一种自适应求三枝决策中决策阈值的算法

贾修一[1,2],李伟▮[3],商 琳[1,2],陈家骏[1,2]
(1.南京大学软件新技术国家重点实验室,江苏南京 210093; 2.南京大学计算机科学与技术系,江苏南京 210093;
3.南京航空航天大学高新技术研究院,江苏南京 210016)

摘 要: 在三枝决策粗糙集模型中,基于贝叶斯决策理论,在给定的损失函数基础上可以计算出不同决策之间的阈值,从而可以推导出各种现有的概率型粗糙集模型,如可变精度粗糙集模型等.但是决策粗糙集模型需要对损失函数预先设定,这就需要合适的先验知识.本文通过研究三枝决策粗糙集模型中的风险损失和建立模型需要的阈值参数之间的关系,提出了一个最优化问题,给出了理论分析,说明解决该优化问题即可求得所需参数,并给出了一种自适应求阈值参数的算法.该算法将每个样本的条件概率作为搜索空间,以决策风险损失最小化为目标,求得的损失函数和阈值能够使得用户基于此作出的风险最小.在部分数据集上的实验也表明了算法的有效性,利用学习到的阈值建立的三枝决策粗糙集模型能够取得更好的分类性能.

关键词: 三枝决策粗糙集; 损失函数; 阈值; 最优化问题
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2011) 11-2520-06

An Adaptive Learning Parameters Algorithm in Three-Way Decision-Theoretic Rough Set Model

JIA Xiu-yi[1,2], LI Wei-wei[3], SHANG Lin[1,2], CHEN Jia-jun[1,2]
(1. *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu* 210093, *China*;
2. *Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu* 210093, *China*;
3. *Academy of Frontier Science, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu* 210016, *China*)

Abstract: Three-way decision-theoretic rough set model is a probabilistic extension of the algebraic rough set model. The required parameters for defining probabilistic lower and upper approximations are calculated based on cost functions through Bayesian decision procedure. Through providing different cost functions, decision-theoretic rough set model can derive many other probabilistic rough set models, such as variable precision rough set model, etc. This paper constructs an optimum problem based on decision-theoretic rough set model. Through solving the optimum problem, one can get the proper cost functions and thresholds without any preliminary knowledge. An adaptive learning parameters algorithm is also proposed to solve the optimum problem. The search space of the algorithm is the set of all instances' probabilities. Under the three-way decision-theoretic rough set model which is based on the learned cost functions and thresholds, the decision cost is minimal and a better classification performance can be gotten from that. The experimental result on some data sets shows the efficiency of our algorithm.

Key words: three-way decision-theoretic rough set model; cost functions; thresholds; optimum problem

1 引言

粗糙集理论作为一种处理不确定性问题的方法,在数据挖掘,机器学习等领域得到了广泛的应用[1~4].Pawlak 的经典粗糙集模型判断一个对象 x 是否属于某一类 X 时,要求对象 x 所属等价类 $[x]$ 的所有对象都要属于 X,也就是 $p(X|[x])=\frac{|X\cap[x]|}{|[x]|}=1$.该模型对噪音数据非常敏感,泛化能力不强.为了解决这个问题,概率型粗糙集模型作为对 Pawlak 粗糙集模型的扩展,很多学者对其进行了研究,提出了如 0.5 概率粗糙集模型[5],决策粗糙集模型[6],可变精度粗糙集模型[7]和 Slezak 提出的贝叶斯粗糙集模型[8]等.这些模型放宽判断 $p(X|[x])=1$ 的条件,改成 $p(X|[x])\geq\alpha$ 这种只要满足一定参数或者阈值的形式.Yao 在文献[9,10]中对于这些概率型粗糙集模型进行了分析,并给出了只要设定合适的损失函数值,决策粗糙集都能推导出现有的这

收稿日期:2011-02-24;修回日期:2011-05-23
基金项目:国家 973 重点基础研究发展规划(No.2010CB327903);江苏省自然科学基金(No.BK2009233)

几种概率型粗糙集模型的结论.决策粗糙集考虑贝叶斯决策理论和假设检验,从理论方面系统的给出了如何在给定损失函数值下计算概率粗糙集模型中的参数值或阈值.同时,Yao 在文献[11]中从三枝决策角度重新审视了决策粗糙集模型,完善了该模型的语义解释.

三枝决策粗糙集模型中一个重要的概念就是损失函数,通过引入损失函数,依据贝叶斯最小风险决策理论,能够计算出决策各个边界的阈值.损失函数的值通常是根据实际情况来给定的,如文献[12]中给出的关于停车场收费的例子,损失函数值是不同停车时段的费用.在很多学习过程中,由于先验知识的缺失,无法获得准确的损失函数值.对于给定的数据,只能依据实验和估计的方法来确定这些值,这可能需要用户多次参与到学习中来,给自动化学习带来一定的困难.在三枝决策粗糙集理论中,目前对于学习损失函数值的研究不是很多,只有 Herbert 从 Game Theory 角度来研究调节损失函数值[13].Herbert 提出了 Game-Theoretic 粗糙集模型,通过对损失函数值的逐步增大或减小来使得所关注的一个或多个目标函数逐渐最优化(如增大正区域等).该方法需要确立优化目标并且在迭代过程中建立多个 payoff 表,并需要用户多次参与到学习当中,这同样要求用户具有合适的领域知识,不利于基于数据的自动化学习.本文给出了一种自适应学习损失函数值和阈值的方法,学习到的损失函数值和阈值只和数据相关,不需要引入其他信息,从而简化了学习过程.

基于贝叶斯最小风险决策理论,每种决策行为都带有相应的风险损失,而通过选择风险最小的行为进行决策,可以计算出决策边界的阈值对(α,β),从而建立三枝决策粗糙集模型.在判断对象 x 是否属于类 X 时,只要通过比较 $p(X|[x])$与阈值 α 和 β 的大小关系,即可作出相应的决策.回顾三枝决策粗糙集模型,我们分析认为风险损失是该模型里最重要的概念,是整个模型的基础,决策时带来的风险损失和预先给定的损失函数,计算出的阈值以及对象 x 属于类 X 的条件概率 $p(X|[x])$等都是相关联的.基于此,我们把整个训练样本集的决策风险损失作为优化目标,构建了一个阈值和风险损失相关的最优化问题,只要解决该最优化问题,就能求得合适的损失函数值和阈值.同时本文给出了一种自适应学习损失函数值和阈值的算法,将每个样本的条件概率作为搜索空间,从中选择合适的概率值作为决策边界阈值对,从而使得基于该阈值对作出的决策风险最小.该算法能够快速的学习到合适的损失函数值和阈值.

2　三枝决策粗糙集模型

信息表 $M=(U,At,\{V_a|a\in At\},\{I_a|a\in At\})$是一四元组,其中有穷集合 U 是对象的论域,At 是所有属性的有穷集合,通常 $At=C\cup D$,C 是条件属性集合,D 是决策属性集合,这类信息表也称之为决策表.V_a 是属性 $a\in At$ 取值的非空集合,$I_a:U\to V_a$ 是从 U 到 V_a 的映射函数.通常 I_a 假设为单值的,任意对象 $x\in U$ 在属性 $a\in At$ 上的取值可以表示为 $I_a(x)$.

在信息表中,等价关系 E 定义如下:

$$E=\{(x,y)\in U\times U|\forall\, a\in At, I_a(x)=I_a(y)\} \quad (1)$$

(U,E)是定义在 At 上的近似空间,对象 x 的等价类可表示为:

$$[x]=\{y\in U|(x,y)\in E\} \quad (2)$$

对于给定数据对象 x,假设 $\Omega=\{\omega_1\cdots,\omega_s\}$是 x 所有可能状态的有穷集合,$A=\{a_1,\cdots,a_t\}$是所有可能动作的有穷集合,$p(\omega_j|x)$表示当对象 x 的状态为 ω_j 的条件概率.设 $\lambda(a_i|\omega_j)$为当对象实际状态为 ω_j 时采取动作 a_i 的损失函数,或简记为 $\lambda_{a_i\omega_j}$.假设对对象 x 采取的动作为 a_i,则该动作所带来的预期风险(损失)可表示为:

$$R(a_i|x)=\sum_{j=1}^{s}\lambda(a_i|\omega_j)\cdot p(\omega_j|x) \quad (3)$$

在近似空间中,等价类$[x]$用来刻画对象 x,状态集合 $\Omega=\{X,X^c\}$用来表示对象是否属于决策类 X,则对象 x 属于 X 和不属于 X 的条件概率分别为:$p(X|[x])=\frac{|X\cap[x]|}{|[x]|}$和 $p(X^c|[x])=1-p(X|[x])$.对对象 x 所有可能的动作集合定义为:$A=\{a_P,a_N,a_B\}$,其中 a_P,a_N,a_B 分别代表将一个对象划分到正区域 POS(X),负区域 NEG(X),边界区域 BND(X)中.正区域表示确定属于 X,负区域表示确定不属于 X,边界区域表示可能属于 X.当对象 x 实际属于 X 时,令 $\lambda_{PP},\lambda_{NP},\lambda_{BP}$ 代表分别采取动作 a_P,a_N,a_B 时的损失函数.反之,当对象 x 实际不属于 X 时,$\lambda_{PN},\lambda_{NN},\lambda_{BN}$代表分别采取动作 a_P,a_N,a_B 时的损失函数.则根据预期风险的公式,对对象 x 采取对应动作时的预期风险为:

$$\begin{aligned}R_P&=R(a_P|[x])=\lambda_{PP}\cdot p(X|[x])+\lambda_{PN}\cdot p(X^c|[x]),\\R_N&=R(a_N|[x])=\lambda_{NP}\cdot p(X|[x])+\lambda_{NN}\cdot p(X^c|[x]),\\R_B&=R(a_B|[x])=\lambda_{BP}\cdot p(X|[x])+\lambda_{BN}\cdot p(X^c|[x]).\end{aligned} \quad (4)$$

依据最小风险准则,贝叶斯决策过程可表示为三枝决策:

(P)如果 $R_P\leqslant R_N$ 并且 $R_P\leqslant R_B$,则 $x\in$POS(X);

(N)如果 $R_N\leqslant R_P$ 并且 $R_N\leqslant R_B$,则 $x\in$NEG(X);

(B)如果 $R_B\leqslant R_P$ 并且 $R_B\leqslant R_N$,则 $x\in$BND(X).

该决策过程的实际意义是,当采取某种动作所带来的风险不超过其他两种动作所带来的风险时,就采取该动作(表现形式为将对象划分到相应区域).

考虑一种特殊情况,假设损失函数满足 $\lambda_{PP}\leqslant\lambda_{BP}<\lambda_{NP}$和$\lambda_{NN}\leqslant\lambda_{BN}<\lambda_{PN}$,其实际意义是:对于一实际属于 X 的对象 x,将其划分到 X 的正区域所带来的风险要小于或等于将其划分到边界区域带来的风险;这两者的风险都小于将其划分到 X 的负区域所带来的风险.同理,对于不属于 X 的对象 x,将其划分到 X 的负区域所带来的风险要小于或等于将其划分到边界区域的风险;这两者的风险都小于将其划分到 X 的正区域所带来的风险.该假设是符合现实意义的.令

$$\alpha=\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})},$$
$$\gamma=\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{NP}-\lambda_{PP})+(\lambda_{PN}-\lambda_{NN})},\tag{5}$$
$$\beta=\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}.$$

则依据损失函数之间的大小关系假设,$\alpha\in(0,1]$,$\gamma\in(0,1)$和 $\beta\in[0,1)$.在文献[9]中 Yao 详细讨论了各损失函数之间的关系及各个模型,本文在此只讨论一种最常用也最具实际意义的情况:基于 $p(X|[x])+p(X^c|[x])=1$ 和等式(5),如果损失函数满足关系$(\lambda_{PN}-\lambda_{BN})\cdot(\lambda_{NP}-\lambda_{BP})>(\lambda_{BP}-\lambda_{PP})\cdot(\lambda_{BN}-\lambda_{NN})$,我们就有 $\alpha>\gamma>\beta$,贝叶斯决策过程(P),(N),(B)可表示为三种规则:

正规则:如果 $p(X|[x])\geqslant\alpha$,则 $x\in\mathrm{POS}(X)$,接受 x 属于 X;

负规则:如果 $p(X|[x])\leqslant\beta$,则 $x\in\mathrm{NEG}(X)$,拒绝 x 属于 X;

边界规则:如果 $\beta<p(X|[x])<\alpha$,则 $x\in\mathrm{BND}(X)$,对 x 是否属于 X 需要进一步的观察.

和经典粗糙集下的三类规则相比,三枝决策粗糙集下的决策规则都是不确定的[11].由于其不确定性,决策的同时也带来了相应的风险.为方便表示,令 $p(X|[x])=p$,则每种规则的风险可表示如下[11]:

正规则的风险:$\lambda_{PP}\cdot p+\lambda_{PN}\cdot(1-p)$,

边界规则的风险:$\lambda_{BP}\cdot p+\lambda_{BN}\cdot(1-p)$,

负规则的风险:$\lambda_{NP}\cdot p+\lambda_{NN}\cdot(1-p)$.

通常在学习过程中设定 $\lambda_{PP}=\lambda_{NN}=0$.

3 决策风险损失最优化问题

为了简化讨论,假定决策表 M 中的论域 $U=\{x_1,\cdots,x_n\}$,决策类只有 2 类:$\{X,X^c\}$.我们可以通过多种理论和方法,如等价类方法,贝叶斯方法等计算得到每个对象 x_i 属于类 X 的概率值,标记为 p_i.依据三枝决策粗糙集模型,对于 $p_i\geqslant\alpha$ 的对象 x_i 采用正规则,对于 $p_j\leqslant\beta$ 的对象 x_j 采用负规则,对于 $\beta<p_k<\alpha$ 的对象 x_k 采用边界规则进行划分.假定 $\lambda_{PP}=\lambda_{NN}=0$,则相应的对整个决策表 M 的每个对象进行划分后所带来的风险损失总和为:

$$R=\sum_{x_i\in\mathrm{POS}(X)}\lambda_{PN}\cdot(1-p_i)+\sum_{x_j\in\mathrm{NEG}(X)}\lambda_{NP}\cdot p_j+\sum_{x_k\in\mathrm{BND}(X)}(\lambda_{BN}\cdot(1-p_k)+\lambda_{BP}\cdot p_k)\tag{6}$$

依据贝叶斯决策理论,该风险损失总和值越小越好.由该公式我们可以构建决策风险损失最优化问题如下:

$$\min_{\alpha,\beta,\gamma}\sum_{p_i\geqslant\alpha}\lambda_{PN}\cdot(1-p_i)+\sum_{p_j\leqslant\beta}\lambda_{NP}\cdot p_j+\varepsilon\cdot\sum_{\beta<p_k<\alpha}(\lambda_{BN}\cdot(1-p_k)+\lambda_{BP}\cdot p_k)\tag{7}$$
$$\text{s.t.}\quad 0\leqslant\beta<\gamma<\alpha\leqslant1,\varepsilon\geqslant1$$

其中 ε 是惩罚因子,用来避免把样本对象过多的划分到边界区域中.

在式(5)中,三个阈值(α,β,γ)由 6 个损失函数值计算得出,我们假定 $\lambda_{PP}=\lambda_{NN}=0$,则剩下的 4 个损失函数值可由式(5)反向推导得出,用阈值(α,β,γ)和 λ_{PN} 表示如下:

$$\lambda_{PN}=\lambda_{PN};\quad \lambda_{NP}=\frac{1-\gamma}{\gamma}\cdot\lambda_{PN};$$
$$\lambda_{BN}=\frac{\beta\cdot(\alpha-\gamma)}{\gamma\cdot(\alpha-\beta)}\cdot\lambda_{PN};\tag{8}$$
$$\lambda_{BP}=\frac{(1-\alpha)\cdot(\gamma-\beta)}{\gamma\cdot(\alpha-\beta)}\cdot\lambda_{PN}.$$

对于所有损失函数的值,我们都可以通过其与 λ_{PN} 的比值及阈值(α,β,γ)的关系求得.假定 $\lambda_{PN}=1$,则最优化问题可重新表示如下:

$$\min_{\alpha,\beta,\gamma}\sum_{p_i\geqslant\alpha}(1-p_i)+\sum_{p_j\leqslant\beta}\frac{1-\gamma}{\gamma}\cdot p_j+\varepsilon\cdot\sum_{\beta<p_k<\alpha}\left[\frac{\beta\cdot(\alpha-\gamma)}{\gamma\cdot(\alpha-\beta)}\cdot(1-p_k)+\frac{(1-\alpha)\cdot(\gamma-\beta)}{\gamma\cdot(\alpha-\beta)}\cdot p_k\right]$$
$$\text{s.t.}\quad 0\leqslant\beta<\gamma<\alpha\leqslant1,\varepsilon\geqslant1\tag{9}$$

至此,整个决策表的风险损失只与阈值(α,β,γ)及每个对象 x_i 的条件概率 p_i 相关,我们可以通过解这个最优化问题来求得合适的阈值(α,β,γ),进而计算出所有的损失函数值.使得学习到的损失函数值和阈值能够使得整个决策表的决策风险损失最小.

4 一种自适应求阈值算法

在上一节中,我们将求阈值和损失函数值的问题转化成了一个最优化问题,本节将给出一个自适应求阈值的算法,使得求得的结果能够近似于最优解.为了给出该算法,我们还需要对上面的最优化问题做进一步的限定.对于所有阈值(α,β,γ),其取值范围在[0,1]之间,因为其值是连续值,无法进行穷举得到最优解,所以限定其搜索空间为决策表中所有对象 x_i 的概率值

所组成的集合,该搜索空间是有穷的.将阈值取为对象的概率值,一是能够和数据相关,二是能够发现那些处于决策边界的对象,便于用户进一步的操作.

该自适应求阈值和损失函数值算法 Alcofa (Adaptive Learning COst Functions Algorithm)的基本思想如下:假定从当前给定的样本 $X=\{x_1,\cdots,x_{i-1}\}$ 学习到的阈值为(α,β,γ),并可计算当前的样本的风险损失总和为 R_X.当新来一个样本 x_i 时,利用其概率值 p_i 和阈值(α,β,γ)计算现在的样本集合 $X'=X\cup\{x_i\}$的风险损失总和 $R_{X'}$,记为 Min_R.然后依次用 p_i 来替代三个阈值(α,β,γ),每次代替都能得到新的阈值(α',β',γ'),重新计算基于新阈值下的当前所有样本的风险损失总和 $R'_{X'}$,如果 $R'_{X'}<Min_R$,则阈值(α,β,γ)更新为(α',β',γ'),否则阈值不变.对下一个样本 x_{i+1}执行同样的步骤,直到所有的样本完成.最后的阈值(α,β,γ)就是我们要求的结果.

该算法步骤中最重要的一步就是用 p_i 来依次替代三个阈值(α,β,γ),举例如下:当用 p_i 替代 α 时,$\alpha'=p_i$,如果 $\gamma<p_i$ 且 $\beta<p_i$,则条件 $\beta<\gamma<\alpha'$ 是满足的,β 和 γ 保持不变,新的候选阈值为(α',β,γ).如果 $p_i\leqslant\gamma$ 或者 $p_i\leqslant\beta$,则条件 $\beta<\gamma<\alpha'$ 不满足,至此可以有多种方法解决,我们采用的方法是也相应的减小 β,γ 值,新的值为 $\beta'=p_i\cdot(1-0.0000005)$ 和 $\gamma'=p_i\cdot(1-0.000001)$,新的候选阈值为$(\alpha',\beta',\gamma')$,也可采用 $\beta'=1-p_i$ 等其他方式.同理,可相应的替代 β 和 γ,必须满足条件 $\beta<\gamma<\alpha$.详细算法描述如下:

Adaptive Learning Cost Functions Algorithm(Alcofa)
Input: training set $X=\{x_1,\cdots,x_n\}$.
Output: three thresholds (α,β,γ).
BEGIN
initialize $\alpha,\beta,\gamma,\chi=\varnothing$, $min=MAXINT$;
FOR each $x_i\in X$
$\chi=\chi\cup\{x_i\}$;
compute the overall cost R_χ based on (α,β,γ);
$min=R_\chi$;
replace α by p_i;
compute the overall cost R'_χ based on (α',β,γ);
IF $R'_\chi<min$ THEN
$min=R'_\chi$;
END IF
replace β by p_i;
compute the overall cost R'_χ based on (α,β',γ);
IF $R'_\chi<min$ THEN
$min=R'_\chi$;
END IF
replace γ by p_i;
compute the overall cost R'_χ based on (α,β,γ');
IF $R'_\chi<min$ THEN
$min=R'_\chi$;
END IF
update the thresholds (α,β,γ) to (α',β',γ') corresponding to the min value;
END FOR
return the current thresholds (α,β,γ);
END BEGIN

该算法的时间复杂度为 $O(n^2)$,只与样本对象个数有关,与属性个数无关.

下面通过一个例子来说明该算法在实际数据集上的应用.我们选用 UCI 数据库[14]中的 Wisconsin Diagnostic Breast Cancer (wdbc) 数据集和 Wisconsin Prognostic Breast Cancer(wpbc)数据集作为训练集,这两个数据集决策类都是 2 类.在求对象是否属于某一类的概率值时我们采用 Naïve Bayes 的方法来计算.决策粗糙集模型在实际使用过程中并没有用到参数 γ,而我们通过式(5)可以得知,$\gamma=\dfrac{\lambda_{PN}}{\lambda_{PN}+\lambda_{NP}}$,对于 2 类问题,不加入领域知识的话,这两类重要性可以认为是相同的,$\lambda_{PN}=\lambda_{NP}$ 是个合理的选择,故我们在实验过程中设置 $\gamma=0.5$. α 初始化为 0.9,β 初始化为 0.1,对于惩罚因子 ϵ,我们也分别实验了不同的值.其具体实验结果如下图所示.

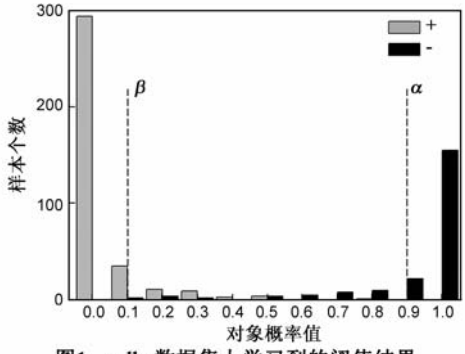

图1　wdbc数据集上学习到的阈值结果

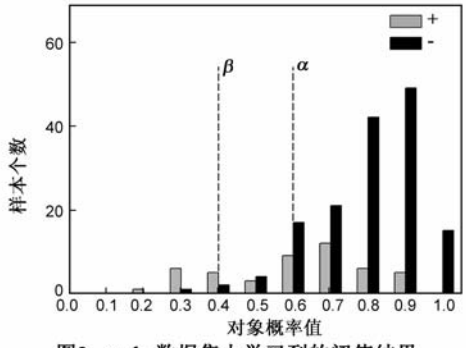

图2　wpbc数据集上学习到的阈值结果

图 1 和图 2 中横坐标表示概率值,纵坐标表示样本的个数,从中我们可以看到,Alcofa 算法学习到的 α,β 值能够将样本分到合适的区域中去.如图 1 所示,对于

那些具有不同类别,概率值相近,和数量少等特点的样本,Alcofa 都将其划分到边界区域中.惩罚因子 ε 我们也实验了从 1 到 50 共 50 个不同的整数值,对于 wdbc 数据集,学习到的 α 的值都为 0.8999,β 的值都为 0.1001.对于 wpbc 数据集,在 $\varepsilon \leqslant 4$ 时,$\alpha = 0.8996$,$\beta = 0.1$;在 $\varepsilon \geqslant 5$ 时,α 的值都如图 2 所示的 0.5999,β 的值都为 0.4001.

为了检验 Alcofa 学习到的阈值是否有效,我们还比较了基于学习到的(α,β)构建的 Naïve Bayes Rough Set (NBRS)分类器[15]和 Naïve Bayes(NB)分类器的分类性能,对于两个数据集都采用 10 倍交叉验证,ε 设定为 20.结果如表 1 所示.

表 1 两种分类器在数据集上的分类性能比较

	wpbc			wdbc		
	p	*r*	*F*	*p*	*r*	*F*
NBRS	0.765	0.895	0.825	0.947	1.0	0.973
NB	0.684	1.0	0.813	0.947	1.0	0.973

从表中的结果可以看出,在 wdbc 数据集上两种分类模型分类能力相同,而在 wpbc 数据集上,基于决策粗糙集模型的 NBRS 分类器在 recall 值上低于 NB 分类器,原因是该模型将部分容易分错的样本都划分到边界区域中去,但同时也提高了 precision 的值,最终的 F 值也要比 NB 分类器的高.可见学习到的阈值是有效的.

5 关于算法的进一步讨论

在本节我们将就风险损失最优化问题和自适应算法中的一些问题做进一步的讨论.

(1)三枝决策粗糙集模型引入了损失函数,并基于损失函数计算得出粗糙集模型所需要的几个阈值.三枝决策粗糙集和经典粗糙集一样,处理的都是以决策表形式存在的数据,并没有限定数据是代价敏感的数据.而损失函数和代价敏感学习中的代价函数,在本质上是一样的,从分类角度来说都是不同分类错误带来的损失,不同的是代价敏感学习中的代价函数只有 2 类错误带来的损失.而三枝决策粗糙集模型由于考虑了粗糙集中的边界区域,则相应的产生了 4 类错误.因此如果数据是代价敏感学习的数据,对于三枝决策粗糙集模型而言,相当于预先给定了部分损失函数值,更加利于建立模型.而对于非代价敏感学习的数据,正如本文所讨论的,完全可以从数据中学习出建立模型所需要的各个参数.

(2)关于学习过程中所需要样本对象的条件概率值,我们可以采用各种方法,并不局限于粗糙集理论中的等价类方法,如本文所采用的就是 Naïve Bayes 方法.本文的算法学习到的阈值是基于分类器提供的概率值,这就要求分类器必须提供较精确的概率值,这点比较困难,很难说哪种分类器好,需要我们进一步研究.对于 Naïve Bayes 分类器,我们知道其求得的概率值本身并不是很精确的,而是比较锐化的,都趋向于 1 或者 0[16].而对于大量概率值趋向于两端的样本来说,通过最优化公式和阈值与损失函数之间关系的公式我们可以得知,在学习时容易将非常极端的值作为学习到的阈值(α,β),从而使得 λ_{BN}和λ_{BP}的值非常小,样本被分到边界区域的损失变得很小,这样大量的样本会被分到边界区域中,而且风险损失总和也比较小.因此需要我们对分到边界区域的样本进行一定的惩罚.为了避免出现这种情况,还可以进一步设定最优化问题的条件,如限定 $0.1 \leqslant \beta < \gamma < \alpha \leqslant 0.9$ 等.对于概率值的精确问题,还可以通过一些映射的方法等得到稍微精确的概率值,如基于 isotonic regression 的 PAV 方法等[16,17].

(3)对于将三枝决策粗糙集模型中的风险损失转化成最优化问题,我们还需对这个最优化公式做进一步的研究,如属于何种优化,可以采用何种解法等等.Alcofa 作为一种自适应学习的算法,只是对这个问题的一个近似的解法,我们还可以使用其他方法进行求解,如遗传算法,模拟退火等其他随机算法来求解.对于是否是最优解或者要学习的阈值是否以区间形式存在而不仅仅是单个值更好等等,都需要我们进一步的研究.

(4)三枝决策粗糙集和普通的决策方法相比,引入了边界决策,这和机器学习中带拒绝的学习类似.不同的是三枝决策对于所拒绝分类的对象给出了语义上的解释和对于拒绝所需的阈值参数给出了理论上的推导.三枝决策应用于分类问题的时候,和带拒绝的学习一样,都会在一定程度上提高分类的准确率,这是因为对于容易分错的对象都将其划分到边界或拒绝区域了,通过引入拒绝率而减小错分率,也就是说以牺牲一部分查全率而提高准确率.在某种程度上这两者可能存在着某种 trade-off 的关系,我们可以通过降低查全率而提高准确率.对于代价敏感学习的任务,如垃圾邮件过滤系统等,引入拒绝率而减小错误率,虽然给用户带来了时间上的消耗,却使得用户最大程度上不错过重要邮件,这是非常合理和有效的.而对于普通的分类问题,如果不需要人工参与,也就是说要求拒绝率为零,则三枝决策可以将边界设定为空,从而变成标准的分类问题.另外也可以融合其他分类方法对三枝决策产生的边界区域对象进行重新分类,以类似集成学习的方式提高总的分类精度,这点也值得我们深入研究.

6 结论

本文针对三枝决策粗糙集模型中的损失函数值和阈值是否能从数据中自动学习做了讨论,通过研究损失函数值和阈值之间的关系,将损失函数值用阈值来

表示,从而直接建立了阈值和风险损失之间的关系,利用这两者之间的关系可以建立一个最优化问题,只要解决风险损失最小化的问题,即可求得相应的阈值,从而求解出合适的损失函数值.为解决这个最优化问题,本文提出了一种自适应学习阈值的算法,通过该算法可以学习出合适有效的阈值.在部分数据集上的实验也表明了算法的有效性,利用学习到的阈值建立的三枝决策粗糙集模型能够取得更好的分类性能.我们将来的工作将从代价敏感学习角度和优化角度来审视三枝决策粗糙集模型,期望得到一些有意义的结果.

参考文献

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

[2] 叶东毅,陈昭炯.一个新的差别矩阵及其求核方法[J].电子学报,2002,30(7):1086 – 1088.
Ye D Y, Chen Z J. A new discernibility matrix and the computation of a core[J]. Acta Electronica Sinica, 2002, 30(7): 1086 – 1088. (in Chinese)

[3] 徐捷,徐从富,耿卫东,潘云鹤.基于粗糙集理论的动态目标识别及跟踪[J].电子学报,2002,30(4):605 – 607.
Xu J, Xu C F, Geng W D, Pan Y H. Dynamic objects identifying and tracing based on rough set theory[J]. Acta Electronica Sinica, 2002, 30(4): 605 – 607. (in Chinese)

[4] 王国胤,姚一豫,于洪.粗糙集理论与应用研究综述[J].计算机学报,2009,32(7):1299 – 1246.
Wang G Y, Yao Y Y, Yu H. A survey on rough set theory and applications[J]. Chinese Journal of Computers, 2009, 32(7): 1299 – 1246. (in Chinese)

[5] Z Pawlak, S K M Wong, W Ziarko. Rough sets: probabilistic versus deterministic approach[J]. International Journal of Man-machine Studies, 1988, 29(1): 81 – 95.

[6] Y Y Yao, S K M Wong. A decision theoretic framework for approximating concepts[J]. International Journal of Man-machine Studies, 1992, 37(6): 793 – 809.

[7] W Ziarko. Variable precision rough set model[J]. Journal of Computer and System Science, 1993, 46(1): 39 – 59.

[8] D Slezak, W Ziarko. The investigation of the Bayesian rough set model[J]. International Journal of Approximate Reasoning, 2005, 40(1 – 2): 81 – 91.

[9] Y Y Yao. Probabilistic rough set approximations[J]. International Journal of Approximate Reasoning, 2008, 49(2): 255 – 271.

[10] Y Y Yao. Probabilistic approach to rough sets[J]. Expert Systems, 2003, 20(5): 287 – 297.

[11] Y Y Yao. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341 – 353.

[12] Y Y Yao, Y Zhao. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356 – 3373.

[13] Joseph P Herbert, J T Yao. Learning optimal parameters in decision-theoretic rough sets[A]. In Proc. RSKT'09[C]. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg, 2009. 610 – 617.

[14] UCI Machine Learning Repository[DB/OL]. http://archive.ics.uci.edu/ ml, 2011-01-15.

[15] Y Y Yao, B Zhou. Naïve Bayesian rough sets[A]. In Proc. RSKT'10[C]. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg, 2010. 719 – 726.

[16] B Zadrozny, C Elkan. Transforming classifier scores into accurate multiclass probability estimates[A]. In Proc. SIGKDD'02[C]. New York: ACM, 2002. 694 – 699.

[17] M Ayer, H Brunk, G Ewing, W Reid, E Silverman. An empirical distribution function for sampling with incomplete information[J]. Annals of Mathematical Statistics, 1955, 26(4): 641 – 647.

作者简介

贾修一　男,1983 年 12 月生于山东日照市,博士研究生,主要研究方向为粗糙集理论与应用、机器学习、自然语言处理.
E-mail: jiaxy@nlp.nju.edu.cn

李伟▮　女,1981 年 10 月生于湖北宜都市,博士研究生,主要研究方向为软件工程、数据挖掘.
E-mail: amy.vivilee@gmail.com

商　琳　女,1973 年 7 月生于河北曲阳县,博士,副教授,主要研究方向为粗糙集理论与应用、机器学习、数据挖掘、人工智能.
E-mail: shanglin@nju.edu.cn

陈家骏　男,1963 年 10 月生于江苏南京市,教授、博士生导师,主要研究方向为自然语言处理、机器翻译、软件工程.
E-mail: chenjj@nju.edu.cn