文章编号:1000-582X(2008)01-0074-03

不完备信息的粗糙集-贝叶斯识别方法

钟　波,罗会亮
(重庆大学 数理学院 重庆 400030)

摘　要:为了对不完备信息进行有效识别,引入粗集贝叶斯定理,结合粗集约简识别方法,建立基于粗糙集的最小错误率贝叶斯决策准则,归纳出不完备信息模式的一种统计意义上的识别方法。实验表明,与区分矩阵属性约简识别法和贝叶斯分类器识别法相比,此方法在识别不完备信息模式时准确率更高,实用性更强。

关键词:粗糙集;不完备信息模式; 粗集贝叶斯定理;识别
中图分类号:TP391　　文献标志码:A

Merged Bayesian Theory of Rough Sets Recognition Method Research In Incomplete Information

ZHONG Bo, LUO Hui-Liang
(College of Mathematics and Science , Chongqing University, Chongqing 400030, P. R. China)

Abstract: This research established a Bayesian decision criterion of minimum error rate based on rough sets in order to recognize effectively incomplete information. The work introduced rough sets Bayes´ theorem and a combining rough sets reduction method. A statistics theory method for recognizing incomplete information pattern was derived. The computing tests of incomplete information recognition show the merged method of rough sets and Bayesian theory is more accurate and practical than either the discernibility matrix attribute reduction or Bayesian classifier recognition methods.

Key words: rough sets ; incomplete information pattern ; rough sets Bayes' theorem; recognition

在许多实际问题的智能专家系统建立过程中[1-3],因受到试验条件和水平的限制,试验结果会存在错误、虚假和缺失数据而常常无法获得完备信息,所以经常需要对不完备信息进行识别。如何合理可行地识别不完备信息,将影响到智能专家系统的应用范围和预测的准确性[4],因此研究对不完备信息的识别具有重要意义。

目前,不完备信息的识别方法主要有粗糙集约简法和贝叶斯网络法[4-6]。由于粗糙集理论在描绘知识表达系统中对不同属性的重要性及在知识表达空间的简化方面具有优势,所以利用粗糙集的属性约简方法进行分类,不需要除问题所需处理数据集合之外的任何信息,从而大大简化了计算的复杂性[7-9]。但是当系统含有不确定决策规则和模式属性信息缺失较多时,粗糙集理论不能很好地进行属性约简和分类,决策效率相对较低[4];贝叶斯分类器具有柔和性、容错性的优点,可以很好地处理上述问题,但是当属性较多时,它的求解规模太大,计算繁琐[1-4]。如果将粗糙集方法和贝叶斯分类理论相结合,利用粗糙集对数据进行约简,利用贝叶斯理论训

收稿日期:2007-10-11
基金项目:重庆市自然基金资助项目(CSTC2005BB3169)
作者简介:钟波(1964-),女,重庆大学副教授,博士,主要从事信息处理技术,系统优化和可靠性方向的研究,(Tel)023-65433169;(E-mail)cdzhongbo@sina.com。

重庆大学期刊社信息网 (http://qks.cqu.edu.cn/)

练约简后的数据,得到粗糙集-贝叶斯模式识别模型,提高了对不完备信息模式的分类能力和决策效率。该方法既能克服粗糙集刚性推理[10-13]的弱点,也能避免单纯贝叶斯理论计算复杂的弊端。

1　粗糙集的有关定义和记号

1)令 $S=<U,A,V,f>$ 是一信息系统,其中:$U\neq\phi$ 是论域;A 是属性集,$A=C\cup D$,C 和 D 分别表示条件属性集与决策属性集,$C\cap D=\phi$;V 是 A 的值域;f:$U\times A\to V$ 是一个信息函数,表示 U 中指定的每一个对象的属性值。

2)设 $\mathrm{Dec}(S)=\{\Phi_i\to\Psi_i\}_{i=1}^{m}(m\geq 2)$ 是 $S=\langle U,C\cup D,V,f\rangle$ 的规则集合。$\|\Phi\|_S$ 表示在 S 中,所有满足 Φ 的 U 中的对象 X,则有以下性质:

$$\|\Phi\vee\Psi\|_S=\|\Phi\|_S\cup\|\Psi\|_S;$$

$$\|\Phi\wedge\Psi\|_S=\|\Phi\|_S\cap\|\Psi\|_S;$$

3)记 $\mathrm{supp}_S(\Phi,\Psi)=\mathrm{card}(\|\Phi\wedge\Psi\|)$ 表示 S 中的决策规则 $\Phi\to\Psi$ 的支持量,则

$$\sigma_S(\Phi,\Psi)=\frac{\mathrm{supp}_S(\Phi,\Psi)}{\mathrm{Card}(U)}=\frac{\mathrm{Card}(\|\Phi\wedge\Psi\|_S)}{\mathrm{Card}(U)},\tag{1}$$

称为决策规则 $\Phi\to\Psi$ 在 S 中的支持度(The strength of the Decision),它说明能被该 S 决策规则分类的对象在决策表中所占的比例。

4)对任意的 $X\subseteq U$,定义概率分布:$P_U(X)=\mathrm{card}(X)/\mathrm{card}(U)$。记 $\pi_S(\Phi)=P_U(\|\Phi\|_S)$,$\forall\Phi\to\Psi$ 定义条件概率:$\pi_S(\Psi/\Phi)=P_U(\|\Psi\|_S/\|\Phi\|_S)$,则 $\pi_S(\Psi/\Phi)=\dfrac{\mathrm{Card}(\|\Phi\wedge\Psi\|_S)}{\mathrm{Card}(\|\Phi\|_S)}$,其中 $\|\Phi\|_S\neq\phi$。

显然,如果 $\pi_S(\Psi/\Phi)=1$,则规则 $\Phi\to\Psi$ 是一个确定的决策规则;如果 $0<\pi_S(\Psi/\Phi)<1$,则规则 $\Phi\to\Psi$ 是一个不确定的决策规则。

5)$\forall\{\Phi\to\Psi_i\}_{i=1}^{n}$有:

$$\pi_S(\Psi_i/\Phi)=\pi_S(\Phi/\Psi_i)\pi_S(\Psi_i)/\sum_{i=1}^{n}\pi_S(\Phi/\Psi_i)\pi_S(\Psi_i)\tag{2}$$

称式(2)为基于粗糙集的贝叶斯公式[11]。

2　不完备信息下的粗糙集与贝叶斯融合的模式识别模型与方法

2.1　基于粗糙集的最小错误率贝叶斯决策准则

粗糙集识别方法是根据决策规则来对待识别对象 x 进行模式识别,$\forall\Phi\to\Psi\in\mathrm{Dec}(S)$及 $x\in\|\Phi\|_S$,当 $0<\pi_S(\Psi/\Phi)<1$ 时,待识别模式 x 将对应着不确定的决策规则,即模式 x 有多种分类的可能性,此时利用粗糙集决策规则将无法直接对模式进行分类。

设 $\mathrm{Dec}(S)$ 是 S 的决策算法,$\{\Phi\to\Psi_j\}_{j=1}^{n}\in\mathrm{Dec}(S)$,$\forall x\in\|\Phi\|_S$,需确定 $i,i\in\{1,2,\cdots,n\}$,使得 $x\in\|\Psi_i\|_S$。考虑对 x 进行模式识别,从尽量减少分类错误的要求出发,根据经典的最小错误率贝叶斯决策准则原理[13],可得基于粗糙集的最小错误率贝叶斯决策准则如下:

若 $\pi_S(\Psi_i/\Phi)=\max\limits_{j=1,\cdots,n}\pi(\Psi_j/\Phi)$时,则 $x\in\|\Psi_i\|_S$。

利用式(2)可得:

若 $\pi_S(\Phi/\Psi_i)\pi_S(\Psi_i)=\max\limits_{j=1,\cdots,n}\pi_S(\Phi/\Psi_j)\pi_S(\Psi_j)$,则 $x\in\|\Psi_i\|_S$。　(3)

2.2　不完备信息的模式识别模型与方法

粗糙集约简是在不丢失信息的前提下,能够与原信息系统表达同样知识的最小条件属性集,它是保持信息系统的相同分类能力的最简形式[8]。上述决策算法的性质及判断准则对决策规则的粗糙集约简仍然适用。

利用粗糙集约简来识别一个不完备信息的新模式 x 时,待识别模式 x 与决策算法 $\mathrm{Dec}(S)$ 可能存在着以下几种关系:1)新模式 x 与一条确定规则匹配;2)新模式 x 与多条确定规则匹配,且结论相同;3)新模式 x 与多条规则相匹配,但结论不同;4)新模式 x 与某些规则部分信息相匹配;5)新模式 x 与任何规则完全不匹配。

①对于不完备信息模式 1)、2),可利用约简决策规则直接进行识别;

②对于模式 3),若是对应唯一的不确定规则,则由式(3)可直接进行识别;若是对应多条不确定规则,则可由式(3)将每一条不确定规则先转为确定规则,然后再考虑与多条确定规则相匹配但结论不同的情况。因此不妨假设相匹配的全为确定规则,即设 $\{\Phi_j\to\Psi_j\}_{j=1}^{k}$ 是模式 x 满足的所有规则集,其中:$\pi_S(\Psi_j/\Phi_j)=1$;$x\in\|\Phi_j\|_S$,$j=1,\cdots,k$。需确定 $i,i\in\{1,\cdots,k\}$,使得 $x\in\|\Psi_i\|_S$。由式(1)可知支持度

$$\sigma_S(\Phi_i,\Psi_i)=\frac{\mathrm{supp}_S(\Phi_i,\Psi_i)}{\mathrm{Card}(U)}=\frac{\mathrm{Card}(\|\Phi_i\wedge\Psi_i\|_S)}{\mathrm{Card}(U)}=\pi_S(\Phi_i/\Psi_i)\pi_S(\Psi_i),\tag{4}$$

度量了决策规则 $\Phi_i\to\Psi_i$ 在决策集中的重要程度,并且充分利用了决策表的先验信息,故可用于对模式 x 的识别,即

若 $\sigma_S(\Phi_i,\Psi_i)=\max\limits_{j=1,\cdots,k}\sigma_S(\Phi_j,\Psi_j)$,则 $x\in\|\Psi_i\|_S$。由式(4)可得:

若 $\pi_S(\Phi_i/\Psi_i)\pi_S(\Psi_i)=\max\limits_{j=1,\cdots,n}\pi_S(\Phi_i/\Psi_j)\pi_S(\Psi_j)$,则 $x\in\|\Psi_i\|_S$。

③对于模式4),由于待识别模式 x 与所有规则都只是部分匹配,在没有其他信息的情况下,只能认为当某条决策规则与待识别模式匹配程度最大时,它所确定的类就是待识别模式 x 应属的类,其中匹配程度可直接使用相匹配的属性个数来度量。若最大匹配规则不唯一,此时可根据待识模式 x 在所有最大匹配规则中分别出现的属性信息,将所有匹配规则重新按匹配信息进行合并,然后仅将待识别模式出现的属性信息带入这些规则中来进行分类,即可将模式4)转化为模式1)、2)、3)的情况。即

设 $\{\Phi'_j\rightarrow\Psi_j\}_{j=1}^m$ 是与模式 x 匹配程度最大的所有规则集,假设待识模式 x 与 Φ'_j 相匹配的属性构成条件 Φ_i,则决策规则 $\{\Phi'_j\rightarrow\Psi_j\}_{j=1}^m$ 变为 $\{\Phi_i\rightarrow\Psi_i\}_{i=1}^k$,利用规则 Φ_i,根据模式 x 所对应1)、2)、3)中的某种情况可对它分别进行识别。

④对于模式5),方法失效。可考虑增加模式的属性信息或对属性缺失值进行数据补齐,然后再用此方法进行识别。

3 实例分析

采用UCI机器学习数据库(http://www.ics.uci.edu/~mlearn/MLRepository.html)中的数据集Iris data、Waveform data和Wine data进行实证,它们均为3分类数据。其中:Iris data含4个条件属性,共有150条数据;Waveform data含21个条件属性,共有300条数据;Wine data含13个条件属性,共有178条数据和3个决策值。

先随机从以上3个数据库分别抽取50条数据作为测试集,再将各自剩余的数据分别作为训练集,对训练集的数据整理后利用区分矩阵进行属性约简,提取决策规则。将50条测试数据随机的去掉1个或2个属性值而得到100条数据,将这100条数据作为测试集分别采用区分矩阵属性约简方法、贝叶斯分类器[13]及笔者的方法进行识别,采用Matlab7.0编程,在硬件环境为PIV2.0下,其识别结果的正确率比较如表1所示,其中括号中的数据表示运行时间(单位:S)。

对所采用的3个数据集,从实验结果可知:1)研究方法总体上精确度均高于另2种方法,而识别速度比区分矩阵属性约简识别法略慢,但比贝叶斯分类器识别法有明显提高,有效地降低了贝叶斯分类器识别法的求解复杂度;2)对Waveform data和Wine data的识别准确率明显优于Iris data,其主要原因是Waveform data和Wine data的训练样本数更充分,缺少的属性个数与总的属性个数比起来相对较少,因而关键信息丢失的可能性也就相对小,计算所得的类后验概率相对更精确,从而提高了识别的准确率。

表1 研究方法同区分矩阵粗集约简方法和贝叶斯分类器的正确识别率之比较 %

识别方法	样本数	Iris data		Waveform data		Wine data	
		缺1个	缺2个	缺1个	缺2个	缺1个	缺2个
区分矩阵约简	50	46(6.7)	26(9.05)	92(7.2)	80(9.5)	84(6.8)	72(8.5)
贝叶斯分类器	50	48(23.8)	28(14.3)	94(60.6)	80(31.3)	88(51.2)	78(34.3)
研究方法	50	82(7.01)	64(11.3)	98(10.3)	92(16.7)	98(8.7)	90(10.1)

4 结 论

1)利用基于粗糙集的最小错误率贝叶斯决策准则处理不确定规则,将约简后的决策表作为识别的先验信息,有效地缩减了问题求解规模,克服了粗糙集和贝叶斯推理在处理不完备信息模式方面的缺陷,增强了处理不确定问题的决策能力。

2)与其它的分类算法相比,粗糙集贝叶斯决策分类理论上具有最小的出错率,且不需要样本训练,具有较强的实用性。

3)当决策表中样本数多,代表性强时,基于粗糙集-贝叶斯理论推理的置信度较高。但当所得几个类的后验概率相差很小时,正确判断率将降低,这是有待于进一步改进的地方。

(下转82页)

[21] 刘疆鹰，赵由才，赵爱华. 大型垃圾填埋场渗滤液COD衰减规律[J]. 同济大学学报:自然科学版,2000,28(3):328-332.
LIU JIANG-YING, ZHAO YOU-CAI, ZHAO AI-HUA, et al. Natural reduction of COD in large-scale landfill leachate[J]. Journal of Tongji University, 2000, 28(3): 328-332.

[22] 刘疆鹰,徐迪民,赵由才,等. 大型垃圾填埋场渗滤水氨氮衰减规律[J]. 环境科学学报, 2001, 21(3):323-327.
LIU JIANG-YING, XU DI-MIN, ZHAO YOU-CAI, et al. Natural reduction of ammonia-N in leachate of large-scale landfill [J]. Acta Scientiae Circumstantiae, 2001, 21(3):323-327.

[23] ARTIOLA F J, FULLER W H. Humic substances in landfill leachates: Humic acid extraction and identification[J]. J Environ Qual, 1982, 11:663-669.

[24] 张徵晟. 生活垃圾焚烧厂渗滤液处理工艺的研究[D]. 上海:同济大学环境工程与科学学院硕士学位论文,2006.

[25] 胡晨燕. 生活垃圾焚烧厂渗滤液物化处理的工艺与机理研究[D]. 上海:同济大学环境工程与科学学院博士学位论文,2006.

[26] 边文骅. 腐植酸形成的生物学机理研究概况[J]. 河北师范大学学报:自然科学版, 2000(4):1-5.
BIAN WEN-HUA. The survey of study of the biological mechanism of humic acid formation[J]. Journal of Hebei Normal Univesity:Natural Science, 2000(4):1-5.

[27] 李吉进，郝晋珉,邹国元,等. 高温堆肥碳氮循环及腐殖质变化特征研究[J]. 生态环境, 2004, 13(3):332-334.
LI JI-JIN, HAO JIN-MIN, ZOU GUO-YUAN, et al. Carbon and nitrogen circulation and humus characteristics of high-temperature composting[J]. Ecology and Environment, 2004, 13(3):332-334.

[28] 楼紫阳，赵由才. 渗滤液处理处置技术与工程实例[M]. 北京:化学工业出版社,2006.

(编辑 张 苹)

(上接第76页)

参考文献:

[1] 王永强,律方成. 基于粗糙集理论和贝叶斯网络的电力变压器故障诊断方法[J]. 中国电机工程学报,2006,26(8):137-141.
WANG YONG-QIANG, LU FANG-CHENG. Synthetic fault diagnosis method of power transformer based on rough set theory and Bayesian network [J]. Proceedings of the CSEE, 2006, 26(8):137-141.

[2] 卢新元，张金隆. 基于粗糙集和贝叶斯理论的IT项目风险规则挖掘[J]. 计算机工程与应用,2006,22:12-15.
LU XING-YUAN, ZHANG JIN-LONG. A method of risk rule mining in IT project based on rough set and bayes theory [J]. Computer Engineering and Applications,2006,22:12-15.

[3] 文琪，彭宏. 基于粗糙集和贝叶斯分类器的病毒程序检测[J]. 西南交通大学学报,2005,40(5):659-662.
WEN QI, PENG HONG. Virus detection based on rough set and bayes classifier [J]. Journal of Southwest Jiaotong University, 2005,40(5):659-662.

[4] 朱永利,吴立增. 贝叶斯分类器与粗糙集相结合的变压器综合故障诊断[J]. 中国电机工程学报,2005,25(10):159-165.
ZHU YONG-LI, WU LI-ZENG. Synthesized diagnosis on transformer faults based on bayesian classifier and rough set [J]. Proceedings of the CSEE, 2005, 25(10):159-165.

[5] 代劲，胡峰. 不完备信息系统下的不确定性度量方法[J]. 计算机应用,2006,26(1):198-201.
DAI JIN, HU FENG. Measurement for the uncertainty of incomplete information system [J]. Computer Applications, 2006,26(1):198-201.

[6] 赵翔，刘同明. 不完备信息系统中基于加权联系度的粗集模型拓展[J]. 计算机应用,2005,25(4):824-826.
ZHAO XIANG, LIU TONG-MING. Extension of rough set model based on weighted connection degree in incomplete information systems [J]. Computer Applications, 2005,25(4): 824-826.

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

[8] 张文修,吴伟志. 粗糙集理论与方法[M]. 北京:科技出版社,2006.

[9] 刘清. Rough集及Rough推理[M]. 北京:科技出版社,2001.

[10] PAWLAK Z. Rough sets approach to knowledge - based decision support[J]. European Journal of Operational Research, 1997 ,99: 48-57.

[11] ZDZISLWA PAWLAK. Rough sets and intelligent data analysis[J]. Information Sciences, 2002, 147:1-12.

[12] ZDZISLWA PAWLAK. Rough sets, decision algorithms and bayes' theorem[J]. European Journal of Operational Research, 2002,136: 181-189.

[13] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社,2000.

[14] 胡彧,李智玲. 一种基于区分矩阵的属性约简算法[J]. 计算机应用,2006,26:80-82.
HU YU, LI ZHI-LING. An attribution reduction method based on discernibility matrix [J]. Computer Applications, 2006,26:80-82.

(编辑 侯 湘)

重庆大学期刊社信息网 (http://qks.cqu.edu.cn/)