一种基于粗糙集理论的启发式特征选择算法

亢 婷[1,2],魏立力[1]
KANG Ting[1,2],WEI Li-li[1]

1.宁夏大学 数学计算机学院,银川 750021
2.宁夏大学 新华学院,银川 750021
1.School of Mathematics & Computer Science,Ningxia University,Yinchuan 750021,China
2.Xinhua College,Ningxia University,Yinchuan 750021,China
E-mail:liliwei@nxu.edu.cn

KANG Ting,WEI Li-li.Heuristic feature selection algorithm based on rough set theory.Computer Engineering and Applications,2008,44(30):77-79.

Abstract: Feature selection is a valid technique for information-preserving data reduction in data analysis.Rough set theory provides a mathematical tool which can be used to discovery all possible feature subsets.This paper proposes a new rough set-based heuristic function called weighted average support heuristic.The main advantage is that it considers the overall quality of the set of potential rules.In another words,it considers the weighted average support of the rules for all decision classes.At last,the example proves this method is valid.

Key words: rough set;feature selection;weighted average support heuristic

摘 要:在数据分析中,特征选择是能够保留信息的数据约简的一个有效方法。粗糙集理论提供了一种发现所有可能的特征子集的数学工具。提出了一种新的基于粗糙集的启发函数叫做加权平均支持启发函数。该方法的优点是它考虑了可能性规则集的整体质量。也就是说,对所有的决策类,它考虑了规则的加权平均支持度。最后,实例表明该方法是有效的。

关键词:粗糙集;特征选择;加权平均支持启发函数
DOI:10.3778/j.issn.1002-8331.2008.30.023 **文章编号**:1002-8331(2008)30-0077-03 **文献标识码**:A **中图分类号**:TP18

在数据分析中,特征选择是能够保留信息的数据约简的一个有效方法。特征选择的目的在于去除冗余特征。

近年来,粗糙集理论在特征选择算法中得到了广泛的应用。粗糙集理论是波兰数学家 Pawlak.Z 于 1982 年提出的一种新的处理模糊和不确定性知识的数学工具[1-2]。其主要思想是在保持分类能力不变的前提下,导出问题的决策或分类规则。这一理论的特点是:除了问题所需处理的数据之外,不需要额外提供任何外界信息或先验知识[3-4]。这一独特优点使粗糙集理论在特征子集选择领域中的应用逐渐受到重视,人们已经成功地提出一些基于粗集理论的相应算法[5-8]。目前,至少存在两种用于特征选择的启发式方法,即显著性定向法[9]和支撑定向法[10]。文献[9]中所提出的启发式方法选择最重要的特征,也就是那些使正域增加较快的特征。Zhong 在[10]中提出的启发式方法不仅考虑了正域增加的快慢,而且考虑了规则的支持度。

本文在[10]所提出的 Maxmium Support Heuristic 的基础上提出了一种新的基于粗糙集理论的启发函数,叫做 Weighted Average Support Heuristic(WASH)。该启发函数的优点在于:它考虑了可能性规则集的整体质量,也就是说,对每个决策类,它考虑了规则的加权平均支持度。实验结果表明,提出的方法是有效的。

1 预备知识

1.1 粗糙集理论的基本概念

定义 1[3] 在粗糙集理论中,信息系统被表示为 $S=(U,A,V,f)$,其中:U 表示对象的非空有限集合,称为论域;A 表示属性的非空有限集合;$V=\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;f 表示 $U\times A\to V$ 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,即 $\forall a\in A,x\in U,f(x,a)\in V_a$。$A$ 可以进一步划分为两个集合:条件属性集 C 和决策属性集 D,并满足 $A=C\cup D$ 且 $C\cap D=\Phi$。$S=(U,C\cup D,V,f)$叫做决策系统或决策表。

定义 2[3] 每一个属性子集 $P\subseteq A$ 决定了一个二元不可分辨关系 $IND(P)$

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

不可分辨关系 $IND(P)$是 U 上的一个等价关系,它构成了 U 上的一个划分 $U/IND(P)$。特别地,$U/IND(C)$和 $U/IND(D)$分

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60663003);宁夏自然科学基金(the Natural Science Foundation of Ningxia of China under Grant No.NZ0725);教育部科学技术研究重点项目(No.206159)。
作者简介:亢婷(1984-),女,硕士,主研方向为应用概率统计;魏立力(1965-),通讯作者,男,博士,教授,主研方向为统计学、人工智能的数学基础。
收稿日期:2007-11-20 **修回日期**:2008-02-29

别叫条件类和决策类。

定义 3[3] 对 $R\subseteq C,X\subseteq U$,$X$ 的 R 下近似 $R(X)$定义为 $R(X)=\cup\{Y\in U/R|Y\subseteq X\}$。

定义 4[3] D 的 C 正域:$POS_C(D)=\bigcup_{X\in U/D}\underline{C}(X)$,其中 $\underline{C}(X)$表示 X 的 C 下近似。D 的 C 正域是 U 中所有根据分类 $U/IND(C)$的信息可以准确地划分到 D 的等价类中去的对象集合。

如果 $POS_P(D)=POS_C(D)$且不存在 $R\subset P$ 使得 $POS_R(D)=POS_C(D)$,则子集 $P\subseteq C$ 叫做 C 的 D 约简。也就是说,约简是保持正域不变的最小的属性集,一个信息系统的约简不唯一。

对于决策系统 $S=(U,C\cup D,V,f)$,若 $B_1\subseteq B_2\subseteq C$,则由不可分辨关系的定义可知关系 B_2 要比关系 B_1 对 U 的分类细,再由定义 4,显然有正域的如下性质:

性质 1 若 $A\subseteq B$,则 $POS_A(D)\subseteq POS_B(D)$。

1.2 特征选择

特征选择是寻找满足一定准则的最优属性子集的过程。本文考虑两个参数:所选择的特征子集的大小和用所选择的特征产生的分类符的精度。目前,用于特征选择的方法主要有三种:穷尽搜索、随机搜索和启发式搜索。穷尽搜索算法是计算每一个可能的特征子集的特征选择度量,找到符合选择判据的最优的特征子集,如分支界限法[11]、Focus[12]和 ABB[13]。随机搜索算法[14]是在规定的时间或者次数内随机地选择特征子集来做判断,以此来找到一个次优的特征子集。启发式算法[15]是根据某种特征选择方向找到一个次优的特征子集。其中,穷尽搜索性能最好,但由于其时间复杂性,所以只有当属性较少时才使用该方法。随机搜索由于限制了搜索时间或次数,所以有可能在最优子集出现后还在进行搜索,因而现在最常用的特征选择方法是启发式搜索。

2 基于粗糙集理论的启发式特征选择算法

本文主要考虑基于粗糙集的启发函数法。这些启发函数通常被用于决定哪个属性与目标概念相关。由于特征选择的最终目的是为了减少用于产生分类规则的特征的数量,所以必须考虑可能的决策规则的质量。规则的质量可以用以下两个参数来评价:(1)规则的覆盖度,也就是协调对象的个数;(2)每个规则的支持度。

2.1 最大支持启发函数法(MSH)

Zhong 等提出了一个考虑上述两个参数的启发函数。利用该启发函数选择特征 a 使得通过将 a 加入当前特征集,能够使协调对象的数目增加较快,而且最重要规则的支持度也比加入任何其它特征大。最重要的规则是具有最大支持度的规则。该启发函数定义如下:

$$F(R,a)=|POS_{R+\{a\}}(D)|\times\text{Max } size(POS_{R+\{a\}}(D)/IND(R+\{a\}))$$

其中:$|POS_{R+\{a\}}(D)|$表明了协调对象的数量,$\text{Max } size(POS_{R+\{a\}}(D)/IND(R+\{a\}))$给出了最重要规则的支持度。

MSH 的局限性在于它所选出的特征只能使最重要的规则具有最大的支持度,而不能使可能性规则集具有最好的整体质量,也就是说,它只考虑了可能性规则的局部最优而不是全局最优。而且,当两个特征集产生的正域的大小和最重要规则的支持度相同时,MSH 就无法在这两个集合中进行选择。

2.2 加权平均支持启发函数法(WASH)

基于上述讨论,提出了一种新的启发函数,叫做加权平均支持启发函数法(WASH)。尽管 WASH 和 MSH 具有相同的时间复杂性,但 WASH 考虑了可能性规则集的整体质量而不是最重要规则的支持度。可能性规则集的整体质量 Q 是对每一个决策类而言,最重要规则的加权平均支持度。而且与 MSH 不同,WASH 考虑了所有的决策类。用 WASH 所选出的特征能在所有的决策类上产生最大的规则的平均支持度。可能性规则集的整体质量定义如下:

$$Q(R,a)=\sum_{i=1}^{n}\alpha_i S(R,a,d_i)$$

其中:$S(R,a,d_i)=\text{Max } size(POS_{R+\{a\}}(D=d_i)/IND(R+\{a\}))$是决策类$\{D=d_i\}$最重要规则的支持度。$D$ 是决策属性集,$D=\{d_1,d_2,\cdots,d_n\}$。$\alpha_i=\frac{|D=d_i|}{|U/A|}$。所以,WASH 定义为:

$$F(R,a)=|POS_{R+\{a\}}(D)|\times Q(R,a)$$

具体算法如下:

输入:信息系统决策表 $S=(U,A,V,f)$,$A=C\cup D$ 是属性集合,$C=\{a_j|j=1,2,\cdots,m\}$,$D=\{d_i|i=1,2,\cdots,n\}$分别为条件属性集和决策属性集。R 为所求特征子集,开始令 $R=\Phi$,$P=C-R$。

输出:信息系统决策表的特征子集。

步骤 1 对任意 $a_j\in P$,计算

$v_{a_j}=|POS_{R+\{a_j\}}(D)|$

$m_{a_j}=\text{Max } size(POS_{R+\{a_j\}}(D=d_i)/IND(R+\{a_j\}))$,$1\leqslant i\leqslant n$

$\alpha_i=\frac{|D=d_i|}{|U/A|}$

$M_{a_j}=\sum_{i=1}^{n}\alpha_i m_{a_j}$,$1\leqslant j\leqslant m$

步骤 2 选择使 $v_{a_j}\times M_{a_j}$最大的 a_j,并且令 $R=R+\{a_j\}$,$P=P-\{a\}$;

步骤 3 若 $POS_R(D)=POS_C(D)$,则终止,输出特征子集 R,否则转步骤 1。

3 实验及结果分析

下面通过具体的例子来说明本文所给出的 WASH 相对于 MSH 的优越性。假设表 1 是一个医疗诊断的信息表,其中 a_1,a_2,a_3是症状,D 是疾病类型,{D=1}和{D=2}分别表示疾病 1 和疾病 2。

表 1 医疗诊断信息表

	size	a_1	a_2	a_3	D
E_1	150	2	0	1	1
E_2	150	0	1	0	2
E_3	40	0	1	2	2
E_4	50	2	1	0	1
E_5	50	0	1	3	1
E_6	170	0	0	2	1
E_7	300	0	2	1	1
E_8	10	1	1	0	2
E_9	250	3	1	1	1
E_{10}	5	2	2	2	1
E_{11}	10	2	2	2	2

首先,将最大支持启发函数用于该信息表。

第一步,令 $R=\Phi$,则有:

$$F(R,a_1)=|POS_{\{a_1\}}(D)|\times \mathrm{Max}\ size(POS_{\{a_1\}}(D)/IND(\{a_1\}))=$$
$$260\times250=65\ 000$$
$$F(R,a_2)=|POS_{\{a_2\}}(D)|\times \mathrm{Max}\ size(POS_{\{a_2\}}(D)/IND(\{a_2\}))=$$
$$320\times170=54\ 400$$
$$F(R,a_3)=|POS_{\{a_3\}}(D)|\times \mathrm{Max}\ size(POS_{\{a_3\}}(D)/IND(\{a_3\}))=$$
$$750\times300=225\ 000$$

所以,$F(R,a_3)>F(R,a_1)>F(R,a_2)$。因此,症状 a_3 是第一个被选出的特征。

第二步,$R=\{a_3\}$有:

$$F(R,a_1)=|POS_{\{a_3,a_1\}}(D)|\times \mathrm{Max}\ size(POS_{\{a_3,a_1\}}(D)/IND(\{\{a_3,a_1\}\}))=$$
$$1\ 060\times300=318\ 000$$
$$F(R,a_2)=|POS_{\{a_3,a_2\}}(D)|\times \mathrm{Max}\ size(POS_{\{a_3,a_2\}}(D)/IND(\{\{a_3,a_2\}\}))=$$
$$1\ 060\times300=318\ 000$$

由于 $F(R,a_1)=F(R,a_2)$所以无法在$\{a_1,a_3\}$和$\{a_2,a_3\}$之间做出选择。但是,从两种疾病的规则支持度上发现$\{a_1,a_3\}$比$\{a_2,a_3\}$重要。表 2 和表 3 分别给出了$\{a_1,a_3\}$和$\{a_2,a_3\}$的正域。在表 2 中,$\{D=1\}$的最重要规则的支持度是 300,$\{D=2\}$的最重要规则的支持度是 150。在表 3 中,$\{D=1\}$的最重要规则的支持度依然是 300,但$\{D=2\}$的最重要规则的支持度只有 40。也就是说,从表 3 中得到的$\{D=2\}$的分类规则没有足够多的支持样本,所以,$\{a_1,a_3\}$比$\{a_2,a_3\}$更优,但是 MSH 却不能指出$\{a_1,a_3\}$和$\{a_2,a_3\}$之间的差别。而且,由于 MSH 基于经典的粗糙集模型,所以它不考虑不协调样本 E_{11}。

表 2　$\{a_1,a_3\}$的正域

	size	a_1	a_2	a_3	*D*
E_1	150	2	0	1	1
E_2	150	0	1	0	2
E_4	50	2	1	0	1
E_5	50	0	1	3	1
E_7	300	0	2	1	1
E_8	10	1	1	0	2
E_9	250	3	1	1	1

表 3　$\{a_2,a_3\}$的正域

	size	a_1	a_2	a_3	*D*
E_1	150	2	0	1	1
E_3	40	0	1	2	2
E_5	50	0	1	3	1
E_6	170	0	0	2	1
E_7	300	0	2	1	1
E_9	250	3	1	1	1

下面,用 WASH 来选择特征,由于不协调样本 E_{11} 具有相当多的支持样本,所以将它考虑在正域里。有:

$$Q(a_3,a_1)=\frac{7\times300}{11}+\frac{4\times150}{11}=\frac{2\ 700}{11}$$
$$Q(a_3,a_2)=\frac{7\times300}{11}+\frac{4\times100}{11}=\frac{2\ 500}{11}$$

所以,$Q(a_3,a_1)>Q(a_3,a_2)$也就是说,从$\{a_1,a_3\}$得到的规则集的整体质量好于从$\{a_2,a_3\}$得到的规则集的整体质量。因此:

$$F(a_3,a_1)=|POS_{\{a_3,a_1\}}(D)|\times Q(a_3,a_1)=\frac{1\ 060\times2\ 700}{11}$$
$$F(a_3,a_2)=|POS_{\{a_3,a_2\}}(D)|\times Q(a_3,a_2)=\frac{1\ 060\times2\ 500}{11}$$

显然。WASH 选择$\{a_1,a_3\}$,因为$\{a_1,a_3\}$产生的规则具有较多的支持样本。

4　总结

本文在 MSH 的基础上提出了 WASH,WASH 克服了 MSH 当两个特征集的正域大小相等时无法做出选择的缺点。而且 WASH 考虑了可能性规则集的整体质量,因此,用 WASH 选出的特征集在一定程度上是最优的。

参考文献:

[1] Pawlak Z.Rough sets[J].International Journal of Computer and Information Science,1982,11(5):341-356.

[2] Pawlak Z.A rough set view on Bayes' theorm[J].International Journal of Intelligent Systems,2003,18(5):487-498.

[3] 张文修,梁怡,吴伟志.信息系统与知识发现[M].北京:科学出版社,2003:7-12.

[4] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2006:12-16.

[5] 常犁云.一种 Rough Set 理论的属性约简及规则提取方法[J].软件学报,1999,10(11):1206-1211.

[6] Nguyen S H.Some efficient algorithms for rough set methods[C]//Proc of the Conf of Information Proceeding and Management of Uncertainty in Knowledge Based Systems,Granada,Spain,1996:1451-1456.

[7] 苗夺谦.知识约简的一种启发式算法[J].计算机研究与发展,1999,36(6):681-684.

[8] 何苗,李春葆.一种结合粗糙集理论和启发式知识的特征选取算法[J].计算机应用,2003,23(2):113-115.

[9] Hu X.Knowledge discovery in databases:An attribute-oriented rough set approach[D].University of Regina,Canada,1995.

[10] Zhong N,Dong J Z,Ohsuga S.Using rough sets with Heuristics for feature selection[J].Journal of Intelligent Systems,2001,16:199-214.

[11] Narendra P,Fukunaga K.A branch and bound algorithm for feature subset selection[J].IEEE Trans on Computer,1977, 26(9):917-922.

[12] Almullim H,Dietterich T G.Learning with many irrelevant features[C]//Proceedings of the 9th National Conference on Artificial Intelligence(AAAI-91),Anaheim,California,1991-07:547-552.

[13] Liu H,Motoda H,Dash M.A monotonic measure for optimal feature selection[C]//Proc of ECML-98,1998.

[14] Liu H R S.A probabilistic approach to feature selection—a filter solution[C]//Proceedings of the 13th International Conference on Machine Learning,1996:319-327.

[15] Kira K,Rendell L.A practical approach to feature selection[C]//Proceedings of the 9th International Conference on Machine Learning,1992:249-256.

[16] 张腾飞,肖健梅,王锡淮.粗糙集理论中属性相对约简算法[J].电子学报,2005,33(11):2080-2083.