)=\frac{|A_j|}{P_L+N_L}I(A_j)+\frac{|B_j|}{P_L+N_L}I(B_j).$$

其中 p_{A_j},n_{A_j}分别表示 A_j 中的正例子数与反例子数;p_{B_j},n_{B_j}表示 B_j 中的正例子数与反例子数;P_L,N_L 表示属于 $L=L_1\times L_2\times\ldots\times L_n$ 的正例子数与反例子数.定义 $I(\varnothing)=0$.

第4步. 在所有划分中选择信息增益最大的划分,即选择信息熵最小的划分.最小的划分对应的属性记为 r.

第5步. 如果 $I(A_r)=0$ 且 $I(B_r)>0$,则 $P(L)=P(L)-P(A_r)$(或者 $N(L)=N(L)-N(A_r)$),$PE=PE-P(A_r)$(或者 $NE=NE-N(A_r)$),$S_r=S_r-\{L_r\}+\{A_r\}+\{B_r\}$,$L_r=B_r$;返回第2步(注意:$I(A_r)=0$,所以,$A_r$ 只包含 L 中的正例子或者只包含 L 中的反例子或者为空;$P(A_r)$,$N(A_r)$分别表示 L 中属于 A_r 的正例子与反例子的集合).

第6步. 如果 $I(A_r)>0$ 且 $I(B_r)=0$,则 $P(L)=P(L)-P(B_r)$(或者 $N(L)=N(L)-N(B_r)$),$PE=PE-P(B_r)$(或者 $NE=NE-N(B_r)$),$S_r=S_r-\{L_r\}+\{A_r\}+\{B_r\}$,$L_r=A_r$;返回第2步.

第7步. 如果 $I(A_r)>0$ 且 $I(B_r)>0$,则 $P(L)=P(A_r)$,$N(L)=N(A_r)$,$L_r=A_r$,$S_r=S_r-\{L_r\}+\{A_r\}+\{B_r\}$,返回第2步.

第8步. 如果 $I(A_r)=0$ 且 $I(B_r)=0$,则 $PE=PE-P(A_r)$,$NE=NE-N(A_r)$,$S_r=S_r-\{L_r\}-\{A_r\}+\{B_r\}$,转第9步.

第9步. 如果 $PE\neq\varnothing$,任意选择 $e^-\in PE$,选择包含 e^+的最小乘积空间 $L=L_1\times L_2\times\ldots\times L_n$,其中 $L_i\in S_i$,并且 $e^+\in L_i(i=1,2,\ldots,n)$,计算 $P(L)$和 $N(L)$,返回第2步;否则(即 $PE=\varnothing$),输出 $S_i(i=1,2,\ldots,n)$,停机.

定理 6. 如果区间分割程序停机,则每一个属性子区间被分割成互不相交的子区间.证明略.

定理 7. 如果 $d(PE,NE)=d>0$,则区间分割算法在不超过 $\sum_{j=1}^{n}\left[2\frac{|D_j|}{d}\right]+n$ 次的区间分割内能够分离出正例集与反例集.证明略.

定理 8. 如果极小空间 $L=L_1\times L_2\times\ldots\times L_n$ 包含一些正例子,则 $C=\bigwedge_{j\in J}(x_j=L_j)$,$J=\{1,2,\ldots,n\}$是一致公式,即公式 C 在反例集 EN 背景下的扩张矩阵有一条路.证明略.

定义 14. X 是 E 的例子集,K 表示某一区间分割算法,I_j^i 表示区间分割算法 K 得到的第 j 个属性的第 i 个子区间,$|X\cap I_j^i|$表示在属于 X 的例子中,第 j 个属性值属于 I_j^i 的例子数目.

$$f_j(t,X)=\begin{cases}|X\cap I_j^1|, & t\in I_j^1\\ |X\cap I_j^2|, & t\in I_j^2\\ \cdots\quad\cdots & \\ |X\cap I_j^{k_j}|, & t\in I_j^{k_j}\end{cases}\quad(j=1,2,\ldots,n),$$

称 $f_j(t,X)$ 为例子集 X 在第 j 个属性区间 D_j 上的分布函数,简称 X 的第 j 个分布函数.

定义 15. $D_j(j=1,2,\ldots,n)$ 的互不相交的并集 $D=D_1\cup D_2\cup\ldots\cup D_n$ 上定义的函数

$$F(t,X)=\begin{cases}f_1(t,X), & t\in D_1\\ f_2(t,X), & t\in D_2\\ \cdots\quad\cdots & \\ f_n(t,X), & t\in D_n\end{cases},$$

称 $F(t,X)$ 为例子集 X 在 D 上的分布函数,简称 X 的分布函数.

定义 16. X 是给定的例子集,I_j^i 表示区间分割算法得到的第 j 个属性的第 i 个子区间,L_j 表示第 j 个属性的子区间中至少包含一个 X 中例子的子区间的并,即 $L_j=\bigcup_i I_j^i\{I_j^i\mid I_j^i\cap X\neq\varnothing\}$,称 $L(X)=L_1\times L_2\times\ldots\times L_n$ 为包含 X 的最小子空间.

定义 17. PE 和 NE 是给定的正例集和反例集,PE' 表示正例集 PE 的子集,$L(PE')=L_1\times L_2\times\ldots\times L_n$ 表示包含 PE' 的最小子空间.对于 NE 的第 $j(j=1,2,\ldots,n)$ 列中所有属于 L_j 的元素都用"*"代替,而得到的矩阵叫做正例子集 PE' 在反例集 NE 背景下的扩张矩阵,记做 $EM(PE')$.

2.3 生成公式的算法

第 1 步. $E=D_1\times D_2\times\ldots\times D_n$ 是连续属性空间,PE 和 NE 是给定的正例集和反例集,I_j^i 表示区间分割算法得到的第 j 个属性的第 i 个子区间,PE' 表示正例集 PE 的子集.SN 表示所有只包含反例子的子区间组成的集合,即只包含扩张矩阵中非死元素的子区间组成的集合.

$$S_N=\{I_j^i\mid I_j^i\cap PE'\text{ 且 }I_j^i\cap NE\neq\varnothing(j=1,2,\ldots,n;i=1,2,\ldots,k)\}.$$

第 2 步. 建立 PE' 在反例集 NE 背景下的扩张矩阵 $EM(PE')$.

第 3 步. $S-\varnothing$,求 S_N.

第 4 步. 如果 S_N 中有相邻的子区间,则合并成一个子区间,直到没有相邻的子区间为止.

第 5 步. 如果扩张矩阵中有必选元素(扩张矩阵的某一行只有一个非死元素),则在 S_N 中选择包含该反例子的子区间 I_j^i,否则,在 S_N 中选择包含反例子数最多的的子区间 I_j^i;NE 中删去属于 I_j^i 的反例子,$S=S+\{I_j^i\}$.

第 6 步. 如果反例集不空,则返回第 5 步;反例集空,转第 7 步.

第 7 步. 生成公式 $L=\bigwedge_{j\in J}\{x_j-D_j-A_j\}$,其中 $A_j=\bigcup I_j^i\{I_j^i\in S\}$;停机.

3 连续属性空间上的规则学习算法

3.1 连续规则学习算法 CRA

下面给出连续属性空间上的规则学习算法(简称连续规则学习算法,记做 CRA).

第 1 步. 对于连续属性空间 $E=D_1\times D_2\times\ldots\times D_n$,根据已知正例集 PE 和反例集 NE,用区间分割算法进行属性区间划分.

$PE=PE'$ 表示正例集,$NE=NE'$ 表示反例集,EM 表示扩张矩阵,$F(t,PE)$,$F(t,NE)$ 分别表示正例集和反例集的分布函数,I_j^i 表示区间分割算法得到的第 j 个属性的第 i 个子区间,$|X\cap I_j^i|$ 表示在属于 X 的例子中,第 j 个属性值属于 I_j^i 的例子数目;

第 2 步. 建立正例集和反例集的分布函数 $F(t,PE')$ 和 $F(t,NE')$;

第 3 步. 所有 I_j^i 中选择使 $(|PE'|-F(t,PE'))F(t,NE')$ 最大的 i 和 j(其中 $t\in I_j^i$);

第 4 步. 在 PE' 和 NE' 中删除第 j 个特征值属于 I_j^i 的例子,删除子区间 I_j^i,调整正例集和反例集的分布函数 $F(t,PE')$ 和 $F(t,NE')$;

第5步. 如果 NE' 不空,则返回第3步;否则,调用生成公式算法;

第6步. 如果 PE 空,停机;否则,$PE-PE-PE'$,$NE'=NE$,$PE'=PE$,并返回第2步.

定理 9. 如果正例集与反例集是可分离的,即 $d(PE,NE)-d>0$,则连续规则学习算法 CRA 在有限步内停机,并生成一致完备规则. 证明略.

3.2 学习算法的时间复杂性分析

P 和 N 分别表示正例集与反例集的个数,$|D_j|$ 表示第 $j(j=1,2,\ldots,n)$ 个属性区间 D_j 的长度,$d=d(PE,NE)$ 是正例集与反例集之间的距离.

区间分割算法的计算时间复杂度为

$$O(nPN)+O\left(n(P+N)\left(\sum_{j=1}^{n}\left[2\frac{|D_j|}{d}\right]+n+\min\{P,N\}\right)\right)=O(nPN)+O\left(n(P+N)\left(\sum_{j=1}^{n}\left[2\frac{|D_j|}{d}\right]+n\right)\right).$$

生成公式算法被 CRA 算法调用的计算时间复杂度为

$$O\left((PN)\left(\sum_{j=1}^{n}\left[2\frac{|D_j|}{d}\right]+n\right)\right).$$

由此可见,使正例集 PE' 空为止,CRA 算法的计算时间复杂度为

$$T(CRA)=O\left((n+P)(P+N)\left(\sum_{j=1}^{n}\left[2\frac{|D_j|}{d}\right]+n\right)\right).$$

4 故障诊断中的应用

表1是文献[12]中给出的汽轮机故障诊断方面的实例. 本文以它为例说明基于信息熵的属性空间分割算法在连续属性空间中进行空间分割的过程,说明连续规则学习算法 CRA 在连续属性空间中生成识别规则的过程(即学习过程).

表1 汽轮机故障诊断实例

	Symptom											Fault
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	D
1	0.8	0	0.1	0.1	1	0	1	0.1	0.9	1	0	1
2	0.8	0	0.1	0.1	0.8	0	0.8	0.1	0.8	0.9	0	1
3	0.5	0	0.1	0.1	0.8	0	1	0.1	0.7	0.9	0	1
4	0.8	0	0.2	0.2	1	0	1	0.1	0.9	1	0	1
5	0.5	0	0.1	0.1	0.8	0	0.9	0.1	0.7	0.9	0	0
6	0.5	0.9	0	0.8	0	0.8	0.1	0.9	0.5	0.1	0.8	0
7	0.5	0.9	0	0.8	0	0.8	0.1	0.5	0.5	0.2	0.9	0
8	0.6	0.7	0	0.9	0	0.5	0.1	0.8	0.4	0.1	0.7	0
9	0.4	0.7	0	0.7	0	0.5	0.1	0.7	0.3	0.1	0.6	1
10	0.3	0.9	0	0.9	0	1	0	0.8	0.1	0	0.9	0
11	0.2	0.7	0	0.8	0	1	0	0.8	0	0	0.8	0
12	0.2	0.6	0	0.6	0	0.9	0	0.7	0	0	0.6	0
13	0.4	0.4	0.3	0.6	0	0	0.1	0.1	0.2	0	0.1	0
14	0.4	0.5	0.3	0.7	0.05	0	0	0.1	0.1	0	0.1	0
15	0.4	0.6	0.4	0.9	0	0.8	0	0.3	0.1	0	0.9	0
16	0.3	0.8	0.3	1	0	1	0	0.1	0	0	1	0
17	0.3	0.4	0.3	1	0	1	0	0.1	0	0	1	0
18	0.6	0.3	0.9	0.3	0.3	0	0	0	0	0	0.6	0
19	0.7	0.3	0.9	0.3	0.2	0	0	0	0	0	0.8	0
20	0.7	0.6	0.9	0.6	0	0	0.2	0.5	0.3	0	0.9	1
21	0.7	0.6	0.9	0.7	0	0	0.3	0.6	0.4	0	0.8	1

本文已经证明,区间分割算法产生的极小乘积空间只包含同一类的例子,所以至少包含1个例子的极小乘

积空间可看做一个例子来处理,目的在于减少参加学习的例子数目,提高学习速度与学习精度.

通过图 1 可以看出 21 个例子的学习问题,用区间分割算法化简成 11 个例子的学习问题.

类	S_1	S_2	S_5	S_7	S_9	Contained Examples
PE	1	0	1	2	1	1,4
	0	1	1	1	1	2
	0	0	1	2	1	3
	0	0	0	0	1	9
	1	0	0	1	1	21,22
NE	0	0	1	1	1	5
	0	1	0	0	1	6,7
	1	0	0	0	1	8
	0	1	0	0	0	10
	0	0	0	0	0	11,12,13,14,15,16,17
	1	0	0	0	0	18,19

图 1　包含例子的极小乘积空间

对于图 1 给出的规则学习问题,规则学习算法 CRA 得到的表示正例子的规则由下列 3 个公式组成.

$$[x_1\neq 0]\wedge[x_2\neq 1]\wedge[x_5\neq 0]\wedge[x_9\neq 0],$$

$$[x_1\neq 1]\wedge[x_2\neq 1]\wedge[x_7\neq 1]\wedge[x_9\neq 0],$$

$$[x_1\neq 0]\wedge[x_2\neq 1]\wedge[x_5\neq 1]\wedge[x_9\neq 0].$$

把上述公式组转化成子区间形式的公式组如下:

$$\{S_1=(0.5,1]\}\wedge\{S_2=[0,0.8]\}\wedge\{S_5=(0.5,1]\}\wedge\{S_9=(0.25,1]\},$$

$$\{S_1=[0,0.5]\}\wedge\{S_2=[0,0.8]\}\wedge\{S_7=[0,0.15]\cup(0.95,1]\}\wedge\{S_9=(0.25,1]\},$$

$$\{S_1=(0.5,1]\}\wedge\{S_2=[0,0.8]\}\wedge\{S_5=[0,0.5]\}\wedge\{S_9=(0.25,1]\}.$$

本文在输入数据的噪音对学习系统的鲁棒性方面进行了比较实验.下面先引进几个相关的概念.

定义 18. 噪音幅度是指,受噪音干扰的输入例子与学习样本中"最相近例子"之间的距离,即受噪音干扰的例子 e^* 的噪音幅度等于 $\mathrm{Min}\{d(e^*,e)\mid e\in E\}$,其中 E 是学习样本集合.

定义 19. 用 e^* 表示一个受噪音干扰的例子,e 表示在学习样本中离 e^* 最近的例子.如果 e^* 与 e 的识别结果是一致的,则称 e^* 为能够被正确识别的例子,否则,称 e^* 为不能被正确识别的例子.

定义 20. 算法的识别率是指,识别系统能够正确识别的例子(受噪音干扰的)数目与所有识别例子(受噪音干扰的)之间的比值.

我们以在采煤机故障诊断系统的研究过程中采集到的数据作为学习样本,进行了识别系统对噪音数据的鲁棒性分析实验,实验结果见表 2.该学习样本的个数为 1 000;输入属性空间的维数为 12,属性值取值区间为[0,1];输出属性空间的维数为 1,属性值只取"正常"与"异常"两个值.识别例子集是对上述学习样本适当地进行人为输入噪音而得到的 5 000 个例子组成的集合.

表 2　输入数据的噪音对学习系统的影响分析

受噪音干扰的幅度≤	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
本文所提出算法的识别率(%)	97	95	94	92	90	89	86	85
Quinlan 算法的识别率(%)	92	89	85	83	81	79	75	72

表 2 表明,在噪音数据(一定范围内的)的鲁棒性方面,本文算法比 Quinlan 算法有比较明显的提高.这表明,本文提出的基于信息熵的区间分割算法及其连续属性空间上的规则学习算法是非常有效的.

5　小　结

本文研究了连续属性空间上的规则学习算法,首先证明了属性空间最小化问题是 NP 困难问题,然后给出

基于信息熵的属性空间极小化算法.在此基础上,提出了连续属性空间上的规则学习算法,并分析了该算法的计算时间复杂性.给出了数字实验结果,结果表明本算法是行之有效的.

参考文献

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

2 Quinlan J R. C4.5: Programs for Machine Learning. Ver. 1. San Mateo, CA: Morgan Kauffmann Publisher, 1993. 170～247

3 Fayyad U M, Irai K B. On the handling of continuous-valued attributes in decision tree generation. Machine Learning, 1992,20(8):88～102

4 Utgoff P E, Berkman N C, Clouse J A. Decision tree induction based on efficient tree restructuring. Machine Learning, 1997,29(1):5～44

5 洪家荣.示例学习的扩张矩阵理论.计算机学报,1991,14(6):37～42
(Hong Jia-rong. Theory of extension matrixes in learning from examples. Chinese Journal of Computers, 1991,14(6):37～42)

6 赵美德,李星原,洪家荣.示例学习的广义扩张矩阵算法及其实现.计算机学报,1994,17(9):83～88
(Zhao Mei-de, Li Xing-yuan, Hong Jia-rong. An algorithm of generalized extension matrixes in learning from examples and implementation. Chinese Journal of Computers, 1994,17(9),83～88)

7 权光日,洪炳熔,叶凤等.集合覆盖问题的启发函数算法.软件学报,1998,9(2):156～160
(Quan Guang-ri, Hong Bing-rong, Ye Feng *et al*. A heuristic function algorithm for minimum set-covering problem. Journal of Software, 1998,9(2):156～160)

8 权光日.基于规则学习的神经网络研究[博士学位论文].哈尔滨工业大学,1998
(Quan Guang-ri. Research on neural networks based on rule learning [Ph.D. Thesis]. Harbin Institute of Technology, 1998)

9 Wu X D. Optimization problems in extension matrixes. Science in China (series A), 1992,35(3):363～373

10 Chen Bin, Hong Jia-rong, Wang Ya-dong. Minimum feature subset selection problem. Journal of Computer Science and Technology, 1997,12(2):123～128

11 陈彬,洪家荣.示例学习的最大复合问题及其算法.计算机学报,1997,20(2):128～131
(Chen Bin, Hong Jia-rong. Maximum composition problem in learning from examples and the algorithm. Chinese Journal of Computers, 1997,20(2):128～131)

12 杨叔子,丁洪,史铁林等.基于知识的诊断推理.北京:清华大学出版社,1993.120～200
(Yang Shu-zi, Ding Hong, Shi Tie-lin *et al*. Diagnosic Inference Based on Knowledge. Beijing: Tsinghua University Press, 1993. 120～200)

A Rule Learning Algorithm on Continuous Attributes Space

QUAN Guang-ri[1] LIU Wen-yuan[2] YE Feng[2] CHEN Xiao-peng[1]

[1](*Weihai Campus Harbin Institute of Technology Weihai* 264200)

[2](*Department of Computer Science and Engineering Harbin Institute of Technology Harbin* 150001)

Abstract The rule learning algorithm on continuous attributes space is studied in this paper. First, the purpose and the importance of studying rule learning algorithm on continuous attributes space are briefly introduced, and then some basic concepts in the theory of rule learning are extended to the continuous attributes space. On this basis, the authors study the problem to divide continuous attributes space, and prove that the problem of min dividing continuous attributes space is a NP hard problem. The concepts of information entropy and infinite normed apply to the problem of dividing continuous attribute space and a new algorithm of dividing continuous attribute space based on the function of information entropy are presented. At last, a rule learning algorithm on continuous attributes space is presented and the data results of the experiments are given.

Key words Rule learning algorithm, continuous attribute space, information entropy, infinite normed, NP hard problem.