文章编号: 1006 - 2475(2016)06 - 0097 - 06

用人工鱼群算法自动确定三支决策阈值

胡 盼,秦亮曦,姚洪曼

(广西大学计算机与电子信息学院,广西 南宁 530004)

摘要: 传统的三支决策粗糙集模型需要设置合适的阈值,需要运用该模型的人员具备相关的专业知识和经验,这阻碍了该模型在实际中的应用。针对此不足,本文提出用人工鱼群算法来自动生成阈值,而不需要先验知识。以样本的条件概率作为解空间,以决策风险最小化为目标,利用人工鱼群算法,能有效地从数据中学习到三支决策粗糙集模型所需要的阈值,使得风险损失最小。在部分 UCI 数据集上的实验表明,该算法在运行时间上和利用学习到的阈值构建的分类器的分类性能都明显优于自适应算法。

关键词: 三支决策粗糙集模型; 人工鱼群算法; 阈值; 代价函数

中图分类号: TP181 文献标识码: A doi: 10.3969/j.issn.1006 - 2475.2016.06.020

Applying Artificial Fish Swarm Algorithm to Automatically Determine Thresholds in Three - way Decision - theoretic Rough Set Model

HU Pan, QIN Liang - xi, YAO Hong - man

(School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract: The traditional three - wany decision rough set model needs to set up appropriate threshold. It requires the user of the model to have the relevant professional knowledge and experience, which hinders the application of the model in practice. To solve this problem, the artificial fish swarm algorithm is proposed to generate the threshold automatically, without requiring priori knowledge. Taking the conditional probability of sample as the target, using the artificial fish swarm algorithm, it can effectively learn from the data to the threshold required by the three - wany decision rough set model. It can make the risk loss minimum. The experimental result in part of UCI data sets shows that the algorithm run much faster than the adaptive learning parameters algorithm, and a three - decision - making classifier was built by using the threshold and this classifier can also make classifier better.

Key words: three - way decision - theoretic rough set model; artificial fish swarm algorithm; thresholds; cost function

0 引 言

三支决策的思想是建立在接受、拒绝和不能做出决定的基础上的,是二支决策的扩展[1-4]。三支决策思想在人们的日常生活中起着重要的作用,也被广泛应用到了众多领域和多个学科中。李建林等[5]在垃圾短信过滤中,将待检测的短信分为: 正常短信、可疑短信和垃圾短信; Yao Jingtao 等[6]将三支决策用到医疗支持系统中,做出 3 种决定: 直接肯定、直接否定和需要进一步观察才能做出决定。黄顺亮等[7]将三支决策用到客户细分中,将待细分的客户划分成: 宝贵客户、普通客户和边界域客户。

利用三支决策进行分类时,每种决策行为都要承担相应的风险损失,如何使需要承担的风险损失总和最小化,是当前需要解决的问题之一[8-9]。如何使三支决策中需要承担的决策风险损失最小,这一问题可以通过设置合适的阈值解决。但是这个阈值往往需要由具备相关领域和具有一定相关专业知识的专家设定,从而阻碍了该模型在实际中的应用。贾修一等[10-11]针对该问题提出了自适应算法进行求解,缺点是学习阈值需要花费较长的时间。本文提出一种人工鱼群求解三支决策粗糙集阈值的算法,该算法以

收稿日期: 2015 - 12 - 08

基金项目: 国家自然科学基金资助项目(61363027)

作者简介: 胡盼(1988 -),男,湖北应城人,广西大学计算机与电子信息学院硕士研究生,研究方向: 数据挖掘,粗糙集; 秦亮曦(1963 -),男,教授,博士,研究方向: 数据挖掘,进化计算,管理信息系统; 姚洪曼(1989 -),女,硕士研究生,研究方向: 数据挖掘,进化计算。

样本的概率值作为解空间,以整个训练样本集的所有决策需要承担的风险损失总和最小化为目标函数,首次利用人工鱼群算法求解该问题,该算法比自适应算法更快学习到合适的阈值。利用人工鱼群算法学习到的阈值构建的三支分类器比自适应算法学习到的阈值构建的三支分类器具有更好的分类精度。

1 三支决策粗糙集模型

先简要介绍决策粗糙集理论的基本内容[12-14]。假设 $S=(U,A,V,f)$ 是一个四元组的信息系统,$\Omega=\{\omega_1,\omega_2,\ldots,\omega_m\}$ 表示该状态集合中有 m 种不同状态;$A=\{a_1,a_2,\ldots,a_n\}$ 表示可能进行 n 种不同决策。$P(\omega_i|x)$ 表示当对象 x 处在状态 ω_i 时的条件概率;$\lambda(a_j|\omega_i)$ 表示在状态 ω_i 时,做出决策 a_j 所需要承担的决策代价;对于对象 x,假设做出了决策 a_j 所产生的期望决策代价为:

$$R(a_j|x)=\sum_{i=1}^{m}\lambda(a_j|\omega_i)P(\omega_i|x) \tag{1}$$

对于对象 x,令 $\tau(x)$ 为从对象空间到 A 的一个决策规则函数,$\tau(x)\in A$。令 f 为在给定的 τ 下的总体期望决策代价,可以表示如下:

$$f=\sum_{x\in U}R(\tau(x)|x)P(x) \tag{2}$$

其中 x 的先验概率表示为 $P(x)$,对象 x 在状态为 $\tau(x)$ 时需要承担的决策代价表示为 $R(\tau(x)|x)$。根据贝叶斯决策过程,如果能找到一种决策方案,使 f 达到最小,该方案就是找到的最优的决策方案。

现在考虑一种简化的情况,设只有 2 个状态集和 3 个行动集[3] 的决策粗糙集模型中,状态集 $\Omega=\{X,\neg X\}$;行动集 $\Omega=\{a_P,a_N,a_B\}$ 中有 3 种不同行为,分别是将状态集中的元素划分到正域、负域和边界域。采取以上的不同划分操作必定会带来相应的损失和代价,当 $x\in X$ 时,采取 a_P,a_N 和 a_B 行动时,相应的代价分别为 λ_{PP},λ_{NP}和 λ_{BP};当 $x\in\neg X$ 时,采取 a_P,a_N 和 a_B 行动时,相应的代价分别为 λ_{PN},λ_{NN}和 λ_{BN}。令 $P(X|[x])=p$,因此式(1)采用 a_P,a_N 和 a_B 这 3 种不同行动时需要承担的代价分别为:

$$\begin{cases}R_P=R(a_P|[x])=\lambda_{PP}p+\lambda_{PN}(1-p)\\R_B=R(a_B|[x])=\lambda_{BP}p+\lambda_{BN}(1-p)\\R_N=R(a_N|[x])=\lambda_{NP}p+\lambda_{NN}(1-p)\end{cases} \tag{3}$$

按照贝叶斯决策过程,最优决策方案是需要承担的代价最小的决策方案。于是得出如下 3 条规则:

1) 规则(P):若 R_P 钬 SymbolcB@ R_N 和 R_P 钬 SymbolcB@ R_B 同时成立,则 $x\in POS(X)$;

2) 规则(B):若 R_B 钬 SymbolcB@ R_P 和 R_B 钬 SymbolcB@ R_N 同时成立,则 $x\in BND(X)$;

3) 规则(N):若 R_N 钬 SymbolcB@ R_P 和 R_N 钬 SymbolcB@ R_B 同时成立,则 $x\in NEG(X)$。

以上的决策过程可以理解为:3 种决策中,执行某一种决策的代价不超过执行其他 2 种决策的代价,就选择这种代价最小的决策。因为 X 与 $\neg X$ 互为补集,所以 $P(X|x)\geq\alpha$,由式(3)可知,上述 3 条决策规则只与概率 $P(X|x)$ 与相应的损失函数 λ 有关。通常情况下,当 $x\in X$,将其划分到正域所需承担的代价小于或等于将其划分到边界所需承担的代价,而将其划分到边界域所需承担的代价小于或等于将其划入负域所需承担的代价。当 $x\in\neg X$,将其划到负域所需承担的代价将小于或等于将其划分到边界域所需承担的代价,将其划分到边界域所需承担的代价将小于或等于将其划分到正域所需承担的代价,则可得到合理假设:0 钬 SymbolcB@ λ_{PP}钬 SymbolcB@ λ_{BP}钬 SymbolcB@ λ_{NP},0 钬 SymbolcB@ λ_{NN}钬 SymbolcB@ λ_{BP}钬 SymbolcB@ λ_{PN}。因此,规则(P)、规则(B)、规则(N)所需要的条件可以推导为:

1) 对于规则(P)来说:

$$R_P\text{ 钬 SymbolcB@ }R_B\Leftrightarrow p\geq\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$

$$R_P\text{ 钬 SymbolcB@ }R_N\Leftrightarrow p\geq\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}$$

2) 对于规则(B)来说:

$$R_B\text{ 钬 SymbolcB@ }R_P\Leftrightarrow p<\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$

$$R_B\text{ 钬 SymbolcB@ }R_P\Leftrightarrow p\geq\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

3) 对于规则(N)来说:

$$R_N\text{ 钬 SymbolcB@ }R_P\Leftrightarrow p<\frac{(\lambda_{PN}-\lambda_{NN})}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}$$

$$R_N\text{ 钬 SymbolcB@ }R_B\Leftrightarrow p<\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

此时,令:

$$\begin{cases}\alpha=\dfrac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BN}-\lambda_{PP})}=\left|1+\dfrac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}\right|^{-1}\\\beta=\dfrac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}=\left|1+\dfrac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}\right|^{-1}\\\gamma=\dfrac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}=\left|1+\dfrac{\lambda_{NP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{NN}}\right|^{-1}\end{cases} \tag{4}$$

由规则(B),可以得出 $\alpha>\beta$,那么 $\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$,$\frac{b}{a}>\frac{d}{c}\Rightarrow\frac{b+d}{a+c}>\frac{d}{c}(a,b,c,d>0)$,有 $\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}<\frac{\lambda_{NP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{NN}}<\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}$,由式(5)可以进一步得出:0 钬 SymbolcB@ $\beta<\gamma<\alpha$ 钬 SymbolcB@ 1。这样,规则(P)、规则(B)、规则(N)可以重写为:

1) (P1):若 $P(X|x)\geq\alpha$,则 $x\in POS(X)$;

2) (B1):若 $\beta<P(X|x)<\alpha$,则 $x\in BND(X)$;

3) (N1):若 $P(X|x)<\beta$,则 $x\in NEG(X)$。

2　决策风险最小化问题

由式(4)可知,由 λ_{PP}、λ_{BP}、λ_{NP}、λ_{NN}、λ_{BN}和 λ_{PN}可以计算出阈值 α、β、γ。在一般情况下,令进行正确分类时需要承担的决策代价为 0,即 $\lambda_{PP}=\lambda_{NN}=0$。这样 3 个阈值 α、β、γ 就只与 λ_{BP}、λ_{NP}、λ_{BN}和 λ_{PN}这 4 个损失函数有关。把式(4)进行反向推导,用阈值 α、β、γ 和 λ_{PN}将其表示如下:

$$\begin{cases}\lambda_{PP}=\lambda_{NN}=0\\ \lambda_{NP}=\dfrac{1-\lambda}{\lambda}\lambda_{PN}\\ \lambda_{BN}=\dfrac{\beta(\alpha-\gamma)}{\gamma(\alpha-\beta)}\lambda_{PN}\\ \lambda_{BP}=\dfrac{(1-\alpha)\cdot(\gamma-\beta)}{\gamma(\alpha-\beta)}\lambda_{PN}\end{cases}\tag{5}$$

为了简化讨论,只考虑一种常见的情况,假设决策表 S 的论域 $U=\{x_1,x_2,\ldots,x_n\}$,其决策类只有 2 类:$\{X,\neg X\}$。记 $x\in X$ 的条件概率为 p_i,可以通过等价类方法或贝叶斯方法计算出 p_i。设 $\lambda_{PP}=\lambda_{NN}=0$,那么整个决策表 S 进行决策时需要承担的风险损失代价总和为:

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

由贝叶斯决策理论可知,该函数的值越小越好。假设 $\lambda_{PN}=1$,由式(5)可知,式(6)可推导为:

$$f=\sum_{p_i\geq\alpha}(1-p_i)+\sum_{p_j\text{钬}SymbolcB@\beta}\frac{1-\gamma}{\gamma}\cdot(1-p_j)+\varepsilon\cdot\sum_{\beta<p_k<\alpha}\left[\frac{\beta\cdot(\alpha-\gamma)}{\gamma\cdot(\alpha-\beta)}\cdot(1-p_k)+\frac{(1-\alpha)\cdot(\gamma-\beta)}{\gamma\cdot(\alpha-\beta)}\cdot(1-p_k)\right]\tag{7}$$

为了防止过多的样本被划分到边界域,引入了惩罚因子 ε。令 $\varepsilon\geq1$。其中 0 钬 SymbolcB@ $\beta<\gamma<\alpha$ 钬 SymbolcB@ 1。此时,整个决策表所需承担的总风险损失只由阈值(α,β,γ)和每个对象 x_i 的条件概率 p_i 决定。这样,求解最优阈值的问题可以转化成求解决策风险最小化问题。

3　基于人工鱼群算法求三支决策粗糙集阈值

3.1　基本人工鱼群算法

人工鱼群算法是由李晓磊等[15]提出的。该算法模拟自然界中鱼的几种常见行为:觅食、聚群和追尾行为以及鱼群之间相互传递信息和相互协作等行为的一种有效寻优算法。该算法鲁棒性强,且对初值和参数的选择不敏感,具有良好的取得全局最优能力,还具有简单,不需要复杂的编程即可实现等优点。人工鱼群的数学模型描述如下:

设 $X=(x_1,x_2,\ldots,x_n)$ 表示鱼群,x_i($i=1,2,\ldots,n$)为人工鱼个体,$Y=f(x)$ 表示个体所在位置的食物浓度,该函数即为目标函数。人工鱼个体之间的距离 $d_{ij}=Distance(x_i,x_j)$,设人工鱼个体的视野范围为 Visual,每条人工鱼的最大移动步长为 Step,鱼群的拥挤度因子为 δ,种群规模 Total,人工鱼的最大试探次数为 Try_number。

3.2　鱼群行为描述

1) 觅食行为:人工鱼个体 x_i 在其视野范围 $Visual_i$ 内随机选择一个位置 x_j,比较 $f(x_i)$ 和 $f(x_j)$ 这 2 个位置上的食物浓度,在求解最小值问题时,若 $f(x_j)<f(x_i)$,则人工鱼 x_i 向着人工鱼 x_j 位置前进一步,可以直接移动到 x_j,以加快搜索速度。反之,重新随机选择新的位置 x_j。若经过 Try_number 次的反复尝试,仍找不到符合条件的位置,则采取 $x_{next}=x_i+rand(Step)$ 的随机游动行为,模拟自然界中的鱼的随机行为。

2) 聚群行为:自然界中,鱼在游动的过程中会自然地聚集成群,这样可以躲避危险和保证群种的生存。在人工鱼群算法中的聚群行为也是尽量朝着附近的伙伴的中心位置移动,还要避免过分拥挤。设人工鱼个体 x_i,在其视野范围 $Visual_i$ 内(即 $d_{ij}<Visual_i$)的伙伴数量为 n_f 及其中心位置 x_c。若 $y_c\cdot n_f<\delta\cdot y_i$,则该中心位置不太拥挤且满足较少食物的条件,则人工鱼个体 x_i 朝着 x_c 中心位置前进一步。

3) 追尾行为:人工鱼个体 x_i 在其视野范围内搜索最优的伙伴 x_{min}。若 $f(x_{min})<f(x_i)$ 且假设该人工鱼个体 x_i 朝着 x_{min} 位置游动,不会造成拥挤,则人工鱼个体 x_i 朝着 x_{min} 位置前进一步,否则执行觅食行为。

4) 随机行为:设人工鱼个体 x_i,随机选择一个新的位置 x_j,并向着新的位置游动。

5) 公告板:记录鱼群中最优人工鱼个体的位置和状态等信息,每次行动结束后,每条人工鱼都会将公告板上记录的最优信息和自身的信息进行对比。如果发现自身的信息比公告板状态的更优,则会用自身的信息去更新公告板的信息。

3.3　人工鱼群算法中距离表示

在计算组合优化问题时[16],人工鱼群算法中的距离往往需要重新进行定义,定义的方法不尽相同。设解空间为样本集中所有样本的概率值集合 $P=\{p_1,p_2,\ldots,p_n\}$,先采用 Naïve Bayes 分类器对样本集中的所有样本进行学习,得到其概率值后,再用 isotonic regression 的 PAVA 算法映射得到稍微精确的概率值作为解空间 Q。

人工鱼个体所在的位置是由阈值(α,β,γ)(限定 0.1 钬 SymbolcB@ $\beta<\gamma<\alpha$ 钬 SymbolcB@ 0.

9[10]) 在解空间 Q 中对应的索引唯一确定的,表示为 (pos_α,pos_β,pos_γ) 。设当前有 2 条人工鱼个体 x_1 和 x_2,当 x_1 的阈值为(α_1,β_1,γ_1) 时,其位置表示为 (pos_α_1,pos_β_1,pos_γ_1) ;同样,当 x_2 的阈值为(α_2,β_2,γ_2) 时,其位置表示(pos_α_2,pos_β_2,pos_γ_2) 。那么,x_1 与 x_2 之间的距离可表示如下:

$$Distance(A,B) = |pos\_\alpha_1 - pos\_\alpha_2| + |pos\_\beta_1 - pos\_\beta_2| + |pos\_\gamma_1 - pos\_\gamma_2| \quad (8)$$

3.4 用人工鱼群算法求三支决策最优阈值算法步骤

输入: 每个对象的条件概率值。

输出: 能使得整个决策表的决策风险损失代价函数 $f_{cost} = f(\alpha,\beta,\gamma)$ 取得最小值,以及其对应的阈值对 (α_{best},β_{best},γ_{best}) 。

Step1 初始化人工鱼群各参数。设鱼群的种群规模 Total,最大迭代次数 N,人工鱼的视野范围 Visual,人工鱼的最大移动步长 StepMax,拥挤度因子 δ 等参数。当前迭代次数 passed_times = 0,生成 Total 个人工鱼群个体,对鱼群初始化,每条人工鱼随机生成阈值对(α_i,β_i,γ_i) 。

Step2 公告板初始化: 计算当前鱼群中所有个体中的信息值,并将最小值记录于公告板,同时也记录下相应人工鱼个体的位置和阈值等状态信息。

Step3 聚集行为: 设人工鱼个体当前的状态 x_i,在其视野范围 $Visual_i$ 内(即 $d_{ij} < Visual_i$) 的伙伴数量为 n_f 及中心位置 x_c。若 $y_c \cdot n_f < \delta \cdot y_i$,则表明中心位置有较少食物且不太拥挤,该人工鱼个体朝着伙伴的中心位置前进一步。否则,转 Step5。

Step4 追尾行为: 人工鱼个体 x_i 在其视野范围内搜索最优的伙伴 x_{min}。若其视野范围内不拥挤且 x_{min} 位置食物浓度 $f(x_{min}) < f(x_i)$,则向着 x_{min} 位置前进一步,否则,转 Step5。

Step5 觅食行为: 人工鱼个体 x_i 在其视野范围 $Visual_i$ 内随机选择一个位置 x_j,当该位置上的食物浓度 $f(x_j) < f(x_i)$ 的时候,当前人工鱼个体向着 x_j 前进一步。反之,重新随机选择新的位置 x_j。若经过 Try_number 次的反复尝试,仍不能满足前进条件,则用该人工鱼的当前状态更新公告板。

Step6 终止条件判断。若 passed_times ≥ N,转 Step7; 否则,passed_times = passed_times + 1,转 Step3。

Step7 算法结束: 输出公告牌中记录的最优结果,包括代价函数的最小值以及其对应的阈值对。

对于本算法中人工鱼群的一些初始参数的设置,人工鱼群的种群规模 Total = 10,最大迭代次数 N = 20,人工鱼群的视野范围 Visual = 3,人工鱼群的移动最大步长 Step = 2,拥挤度因子 δ = 9,关于该算法中的人工鱼群的参数设置可以参考文献[7]。

4 实验结果与分析

将人工鱼群求三支决策最优阈值算法分别和自适应算法 Alcofa[10]、模拟退火算法求解三支决策阈值[18]从运行时间和学习到的最小风险代价 2 方面进行比较。同时,为了检验利用人工鱼群算法学习到的阈值的有效性,利用 3 种算法学习到的阈值构建了分类器,并计算其准确率 Precision(P) 、召回率 Recall (R) 和 F1 值,计算方法如下:

Preciosn = 系统检索到的相关文件数/相关文件总数

Recall = 系统检索到的相关文件数/系统返回的文件总数

F1 = 2Precision × Recall/(Precision + Recall)

人工鱼群算法和模拟退火算法都是一种随机算法,每次运行的状态和结果可能不一样。对每个数据集都运行 50 次取平均值作为算法的运行结果。先需要对数据集进行预处理,对于数据集中的 missing 值,采取的是直接删除的方法。

实验环境如下: CPU 是 Intel 的 I3 - 230M,主频 2.3 GHz,4 G 的内存,64 位的 Windows7 系统,3 种算法在 Matlab R2012a 上实现。实验的数据集为 UCI 数据库中的 10 个数据集。

表 1 3 种算法的运行时间结果

	样本个数	运行时间/s		
		自适应算法	模拟退火算法	人工鱼群算法
wdbc	569	11.1	1.7	0.6
agaricus	8124	103.2	2.4	0.8
wpbc	192	3.2	2.1	0.6
monks - 1	432	5.4	1.9	0.5
monks - 2	432	3.4	2.2	0.7
monks - 3	432	2.7	2.1	0.7
transfusion	748	8.0	1.8	0.6
credit	690	7.9	1.8	0.7
ionosphere	352	3.1	1.3	0.7
bank	45212	1229.0	2.8	2.0

3 种算法的运行时间结果如表 1 所示。从表 1 来看,本文提出的人工鱼群算法在大部分数据集上要明显快于自适应算法,略快于模拟退火算法。并且,随着数据集中样本数量的增加,人工鱼群算法运算时间无明显增加。由于自适应算法 Alcofa 是一种迭代算法,其时间复杂度为 $O(n^2)$,其运算时间与随着样本个数的提高而明显增加。人工鱼群算法与模拟退火算法都是一种随机优化算法,运算时间与参数的设置相关,如果参数设置得当,运算时间比较快。对于模拟退火算法的参数设置,读者可以参阅文献[8]。

从表 2 中 3 种算法的决策风险分析的结果看,人工鱼群算法在大部分数据集上取得较小的决策风险代价。仅仅通过比较基于学习到阈值所计算得到的决策风险代价的大小,来判断学习到的阈值的好坏是

不完备的。

表 2　3 种算法的决策风险代价

	自适应算法	模拟退火算法	人工鱼群算法
wdbc	4.53×10^{-2}	4.53×10^{-2}	1.49×10^{-4}
agaricus	4.42	4.00	2.22
wpbc	9.86×10^{-11}	6.72×10^{-23}	3.51×10^{-23}
monks－1	2.30×10^{-19}	1.31×10^{-60}	1.51×10^{-64}
monks－2	1.08×10^{-11}	2.21×10^{-16}	4.01×10^{-36}
monks－3	3.07×10^{-18}	6.69×10^{-18}	1.19×10^{-32}
transfusion	9.99×10^{-26}	1.28×10^{-61}	1.41×10^{-62}
credit	1.77×10^{-39}	3.57×10^{-41}	5.18×10^{-57}
ionosphere	2.82×10^{-9}	1.28×10^{-13}	5.32×10^{-56}
bank	2.18	1.57×10^{-80}	9.88×10^{-323}

为了检验人工鱼群算法学习到的阈值的有效性，如表 3、表 4 和表 5 所示，本文比较了人工鱼群算法学习到的阈值构建的 Naïve Bayes Rough Set (NBRS) 分类器和普通的 Naïve Bayes(NB) 分类器、自适应算法学习到的阈值构建的 NBRS 分类器和模拟退火算法学习到的阈值构建的 NBRS 分类器的分类性能。

表 3　NB 分类器与人工鱼群算法构建的分类器的性能比较

	NB 分类器			人工鱼群算法		
	P	R	F1	P	R	F1
wdbc	0.920	0.880	0.88	0.939	0.955	0.947
agaricus	0.886	0.878	0.874	0.886	0.878	0.874
wpbc	0.754	0.722	0.734	0.880	0.866	0.872
monks－1	0.667	0.667	0.667	0.995	0.995	0.995
monks－2	0.450	0.669	0.538	0.947	0.924	0.928
monks－3	0.935	0.928	0.928	0.935	0.928	0.928
transfusion	0.708	0.753	0.714	0.939	0.939	0.939
credit	0.765	0.732	0.716	0.971	0.962	0.964
ionosphere	0.629	0.800	0.704	0.923	0.796	0.854
bank	0.865	0.863	0.864	0.997	1.000	0.998

表 4　自适应算法和人工鱼群算法构建的分类器的性能比较

	自适应算法			人工鱼群算法		
	P	R	F1	P	R	F1
wdbc	0.922	0.896	0.909	0.939	0.955	0.947
agaricus	0.886	0.878	0.874	0.886	0.878	0.874
wpbc	0.821	0.799	0.808	0.880	0.866	0.872
monks－1	0.995	0.995	0.995	0.995	0.995	0.995
monks－2	0.891	0.891	0.891	0.947	0.924	0.928
monks－3	0.780	0.824	0.800	0.935	0.928	0.928
transfusion	0.854	0.803	0.808	0.939	0.939	0.939
credit	0.673	0.648	0.656	0.971	0.962	0.964
ionosphere	0.715	0.735	0.725	0.923	0.796	0.854
bank	0.973	0.965	0.965	0.997	1.000	0.998

表 5　模拟退火和人工鱼群算法构建的分类器的性能比较

	模拟退火算法			人工鱼群算法		
	P	R	F1	P	R	F1
wdbc	0.939	0.955	0.947	0.939	0.955	0.947
agaricus	0.886	0.878	0.874	0.886	0.878	0.874
wpbc	0.880	0.866	0.872	0.880	0.866	0.872
monks－1	0.995	0.995	0.995	0.995	0.995	0.995
monks－2	0.933	0.912	0.915	0.947	0.924	0.928
monks－3	0.791	0.845	0.813	0.935	0.928	0.928
transfusion	0.862	0.814	0.819	0.939	0.939	0.939
credit	0.749	0.764	0.755	0.971	0.962	0.964
ionosphere	0.845	0.826	0.829	0.923	0.796	0.854
bank	0.988	0.987	0.987	0.997	1.000	0.998

表 3 为 NB 分类器与人工鱼群算法构建的分类器的性解比较。从表 3 可以看出，人工鱼群算法学习到的阈值构建的分类器除了在 agaricus 数据集上的分类器性能和普通贝叶斯分类器分类性能相同，即 P 值、R 值和 F1 相同外，在其他数据集上的 P 值、R 值和 F1 均明显高于普通的贝叶斯分类器。

表 4 为自适应算法与人工鱼群算法构建的分类器的性能比较。表 5 为模拟退火和人工鱼群算法构建的分类器的性能比较。从表 3 ~ 表 5 中的结果来看，在 UCI 数据库中的 10 个数据集上 NB 分类器和基于 3 种算法学习到的阈值构建的 NBRS 分类器的分类能力的进行对比。在 agaricus 数据集上，NB 分类器和 3 种算法学习到的阈值构建的三支决策分类器分类能力相同。而在其他数据集上，3 种算法学习到的阈值构建的三支决策分类器的 F1 值都比 NB 分类器的 F1 值要高，可以证明人工鱼群算法学习到的阈值是有效的。在大部分数据集上，人工鱼群算法学习到的阈值构建的分类器的 P 值和 F1 值等于或高于自适应算法学习到的阈值构建的分类器的 P 值和 F1 值，也高于模拟退火算法学习到的阈值构建的分类器的 P 值和 F1 值。

通过表 2 可以看到，在数据集 agaricus 上，基于 3 种算法计算得到的决策风险损失的值比较接近。通过表 3、表 5 可以看到，在数据集 agaricus 上，基于 3 种算法学习到的阈值构建的分类器的分类能力相同。再联系式(7)，可以看到学习到的决策风险损失的值和阈值是存在一定联系的，决策风险损失的值越小，阈值就越接近最优解，分类的准确率就越高。

5　结束语

本文重新定义了人工鱼群算法的距离，并提出了基于人工鱼群算法的最优阈值生成算法，该算法无需先验知识。本文将该算法应用到部分 UCI 数据集上学习阈值和计算决策风险损失值，最后将学习到的阈

值构建出三支决策分类器,并检验其分类效率。实验结果表明,通过人工鱼群算法构建的三支决策分类器分类性能有所提高,且运算时间明显快于自适应算法求解三支决策粗糙集阈值。因此可以认为人工鱼群算法在三支决策的阈值学习方面是一个可以选择的算法。

参考文献:

[1] Yao Yiyu. Three – way decision: An interpretation of rules in rough set theory, rough sets and knowledge technology [C] // Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology. 2009: 642 – 649.

[2] Li Huaxiong, Zhou Xianzhong. Risk decision making based on decision – theoretic rough set: A three – way view decision model [J]. International Journal of Computational Intelligence Systems, 2011, 4(1): 1 – 11.

[3] Yao Yiyu. The superiority of three – way decisions in probabilistic rough setmodels [J]. Information Sciences, 2011, 181(6): 1080 – 1096.

[4] 于洪,王国胤,姚一豫. 决策粗糙集理论研究现状与展望 [J]. 计算机学报, 2015, 38(8): 1628 – 1639.

[5] 李建林,黄顺亮. 多阶段三支决策垃圾短信过滤模型 [J]. 计算机科学与探索, 2014(2): 226 – 233.

[6] Yao Jingtao, Azam N. Web – based medical decision support systems for three – way medical decision making with game – theoretic rough sets [J]. IEEE Transactions on Fuzzy Systems, 2015, 23(1): 3 – 15.

[7] 黄顺亮,李建林,王琦. 客户细分的三支决策方法 [J]. 计算机科学与探索, 2014(6): 743 – 750.

[8] Liu Dun, Yao Yiyu, Li Tianrui. Three – way investment decisions with decision – theoretic rough sets [J]. International Journal of Computational Intelligence Systems, 2011, 4(1): 66 – 74.

[9] 张燕平,邹慧锦,赵姝. 基于 CCA 的代价敏感三支决策模型 [J]. 南京大学学报(自然科学版), 2015(2): 447 – 452.

[10] 贾修一,李伟湋,商琳,等. 一种自适应求三枝决策中决策阈值的算法 [J]. 电子学报, 2011, 39(11): 2520 – 2525.

[11] Jia Xiuyi, Tang Zhenmin, Liao Wenhe, et al. On an optimization representation of decision – theoretic rough set model [J]. International Journal of Approximate Reasoning, 2014, 55(1): 156 – 166.

[12] Yao Yiyu. Decision – theoretic rough set models, rough sets and knowledge technology [C] // Proceedings of the 2nd International Conference on Rough Sets and Knowledge Technology. 2007: 1 – 12.

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

[14] 钱进,吕萍,岳晓冬. 决策粗糙集属性约简算法与属性核研究 [J]. 计算机科学与探索, 2014, 8(3): 345 – 351.

[15] 李晓磊,邵之江,钱积新. 一种基于动物自治体的寻优模式:鱼群算法 [J]. 系统工程理论与实践, 2002, 22(11): 32 – 38.

[16] 黄光球,刘嘉飞,姚玉霞. 求解组合优化问题的鱼群算法的收敛性证明 [J]. 计算机工程与应用, 2012, 48(10): 59 – 63, 88.

[17] 李晓磊. 一种新型的智能优化方法——人工鱼群算法 [D]. 杭州: 浙江大学, 2003.

[18] 贾修一,商琳. 一种求三支决策阈值的模拟退火算法 [J]. 小型微型计算机系统, 2013, 34(11): 2603 – 2606.

(上接第 18 页)

参考文献:

[1] 黄强,陶正苏,宋浩,等. 基于 ARM 的 GPRS 远程数据传输模块设计 [J]. 电子器件, 2008, 31(4): 1214 – 1218.

[2] 翟顺,王卫红,张衍,等. 基于 SIM900A 的物联网短信报警系统 [J]. 现代电子技术, 2012, 35(5): 86 – 89.

[3] 左兆辉,孙耀杰,马晓峥,等. 基于 PPI 协议与 SIM900A 的抽油机监控系统 [J]. 仪表技术与传感器, 2014(4): 50 – 52.

[4] 张晶如,邵建华,于笃发,等. 基于 SIM900A 的智能心电监护系统客户端 [J]. 通信技术, 2013, 46(7): 109 – 111.

[5] 朱永辉,白征东,过静珺. 基于北斗一号的地质灾害自动监测系统 [J]. 测绘通报, 2010(2): 5 – 7.

[6] 李成大. Windows 下 TCP/IP 协议分析软件的设计开发 [J]. 计算机应用研究, 2002, 19(2): 133 – 135.

[7] 卢华伟,秦品健,郑锐. 基于 Qt/Qwt 的操作监控系统的设计与实现 [J]. 微计算机信息, 2010, 26(1): 72 – 74.

[8] 霍涛,贾振堂. 基于 STM32 和 SIM900A 的无线通信模块设计与实现 [J]. 电子设计工程, 2014, 22(17): 106 – 110.

[9] 胡爱军,张瑞卿,王聪. 基于 ARM9&Linux 的 AD 转换的实现 [J]. 机械设计与制造, 2011(6): 97 – 99.

[10] 狄辉辉,李京华,刘景桑,等. 基于 Qt/E 的嵌入式实时曲线显示界面设计与实现 [J]. 电子测量技术, 2011, 34(12): 76 – 79.

[11] 陈周国,王胜银,付国晴,等. 基于 LinuxQT 技术的远程监控 GUI 设计 [J]. 通信技术, 2009, 42(12): 234 – 236.

[12] 刘伟民,韩斌,李征. 基于 linux 的数据采集及在 QT 界面的显示 [J]. 微计算机信息, 2008, 24(22): 97 – 99.

[13] 陈琦,丁天怀,李成,等. 基于 GPRS/GSM 的低功耗无线远程测控终端设计 [J]. 清华大学学报(自然科学版), 2009, 49(2): 223 – 225.

[14] 刘爽,史国友,张远强. 基于 TCP/IP 协议和多线程的通信软件的设计与实现 [J]. 计算机工程与设计, 2010, 31(7): 1417 – 1420.

[15] 束长宝,于照,张继勇. 基于 TCP/IP 的网络通信及其应用 [J]. 微计算机信息, 2006, 22(36): 157 – 159.