基于决策分类的决策表分解方法研究 *

王加阳, 胡 沛
(中南大学 信息科学与工程学院, 长沙 410083)

摘 要: 针对当前基于属性重要性的决策表属性集分解方法存在的不足,提出了一种新型的基于决策分类的决策表属性集分解方法。分析了近似分类质量和属性重要性与决策分类之间的关系,利用粗糙集理论,从提高子决策表中决策分类正确性的角度出发考虑条件属性与决策属性之间的关系,提出了决策表分解的条件属性选择量度并对决策表实施属性集分解。

关键词: 决策表; 分解; 粗糙集理论
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2009)01-0108-03

Research on decomposition of decision table based on decision making

WANG Jia-yang, HU Pei
(*College of Information Science & Engineering, Central South University, Changsha* 410083, *China*)

Abstract: There are some defects in the attribute set decomposition of decision table based on the attribute significance. The paper introduced a novel feature decomposition of decision table approach based on decision making. It analyzed the differences between the quality of approximation and the attribute significance to decision making. With the rough set theory, the relationships of condition attributes and decision attributes were considered increasing the attribute classification rate in the decision table, and a new attribute selection criterion was proposed. Then decision table was decomposed with the attribute selection criterion.

Key words: decision table; decomposition; rough set theory

随着计算机和数据采集技术的进步,数据的积累无论是在数据对象个数上还是在数据维度上都以无法估算的速度迅速增长。数据量的不断增大给现有的数据分析和处理技术提出了挑战,许多实际的决策表包含了大量的对象和属性,如何从这些堆积如山的数据中挖掘有用的信息已成为当今数据挖掘领域的热点[1]。决策表分解是解决大型决策表数据海量性和复杂性问题的一种有效手段。其思想是将一个大型决策表分解为若干规模较小且易于处理的子表,在子表中进行规则获取,可以减少每次处理的数据量,避免直接在复杂系统中建模的困难和缺陷,提高数据分析的效率和质量[2]。

本文提出了一种基于决策分类的决策表分解方法。首先从条件属性与决策属性之间的关系出发,分析了近似分类质量和属性重要性与决策分类的关系,根据提高子决策表条件属性正确分类能力的标准提出了一种新的条件属性选择量度,按照属性选择量度指导子决策表的属性选取,将决策表分解为几个子决策表。

1 粗糙集理论基本概念

粗糙集理论是波兰数学家 Z. Pawlak 于 1982 年提出的一种处理含糊和不确定性问题的新型数学工具,已经在机器学习、数据挖掘等领域中得到广泛应用,决策表信息系统是粗糙集理论的主要研究对象[3]。

定义 1 一个决策表信息系统 $S=\langle U,R,V,f\rangle$。其中:U 是非空有限集合,即为论域;$R=C\cup D$,C 和 D 分别称为条件属性集和决策属性集,$D\neq\varnothing$是非空有限集合;$V=\bigcup_{r\in R}V_r$ 是属性值的集合,V_r 表示属性 $r\in R$ 的值域;$f:U\times A\rightarrow V$ 是一个信息函数,它指定 U 中每一个对象 x 的属性值。

定义 2 在决策表信息系统 $S=\langle U,R,V,f\rangle$ 中,对于 $B\subseteq R$,则 B 在 U 上的不可分辨关系定义为

$$\mathrm{IND}(B)=\{(x_1,x_2)\in U\times U\mid b(x_1)=b(x_2)\ \forall b\in B\}=\mathop{\mathrm{I}}_{b\in B}\mathrm{IND}(b) \tag{1}$$

显然 $\mathrm{IND}(B)$是一个等价关系,它的所有等价类的集合记为 $U/\mathrm{IND}(B)$,简记为 U/B。

定义 3 给定决策表信息系统 $S=\langle U,R,V,f\rangle$,对于每个子集 $X\subseteq U$ 和不可分辨关系 B,X 的下近似集和上近似集可以分别定义为

$$B_-(X)=\cup\{Y_i\mid(Y_i\in U/\mathrm{IND}(B)\wedge Y_i\subseteq X)\} \tag{2}$$

$$B^-(X)=\cup\{Y_i\mid(Y_i\in U/\mathrm{IND}(B)\wedge Y_i\,\mathrm{I}\,X\neq\varnothing)\} \tag{3}$$

定义 4 设 U 为一个论域,P、Q 为 U 上的两个等价关系簇,Q 的 P 正域 $\mathrm{POS}_p(Q)$定义为

$$\mathrm{POS}_p(Q)=\bigcup_{X\in U/Q}P_-(X) \tag{4}$$

相对正域 $\mathrm{POS}_p(Q)$是论域 U 的所有那些使用分类 U/P 所表达的知识中能够正确地分类到 U/Q 的等价类之中的对象的集合。

定义 5 设集合族 $F=\{X_1,X_2,\cdots,X_n\}$($U=\cup_{i=1}^{n}X_i$)是论域 U 上定义的知识,B 是一个属性子集,定义 B 对 F 近似分类质量 $\gamma_B(F)$为

$$\gamma_B(F)=\sum_{i=1}^{n}|B_-(X_i)|/|U| \tag{5}$$

收稿日期: 2008-04-07; **修回日期**: 2008-06-23 **基金项目**: 湖南省自然科学基金资助项目(06JJ20075);湖南省科技计划资助项目(2008FJ3184)

作者简介:王加阳(1963-),男,教授,博士,主要研究方向为计算智能与信息融合;胡沛(1983-),男, 硕士研究生,主要研究方向为智能信息处理、粗糙集理论及应用(huxiaopei163@163.com).

定义 6　F 是属性集 D 导出的分类,C 是条件属性集合,$D=\{d\}$ 是决策属性集合,且 $A\subset C$ 则对于任意属性 $a\in C-A$ 的重要性 $\mathrm{SGA}(a,A,D)$ 为

$$\mathrm{SGA}(a,A,D)={}_{rA\cup\{a\}}(F)-{}_{rA}(F) \tag{6}$$

2　决策表分解

现实应用中,决策表的复杂性主要来自于属性数量的增长。随着属性集的不断扩大,为了建立有效的分类模型,避免出现过度拟合现象,训练集中所需的对象数需呈指数级增长,针对决策表的复杂性首先考虑对属性集的处理,力求减小属性集的规模。

属性集分解方法是从给定原始属性集中按照一定的标准选取属性结成一组,然后按照一定的算法来识别定义在属性组上的中间概念层次,从而达到降维、增强模型可理解性的目的。这一方法要求在保持模型的分类能力的前提下,简化计算并增强模型的可理解性。文献[4,5]提出了基于属性重要性的决策表属性集分解方法,优先选择那些属性重要性较大的条件属性对决策分类;对于不能正确分类的对象,属性重要性小的条件属性可以对决策分类起辅助作用。由于属性重要性反映的是去掉某条件属性后对决策分类能力的影响,文献[4,5]提到的决策表分解方法可能会导致属性重要性较大的条件属性所形成的子决策表对决策属性具有较差的正域分类能力,不利于知识获取。本文根据近似分类质量和属性重要性与决策分类的关系,从提高子决策表条件属性分类能力出发提出了一种新的决策表分解属性选择量度,使其能够有效地提高子决策表的决策分类质量。

2.1　决策表条件属性与决策属性之间的关系研究

近似分类质量和属性重要性作为粗糙集理论中描述条件属性与决策属性之间关系的两个重要概念,它们从不同的角度刻画了条件属性与决策属性的关系。近似分类质量只关注条件属性对决策属性正区域的变化情况;值越大表明该属性集对决策属性的正域分类能力越强。而属性重要性侧重反映在属性集中增加条件属性对近似分类质量的影响,值越大说明去掉该属性后引起的决策正域变化越大,即能正确分类到决策属性划分的等价类中的元素越少,该属性对决策属性越重要。

设 $B=\{x\}$ 为非空单属性集,决策表中的条件属性可能会出现 B 的近似分类质量较大,而属性 x 重要性却很低的情况;也可能会出现属性 x 重要性很高而 B 的近似分类质量较小的情况。如表 1 所示,条件属性集 $\{a\}$ 和 $\{b\}$ 具有最大的近似分类质量,属性 a,b 却具有最小的属性重要性;条件属性 c 虽具有最大的属性重要性,条件属性集 $\{c\}$ 却具有最小的近似分类质量。

表 1　决策表 1

U	a	b	c	d	U	a	b	c	d
1	1	0	3	1	6	3	2	0	0
2	0	1	3	0	7	3	3	0	1
3	1	0	0	1	8	3	3	1	0
4	0	1	0	0	9	3	3	2	1
5	2	3	0	0	10	3	3	3	0

2.2　决策表分解条件属性的选择研究

为了提高子决策表条件属性对论域的正确分类能力,子表中的条件属性进行选取时需从近似分类质量和属性重要性两方面考虑。

1)子决策表中的对象能被尽可能多地正确分类

选取到子表中的第一个条件属性 b 要有对决策属性较强的正域分类能力。如果从属性重要性的角度选取条件属性,可能会出现由属性重要性较大的条件属性所形成的子表对决策属性具有较差的正域分类能力。条件属性 b 应从近似分类质量考虑,单条件属性集 $\{b\}$ 需有最大的近似分类质量 ${}_{rB}(F)$;如果有多个单条件属性集的近似分类质量同为最大,从它们所包含的单条件属性与决策属性的重要性考虑,优先选择具有较大属性重要性的。

如果一个决策表中每个条件属性都不能对决策进行正确分类,即 ${}_{rC_i}(F)=0(C_i\in C)$;如表 2 中所示的类似情况,应考虑使用条件属性组合求其与决策属性的近似分类质量。具体方法为首先两两属性组合,找出近似分类质量值最大的组合;如果近似分类质量仍为 0,应使用三个属性组合,依此类推。

2)子决策表的条件属性对决策正域的划分能很好地互补

选取到子表中的其他条件属性能对决策分类起到很好的辅助作用。如果从近似分类质量的角度选取条件属性,虽可以保证条件属性对论域有较强的正确分类能力,但可能会出现子表中的条件属性对决策属性的正域划分有大量相交子集的情况,即子表中的各条件属性的属性重要性均较低,其余的条件属性不能对现有的决策分类起到很好的辅助作用。选取到子表中的条件属性 a 应满足加入该属性后子表的条件属性对决策正域分类有较大增加,即相对于当前条件属性集 A,a 属性需有较大的属性重要性 $\mathrm{SGA}(a,A,D)$。如果有多个属性的属性重要性同为最大,则从它们的分类能力考虑,优先选择具有较大正域分类能力的条件属性。

定义 7　F 是属性集 D 导出的分类,C 是条件属性集合,$D=\{d\}$ 是决策属性集合,则对于任意属性集 $A\subset C$ 的重要性 $\mathrm{SGA}(A,C,D)=rc(F)-rc-A(F)$。

如果条件属性的属性重要性均为 0,如表 3 所示的类似情况,属性 b、c、e 相对于 $\{a\}$ 的属性重要性均为 0,此时应考虑使用条件属性组合求属性集的重要性。具体方法为首先两两属性组合,找出属性集重要性值最大的组合;如果属性集重要性仍为 0,应使用三个属性组合,依此类推。

表2　决策表2

U	a	b	c	D
1	0	0	1	1
2	1	0	1	0
3	1	0	0	1
4	0	0	0	0
5	1	1	1	1
6	0	1	1	0

表3　决策表3

U	a	b	c	e	D
1	2	0	0	1	1
2	0	1	0	1	0
3	0	1	0	0	1
4	0	0	0	0	0
5	0	1	1	1	1
6	0	0	1	1	1

2.3　基于决策分类的决策表分解算法描述

对于含有多个决策属性的决策属性集 D,可以将其转换为单一的决策属性,本文只考虑决策属性集包含一个决策属性 d 的情况,即 $D=\{d\}$,$S=\langle U,C\cup D,V,f\rangle$。

算法 1　DDBDM

输入:大型决策表 $S=\langle U,C\cup D,V,f\rangle$。其中 U 为论域,C,D 分别为条件属性集和决策属性集。

输出:一系列符合要求的子表。

a)设置每个子表所允许的最大条件属性集个数为 n'(n'由用户实际需要设置或由专家结合问题的先验知识确定)。

b)设决策属性 D 的等价类 $F=U/D=\{D_1,D_2,\cdots,D_m\}$,用式(5)计算 C 中各单属性集的近似分类质量,取满足 $\mathrm{MAX}({}_{rB}(F))$ 的属性 x。如果值为 0,使用属性组合求值最大的 ${}_{rB}(F)$ 属性集 x,$j=0$,$R_j=\varnothing$,$k=0$。

c)将 x 加入集 R_j 中,$k++$。若 $C=\varnothing$,跳到 g)执行;若 $C\neq\varnothing$,判断是否满足 $k\leqslant n'$,如果满足的话,重复执行 d);如果不满足,执行 b)。

d)用式(6)计算 C 中各属性的属性重要性,取满足 $\mathrm{MAX}(\mathrm{SGA}(a,R_j,D))$ 的条件属性 y。如果值为 0,使用属性组合求值最大的 $\mathrm{SGA}(a,A,D)$ 属性集 y;将 y 加入 R_j。

e) $C = C - R_j$,生成决策表 $S_j = \langle U, R_j, V, f\rangle$ 和 $T'_j = \langle U, C, V, f\rangle$。其中:$S_j$ 为新生成的子决策表;T'_j 为中间决策表; $++j$。

f)对中间决策表 T'_j 重复 b) ~e)。

g)输出各子表 $S_i(0 \leqslant i \leqslant j)$。

3　基于决策分类的决策表分解实例

如表4 所示的决策表,$C = \{a, b, c, e, f, g\}$ 是条件属性集合,$D = \{d\}$ 是决策属性集合。下面利用本文所提到的基于决策分类的决策表分解方法对决策表4 进行属性集分解。

表4　决策表4

U	a	b	c	e	f	g	d	U	a	b	c	e	f	g	d
1	1	1	2	0	3	1	1	6	2	2	1	2	2	2	0
2	0	0	2	0	1	0	0	7	2	2	0	0	2	2	1
3	1	1	2	0	0	0	1	8	2	2	1	0	2	0	0
4	0	0	2	2	0	1	0	9	2	2	2	0	0	1	1
5	1	2	2	1	2	0	1	10	2	2	2	0	1	0	0

a)定义子表中所允许的最大条件属性个数 $n' = 3$。

b)计算各单条件属性集的近似分类质量:

$$r_{\{a\}}(F) = 0.5 \quad r_{\{b\}}(F) = 0.4 \quad r_{\{c\}}(F) = 0.3$$

$$r_{\{e\}}(F) = 0.3 \quad r_{\{f\}}(F) = 0.1 \quad r_{\{g\}}(F) = 0$$

c) $\{a\}$ 的近似分类质量最大,选 a 进入子决策表1(表5),求此时基于属性集 $\{a\}$ 的属性重要性:

$$SGA(b, \{a\}, D) = 0 \qquad SGA(c, \{a\}, D) = 0.3$$

$$SGA(e, \{a\}, D) = 0.1 \qquad SGA(f, \{a\}, D) = 0.2$$

$$SGA(g, \{a\}, D) = 0.3$$

d)属性 c、g 相对于 $\{a\}$ 的属性重要性最大,选 c 进入子决策表1,求此时基于属性集 $\{a, c\}$ 的属性重要性:

$$SGA(b, \{a, c\}, D) = 0 \qquad SGA(e, \{a, c\}, D) = 0$$

$$SGA(f, \{a, c\}, D) = 0.2 \qquad SGA(g, \{a, c\}, D) = 0.2$$

e)属性 f、g 相对属性 $\{a, c\}$ 的属性重要性最大,选 f 进入子决策表1;

f)同理,属性 b、e、g 选入子决策表2(表6),分解结束。

表5　子决策表1

U	a	c	f	d	U	a	c	f	d
1	1	2	3	1	6	2	1	2	0
2	0	2	1	0	7	2	0	2	1
3	1	2	0	1	8	2	1	2	0
4	0	2	0	0	9	2	2	0	1
5	1	2	2	1	10	2	2	1	0

表6　子决策表2

U	b	e	g	d	U	b	e	g	d
1	1	0	1	1	6	2	2	2	0
2	0	0	0	0	7	2	0	2	1
3	1	0	0	1	8	2	0	0	0
4	0	2	1	0	9	2	0	1	1
5	2	1	0	1	10	2	0	0	0

4　结束语

从决策表中快速有效地提取规则是决策表分解的主要目的。基于决策分类的决策表方法从决策分类和知识发现的角度考虑条件属性与决策属性之间的关系,从属性重要性和近似分类质量两个角度对决策表进行分解,能使子决策表中被正确分类的对象尽可能多,提高了决策分类质量,有利于发现知识之间的隐含关系。本方法结构简单,能够很好地解决大型决策表的分解问题。

参考文献:

[1] PAWLAK Z. Theorize with data using rough sets[C]//Proc of the 26th Annual International Computer Software and Applications Conference. Los Alamitos:IEEE Press, 2002:1125-1128.

[2] 王加阳,刘柳明,罗安. 大型决策表分解方法研究[J]. 计算机科学,2007,34(8):211-214.

[3] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982,11(5):341-356.

[4] 樊群,赵卫东,达庆利. 一种基于粗集的实例分解归纳学习方法[J]. 管理工程学报,2001,15(2):79-81.

[5] 赵卫东,李旗号. 复杂决策表的特征提取方法研究[J]. 小型微型计算机系统,2002,10(23):1241-1244.

(上接第 78 页) 20 个分量分类器分别对各样本点进行预测,并根据类别投票的多少来决定样本点最终的分类类别,将投票数目多的分类作为该样本点的类别。

4.2　Boost-SVM 算法的实验结果

本文的样本训练集采用 UCI[8] 的五组数据,对其进行了训练和测试,并与 LIBSVM 的学习性能进行了比较,如表1 所示。

表1　Boost-SVM 算法与 LIBSVM 性能比较

UCI datasets	LIBSVM/%	Boost-SVM/%
diabetes	78.125 0	78.515 6
liver-disorders	77.391 3	93.333 3
breast-cancer	97.744 4	97.932 3
Australian	85.797 1	87.681 1
German	78.900 0	79.600 0

表1 中列出了将五个 UCI 数据集用于 LIBSVM 以及 Boost-SVM 算法所得到的分类精度。本文将训练这五组数据集时所使用的迭代次数都设定为 20 次。通过与 LIBSVM 学习精度相比,五组数据集的学习精度后者均高于前者。从以上实验可以看出,Boost-SVM 算法的学习精度与 LIBSVM 相比,具有更好的学习性能。

5　结束语

本文提出了 Boost-SVM 学习算法,将支持向量机作为基础学习器应用于 AdaBoost 的迭代学习框架中。实验结果表明,此算法将支持向量机和 AdaBoost 算法的优点很好地集于一身,改进了学习性能,提高了学习精度。另外,文献[9]提出了一种 σ-AdaBoostRBFSVM 算法,通过使用变化着的参数 σ 试图更好地适应每次迭代学习的 SVM 学习器,同样达到了提高分类器的分类精度和泛化性的目的。由于缺乏该算法的实现程序,本文并未与 σ-AdaBoostRBFSVM 进行比较研究。现阶段 Boost-SVM 算法还仅局限于解决两类分类问题,笔者将来的工作是解决多类问题以及回归问题,并且将其应用于大规模数据学习问题中。

参考文献:

[1] VAPNIK V N. The nature of statical learning theory[M]. London: Springer-Verlag, 1995.

[2] VAPNIK V N. Principles of risk minimization for learning theory[C]// Advances in Neural Information Processing Systems 4. San Francisco: Morgan Kaufmann Publishers,1992:831-838.

[3] FREUND Y, SCHAPIRE R E. A decision-theretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997,55(1) : 119-139.

[4] DUDA R O, HART P E. 模式分类[M]. 李宏东,等译. 北京:电子工业出版社, 2001.

[5] CHANG C C, LIN C J. LIB SVM:a library for support vector machines[EB/OL]. (2002-03-10). http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[6] LIN C J. LIBSVM[EB/OL]. (2003-06-23). http://www.csie.ntu.edu.tw/~cjlin/.

[7] CHEN P H, FAN Rong-en, LIN C J. A study on SMO-type decomposition methods for support vector machine[J]. IEEE Trans on Neural Networks, 2006,17(4):893-908.

[8] Machine learning repository[EB/OL]. (2002-04-13). http://archive.ics.uci.edu/ml/.

[9] 王晓丹, 孙东延. 一种基于 AdaBoost 的 SVM 分类器[J]. 空军工程大学学报, 2006,7(6):54-57.