ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 3,May 2021,pp. 498－508
DOI:10. 16337/j. 1004-9037. 2021. 03. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于属性约简的自采样集成分类方法

李朋飞[1,2]，于　洪[1,2]

(1. 重庆邮电大学计算智能重庆市重点实验室,重庆 400065;2. 重庆邮电大学计算机科学与技术学院,重庆 400065)

摘　要:现有的集成技术大多使用经过训练的各个分类器来组成集成系统,集成系统的庞大导致产生额外的内存开销和计算时间。为了提高集成分类模型的泛化能力和效率,在粗糙集属性约简的研究基础上,提出了一种基于属性约简的自采样集成分类方法。该方法将蚁群优化和属性约简相结合的策略应用在原始特征集上,进而得到多个最优的特征约简子空间,以任意一个约简的特征子集作为集成分类的特征输入,能在一定程度上减少分类器的内存消耗和计算时间;然后结合以样本的学习结果和学习速度为约束条件的自采样方法,迭代训练每个基分类器。最后实验结果验证了本文方法的有效性。

关键词:集成学习;属性约简;蚁群优化;粗糙集;自采样

中图分类号:TP391　　**文献标志码**:A

Self-sampling Ensemble Classification Method Based on Attribute Reduction

LI Pengfei[1,2], YU Hong[1,2]

(1. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: The ensemble learning technology often uses each basic classifier that has been trained to form a complete ensemble system, and the largeness of the ensemble system easily leads to more memory and time. So as to gain the high prediction correctness and low classification time of the ensemble classification model, according to the research of attribute reduction in rough sets, this paper proposes a self-sampling ensemble classification method based on the attribute reduction. This method applies the strategy of combining ant colony optimization and attribute reduction to the original feature data set, and then multiple optimal feature reduction subspaces are obtained. Taking any feature subset after reduction as the feature input of the integrated classifier can reduce the memory usage and classification time of the classifier to some extent. And then each self-sampling method taking the learning results and learning speed of the samples as constraints is combined to iteratively train each base classifier. Finally, the feasibility of the proposed method is further proved by experimental results.

Key words: ensemble learning; attribute reduction; ant colony optimization; rough set; self-sampling

---

基金项目:国家自然科学基金(61876027, 61751312, 61533020)资助项目。
收稿日期:2020-05-23;**修订日期**:2020-11-01

引　言

分类在帮助人类理解事物和理解自然方面起着重要的作用,鉴于这种情况,集成学习可以提供一个有效的学习框架,将多个具有不同优点的子模型组合成一个效果更好的模型。集成学习作为一种高效的分类方法,其性能往往高于单个最优的分类器[1],在故障检测、文本处理和统计学等诸多领域引起了很大关注[2-5]。然而,集成学习方法有两个重要的缺陷。首先,在集成系统中,它需要更多的内存去存储所有学习模型;其次,它需要更多的计算时间来对未标记的样本进行预测打标。在整个集成分类过程中,存储和计算时间随着参与集成的分类器数量的增加而增加。现有的集成技术大多使用经过训练的各个分类器来组成集成系统,这使集成系统不必要地增大,导致了额外的内存开销和计算时间。为了弥补现有分类方法的不足,提高集成学习的预测能力和效率,本文提出了一种基于属性约简的自采样集成分类方法。根据粗糙集理论[6-8],采用属性约简的方法[9]可以得到多个特征约简子空间,基于每个特征子空间都可以得到知识。从属性约简的定义来说,约简的目的就是以较少的特征(属性)来描述这个信息系统[10],而这个较小的系统反映的知识与原来的系统是一样的。所以,基于这种认识,本文认为基于这些属性约简后可以得到多个信息系统来获取分类知识,这正好也体现了集成学习的特点。所以本文方法的基本思路就是找到反映信息系统知识的多个代表性特征子集,从而由这些子集对应形成的子空间形成集成学习中各分类器的输入空间,通过输入空间的减小来提高分类效率。为了实现这种思路,本文采用蚁群算法思想求取多个属性约简集合,从而在包含全局知识的各个特征子空间上获取分类知识。以属性约简得到的多个特征子集作为各个分类器的输入,不仅可以获取到等价的知识,而且还能在一定程度上减少分类器的内存消耗和计算时间;通过以样本的学习结果和学习速度为约束条件的自采样方法[11],在每次迭代过程中从所有训练数据中过滤掉噪声点,使模型的训练过程保持平滑,以此提升集成模型的分类效率和性能。

1　预备知识

1.1　属性约简

属性约简也称特征选择,在粗糙集和知识获取的研究工作中占有重要地位。属性约简算法从代数观[12-13]和信息论[14]两个角度出发,通过简化知识表示过程来获取实用规则。本文主要采用基于可辨识矩阵的属性约简算法,同时结合蚁群优化策略得到多个属性约简子集。

定义1(可辨识矩阵) 给定一个知识表达系统$S=(U,C\cup D,V,f)$,论域为U,条件属性集$C=\{a_1,a_2,\cdots,a_m\}$,决策属性集$D=\{d\}$,V为属性值的集合,f为具有映射关系的信息函数。关于知识表达系统S的可辨识矩阵$\boldsymbol{M}$的各个元素$M_{i,j}$可被表示为

$$M_{i,j}=\begin{cases}\{a|a\in C\wedge a(x_i)\neq a(x_j)\} & d(x_i)\neq d(x_j)\\ \varnothing & \text{其他}\end{cases} \tag{1}$$

表1是一个知识表达系统,根据式(1)可得到该知识表达系统对应的可辨识矩阵,如表2所示。

应用基于可辨识矩阵的约简算法对该知识表达系统进行属性约简的推导过程为:$R=(a\vee c\vee d)\wedge(b\vee c\vee d)\wedge(b\vee d)\wedge(a\vee b\vee d)\wedge(a\vee b)\wedge b\wedge(c\vee d)\wedge(a\vee b\vee d)\wedge b\wedge(a\vee b\vee c\vee d)\wedge(b\vee c\vee d)\wedge(b\vee c\vee d)\wedge(c\vee d)\wedge(a\vee b\vee c\vee d)\wedge(a\vee c\vee d)=b\wedge(c\vee d)=(b\wedge c)\vee(b\wedge d)$,由可辨识矩阵得到的约简为:$\{b,c\}$,$\{b,d\}$。

本文需要在关系模型图R-Graph中进行属性约简,因此首先要将其引申为完全图上的约束满足问题(Constraint satisfaction problem, CSP)。

表1 信息表

Table 1 Table of information

Object	*a*	*b*	*c*	*d*	Class
1	0	0	0	0	0
2	1	0	1	1	1
3	1	1	0	0	0
4	0	2	0	1	1
5	1	2	0	0	1
6	1	0	0	0	0
7	1	2	1	1	1
8	0	0	1	1	1

表2 表1的可辨识矩阵

Table 2 Discernibility matrix of Table 1

Object	1	2	3	4	5	6	7	8
1	0							
2	*acd*	0						
3		*bcd*	0					
4	*bd*		*abd*	0				
5	*ab*		*b*		0			
6		*cd*		*abd*	*b*	0		
7	*abcd*		*bcd*			*bcd*	0	
8	*cd*		*abcd*			*acd*		0

定义2(约束满足问题) CSP[15]被定义为一个三元组$<B,D,C>$,有限集$B=\{B_1,B_2,\cdots,B_k\}$,D为B_p和论域之间的关系函数,$D(B_p)$表示变量B_p的论域,C为有限的约束条件集合。

求解约束满足问题的最优解的过程,即是对有限集B中变量满足约束条件集合C的限制条件的最小代价。由属性约简与约束满足问题之间的转化关系可知,辨识矩阵的元素$M_{i,j}$可转化为变量B_p。因辨识矩阵中存在一定冗余元素,可定义吸收算子规则来减小矩阵空间规模。

定义3(吸收算子(&&)) 对于可辨识矩阵$\boldsymbol{M}$,由相异的可辨识矩阵元素组成约简的空间S是$\boldsymbol{M}$的子集,表示为$S=\{B_k\in\boldsymbol{M}|B_k\neq B_s,\forall B_k,B_s\in\boldsymbol{M}\}$。当$B_i\in S,B_j\in S$时,若$B_j\subseteq B_i$,则$B_i\&\&B_j=B_j$,称$B_i$是可被$B_j$吸收的元素。

约简运算是基于可辨识矩阵的布尔运算[16],辨识矩阵中冗余元素的吸收不会改变原有知识表达系统的约简效果,当所有的冗余元素都不存在时,得到的约简空间为最小约简空间(Minimal reduction space, MRS)。

1.2 自采样学习

对于二分类训练数据$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,$y_i\in\{-1,1\}$为$x_i$的标签。自采样学习方法同时关注每个训练样本的学习结果和学习速度[11]。学习结果以损失来表示,而学习速度则以每次迭代中损失的变化来表示。AdaBoost算法主要构建可加性逻辑回归模型[17-18],在每轮迭代中优化指数损失函数并训练一个弱分类器$f(x)$,自采样方法通过在训练弱分类器$f(x)$时增加约束函数来迭代权重c,优化的目标函数为

$$J(c,f,v)=\arg\min\sum_{i=1}^{n}v_i\mathrm{e}^{-y_i(F(x_i)+cf(x_i))}-\alpha v_i\mathrm{e}^{-y_iF(x_i)}-\lambda v_i \tag{2}$$

式中:$\alpha\in[0,1]$为一个平衡参数,λ为一个采样率参数。更新$F(x)$意味着去训练并且添加一个新的加权弱分类器到集成模型中。优化目标函数的实质就是自采样权重v随着弱分类器$f_m(x)$的更新而改变。权重v在每轮迭代的开始被固定,则$\alpha v_ie^{-y_iF(x_i)}-\lambda v_i$是一个常数值,由此式(2)成为一个加权指数损失最小化问题,即有

$$J(c,f)=\arg\min\sum_{i=1}^{n}v_i\mathrm{e}^{-y_i(F(x_i)+cf(x_i))} \tag{3}$$

将式(2)用泰勒公式展开到二次平方项时,因y^2和$f(x)^2$的值为1,式(3)转化为

$$L(F+cf)=\sum_{i=1}^{n}\mathrm{e}^{-y_i(F(x_i)+cf(x_i))}\approx\sum_{i=1}^{n}v_i\mathrm{e}^{-y_iF(x_i)}(1-y_icf(x_i)+c^2y_i^2f(x_i)^2/2)=\sum_{i=1}^{n}v_i\mathrm{e}^{-y_iF(x_i)}(1-y_icf(x_i)+c^2/2) \tag{4}$$

因为$f(x)\in\{-1,1\}$,$\omega_i=\mathrm{e}^{-y_iF(x_i)}$,对于$c>0$,将最优化问题进一步化简为

$$f=\arg\min\sum_{i=1}^{n}v_i\mathrm{e}^{-y_iF(x_i)}(1-y_icf(x_i)+c^2/2)=\arg\min\sum_{i=1}^{n}v_i\mathrm{e}^{-y_iF(x_i)}(y_i-f(x_i))^2=\arg\min\sum_{i=1}^{n}v_i\omega_i(y_i-f(x_i))^2 \tag{5}$$

当式(5)中的$v_i\omega_i$作为样本权重时,最小化准则的二次近似仍能得到$f(x)$的加权最小二乘,因此可以用样本的权值$v_i\omega_i$训练弱分类器得到可加性模型f。在固定c和f后,式(2)是一个关于自采样权重v的优化问题,即

$$v_i=\begin{cases}1 & \alpha(\mathrm{e}^{-y_i(F(x_i)+cf(x_i))}-\mathrm{e}^{-y_iF(x_i)})+(1-\alpha)\mathrm{e}^{-y_i(F(x_i)+cf(x_i))}<\lambda\\ 0 & \text{其他}\end{cases} \tag{6}$$

从式(6)可知,自采样集成方法选择训练样本是基于上一轮的学习结果$\mathrm{e}^{-y_i(F(x_i)+cf(x_i))}-\mathrm{e}^{-y_iF(x_i)}$和最新学习速度$\mathrm{e}^{-y_i(F(x_i)+cf(x_i))}$,此时最新学习速度也相当于新的弱分类器的加权损失。

超参数$\alpha\in[0,1]$在自采样学习中很重要。α设置为0时,它与自定步长学习相似[19,20],α设置为1时,它仅受到学习速度的影响。适当的更新策略是通过设置从0.5到一个更小的值来逐步改变α。对于超参数λ,λ的值越大,训练集中可能是噪声的数据越少,通过在每轮迭代中固定采样比参数$\mu\in(0,1]$来计算λ。在第m轮迭代时,在计算弱分类器权重c_m后,先标准化$\mathrm{e}^{-yF(x)}$和$\mathrm{e}^{-yc_mf_m(x)}$,防止α不能有效控制两种约束的平衡。然后升序排列$L_{\mathrm{sort}}=\mathrm{e}^{-y_iF(x_i)}(\mathrm{e}^{-y_ic_mf_m(x_i)}-\alpha)$的值,选择第$\mu n$个$L_{\mathrm{sort}}$值作为$\lambda_m$的值。这样在训练的早期阶段,分类器就可以关注训练数据中的异类模式,避免在后期迭代中过多关注可能是噪声点的复杂样本。

2 基于属性约简的自采样集成分类模型

本文提出基于属性约简的自采样集成分类方法,首先系统地总结了该分类模型的基本框架,然后分别描述了蚁群优化的属性约简策略和自采样提升的SS-AdaBoost方法,随后给出了模型的完整分类算法。

2.1 模型框架

图1描述了所提模型的基本框架,主要包括两个阶段:第一阶段是蚁群优化的属性约简阶段,其目的是从原始特征集中获取多个约简后的特征子集。根据1.2节可辨识矩阵定义将信息表转化为一个$|U|\times|U|$的可辨识矩阵。吸收算子的定义使得在提高时空效率的同时得到简化后的可辨识矩阵。多数情况下,多个最小属性约简能帮助用户做出更好的决策,启发式蚁群算法结合约束满足问题在关系图模型R-Graph中寻找最小代价路径来解决属性约简问题。第二阶段是集成分类阶段,除了采用上一阶段的任意特征子集作为分类器的特征输入,自采样集成分类在Boosting框架基础上,增加了采样比参数μ、平衡参数α和时效参数δ。为避免传统Boosting算法过多关注误分类样本,没有在样本权重上增加约束,导致可能是噪声的异常数据权重一直提升,发生过拟合现象[17]。自采样集成方法通过以训练样本的学习结果和学习速度为约束条件,每轮迭代更新样本权重w与自采样权重v训练基分类器,准确地去除噪声点,使模型的训练过程保持平滑。

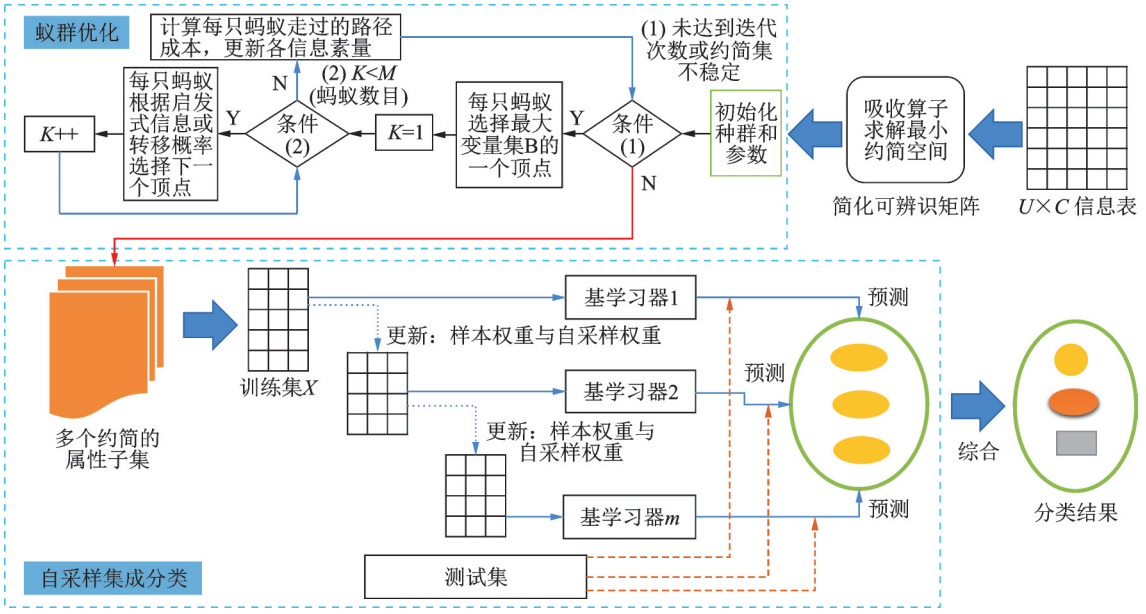

图1 基于属性约简的自采样集成分类模型

Fig.1 Self-sampling ensemble classification model based on attribute reduction

2.2 蚁群优化的属性约简

当属性约简的问题转化为CSP问题，关联CSP(B, Dom, Con)的R-Graph模型$G=(V,E)$就转化为一个完全有向图，顶点V和有向边E分别表示为

$$\begin{aligned} V&=\{<B_i,u>|B_i\in B \text{ and } u\in Dom(B_i)\}\\ E&=\{(<B_i,u>,<B_j,w>)\in V^2|B_i\neq B_j\}\end{aligned} \tag{7}$$

经过顶点的一个完整序列构成$G=(V,E)$中的一条路径，关系图模型R-Graph的目标就是在蚂蚁经过每个变量B_i时，通过选取不同的顶点v使符合约束满足条件情况下的路径代价最小。

在结合R-Graph图的蚁群算法中[21]，蚂蚁将信息素τ_{u_i,u_j}沉淀在图的边(u_i,u_j)上，蚂蚁从蚁穴出发，然后访问图中其余k个顶点，就构建完成一条路径，即一个解。为获得更多的约简集，将蚂蚁分配在具有最多节点的B_i的每个节点之上，通过利用启发信息构造条件转移概率函数为

$$P_{uu'}^{k}=\begin{cases}\dfrac{\tau_{uu'}^{\alpha}(\eta_{u'}+\Delta\eta_{u'})^{\beta}}{\sum\limits_{r\in B_i}\tau_{ur}^{\alpha}(\eta_r+\Delta\eta_r)^{\beta}} & u\in B_i\\ 0 & \text{其他}\end{cases} \tag{8}$$

式中：u为当前B_{i-1}中的值，u'为下一个节点B_i中的值，τ为信息素，η为启发信息，$\Delta\eta_{u'}$为节点u'的启发信息的修正量。α和β分别决定信息素和启发式因子的重要程度，α过大易使蚂蚁选择重复路径，搜索的随机性减弱，β过小易使蚂蚁进入完全随机搜索而无法找到最优解，为保证算法的综合求解性能较好，因而将α与β值分别设置为1和5。符合常取的经验值$\alpha\in[1,5]$,$\beta\in[1,5]$范围之内。

在每轮所有蚂蚁走完各自的路径之后，需要进行一次信息素更新，信息素更新公式为

$$\begin{aligned}&\tau=(1-\rho)\tau_{uu'}+\Delta\tau_{uu'}\qquad u\in B_{i-1},u'\in B_i\\ &\Delta\tau_{uu'}=\sum_{k=1}^{m}\Delta\tau_{uu'}^{k}\qquad \Delta\tau_{uu'}^{k}=\begin{cases}\dfrac{Q}{\text{cost}(SP_k)} & u,u'\in SP_k\\ 0 & \text{其他}\end{cases}\end{aligned} \tag{9}$$

式中:Q为正常数值,ρ为信息素的消退速度,ρ过小则信息素残留过多,蚂蚁很可能再次选择重复路径而得到局部最优解;ρ过大则信息素挥发较快,影响蚂蚁寻找全局最优路径。所以本文结合已有文献将ρ的值设置为0.5。SP_k指蚂蚁k走过的路径,cost(SP_k)表示蚂蚁k走过的路径成本。

2.3 RSS-AdaBoost算法

为提高集成分类算法的效率,本文将自采样的学习方法与属性约简的思想相结合,提出一种基于属性约简的自采样集成分类算法RSS-AdaBoost,通过蚁群优化的属性约简策略可以得到多个近似等效的最优特征子集,将任意特征子集作为自采样集成分类的输入特征,特征空间由原来的$|C|$减小到$|C'|$,在训练基分类器的过程中,特征由$|C|$到$|C'|$的减少数量和参与训练的基分类器数量越多,由约简的特征子集得到的性能提升越明显。本文所提方法过程见算法1。

算法1 RSS-AdaBoost算法

输入:训练数据$\{(x_1,y_1),\cdots,(x_n,y_n)\}$,蚁群寻径轮数$R$,弱分类器个数$N$,采样比$\mu$,平衡因子$\alpha$,时效参数$\delta$

输出:属性约简集合SG和强分类器sign($F(x)$)

(1) 初始化:信息素矩阵$\boldsymbol{E}_p$,边覆盖矩阵$\boldsymbol{E}_c$,信息素挥发速率$\rho=0.5$,随机数$q_0=5$,属性约简子集$SG=\varnothing$,学习模型$F_0(x_i)=0$,样本权重$\omega_i=1/n$,自采样权重$v_i=1$;

(2) 对每一对样本,计算可辨识矩阵元素$M_{i,j}$;

(3) 利用吸收算子过滤掉已在MRS中存在的非空$M_{i,j}$/*MRS为不存在可被吸收项的最小约简空间*/;

(4) 将k只蚂蚁分配在具有最多属性个数的变量集B_1的每个顶点上;

(5) 若B_1顶点上有蚂蚁存在时,开始当前蚂蚁的路径搜索;否则,转到⑩;

(6) 对于每个$B_i\rightarrow B_{i+1}$,随机产生一个常数$q\in(0,10)$,若$q>q_0$,执行(a);否则,执行(b);

(a) 选择B_{i+1}中具有最大启发信息的顶点v^*;

(b) 根据式(8)计算B_{i+1}中每个顶点的转移概率$P_{uu'}^k$,选择具有最大转移概率和最小边覆盖的顶点v^*;

(7) 将⑥中得到的顶点v^*添加到路径SP_k中,该边对应的边覆盖矩阵元素$E_{cuu'}$加1,若顶点v^*是新顶点,更新路径SP_k的路径成本cost(SP_k)。若cost(SP_k)$>$mincost,删除该条SP_k路径,并转到⑤;否则,顶点v^*的启发信息修正量$\Delta\eta_v$减1;

(8) 若当前B_{i+1}是蚂蚁最后一个访问的变量集,转到⑤;否则,转到⑥;

(9) 找到蚂蚁走过的所有路径中具有最小成本路径$SP_{\min}$,若$|SP_{\min}|<|SG|$,则更新属性约简集合$SG\leftarrow SP_{\min}$,最小路径成本mincost$\leftarrow$cost($SP_{\min}$);否则,将$SP_{\min}$中包含但SG中不包含的路径添加到SG中;

(10) 根据式(9)更新信息素矩阵$\boldsymbol{E}_p$,R减1。若属性约简集合SG稳定或者蚁群寻径轮数M减为0,蚁群优化结束,将得到的约简集合SG中任意子集作为输入特征执行后面过程;否则,转到④;

(11) 对于迭代训练$m=1\rightarrow N$,用数据样本权重$v_i\omega_i/\sum_i v_i\omega_i$训练分类器$f_m(x)\in\{-1,1\}$;

(12) 计算分类误差率$err=\sum_{y_i\neq f_m(x_i)}v_i\omega_i/\sum_i v_i\omega_i$和分类器权重$c_m=0.5\times\log((1-err)/err)$;

(13) 根据采样比μ计算采样比率参数λ_m,根据公式$v_i=\begin{cases}1 & e^{-y_i\cdot F_{m-1}(x_i)}(e^{-y_i\cdot c_m\cdot f_m(x_i)}-\alpha)<\lambda_m\\ 0 & \text{其他}\end{cases}$更新$v_i$;

(14) 更新:样本权重$\omega_i=\omega_i\cdot e^{c_m\cdot 1_{y_i\neq f_m(x_i)}}$,学习模型$F_m(x)=F_{m-1}(x)+c_m\cdot f_m(x)$,平衡因子$\alpha=\delta\cdot\alpha$。

2.4 复杂度分析

本节将RSS-AdaBoost算法与文献[11]中的SS-AdaBoost方法在时间复杂度上进行对比分析。SS-AdaBoost方法基于数据集的原始特征空间,即数据样本的所有特征都要作为每个基分类器的特征输入来参与分类。而RSS-AdaBoost算法是基于数据集的特征子空间,即每个基分类器的特征输入是经过属性约简后的特征子集,通过分类器特征输入的特征空间的减小,能在一定程度上减少分类器的内存消耗和计算时间;另一方面,从属性约简的定义来说,约简的目的就是以较少的特征来描述这个信息系统,而这个较小的系统反映的知识与原来的系统是一样的,通过属性约简后得到的多个信息系统来获取分类知识,也正好体现了集成学习的特点。总体时间复杂度与所采用的弱分类器有关,不同的弱分类器计算代价不同。以决策树为例进行分析,假设数据集的实例总数目为$|U|$,实例的属性数目为$|C|$,基分类器数为M。实例权重归一化的复杂度为$O(|U|)$,用决策树训练基分类器时对属性排序的复杂度为$O(|U|\cdot\log_2|U|)$,实例权重在更新时的复杂度为$O(|U|)$,更新样本自采样权重的复杂度为$O(|U|)$,则训练整个强分类器的复杂度为$O((|U|\cdot\log_2|U|+3|U|)\cdot|C|\cdot M)$。

本文方法分为两个阶段。在属性约简阶段,当辨识矩阵包含的元素互不相同时,在R-Graph图中需要进行的决策搜索为$O((|U|^2+|U|)/2)$次。但通常可辨识矩阵中很多元素都是冗余的,很多项都可被吸收,所以该阶段的复杂度远小于$O(((|U|^2+|U|)/2)\cdot|C|)$。在后一阶段,主要区别在于特征空间由$|C|$减小到$|C'|$,训练所有分类器的时间复杂度变为$O((|U|\cdot\log_2|U|+3|U|)\cdot|C'|\cdot M)$。本文方法总的时间复杂度为$O(((|U|^2+|U|)/2)\cdot|C|)+O((|U|\cdot\log_2|U|+3|U|)\cdot|C'|\cdot M)$。从时间复杂度的表现形式上看,当训练强分类器所需的弱分类器数量和特征由$|C|$到$|C'|$的减少数量越多,RSS-AdaBoost算法的效率比SS-AdaBoost越高。

3 实验分析

本文算法采用Python语言在Python 3.7解释器上进行实现,所有实验都在内存为8 GB RAM,CPU频率为1.70 GHz计算机上运行。

3.1 评价指标和数据集

实验采用3个分类指标对本文方法进行性能评估,指标分别为准确率(Acc)、F_1-score和时间。本文所使用的数据集为二分类数据集,表3表示通过真实样本和预测样本的关系所形成混淆矩阵,根据混淆矩阵的形式给出所用评价指标的公式。

表3 混淆矩阵

Table 3 Confusion matrix

样本	预测正例	预测反例
真实正例	TP: True positive	FN: False negative
真实反例	FP: False positive	TN: True negative

分类准确率是指数据样本由分类器预测得到的标签和真实标签相同的数目占总样本数目的百分比。由混淆矩阵得到的分类准确率和F_1-score的表示形式为

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{10}$$

$$\mathrm{precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\quad \mathrm{recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\quad F_1\text{-score}=\frac{2\times\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \tag{11}$$

为了便于实验,本文在UCI的一些真实数据集上验证了本文提出的方法,表4给出了6个UCI数据集的基本信息。

Ionosphere数据集包含351个观察值和34个属性,根据雷达回波数据来预测大气结构特征。Mammographic数据集可根据BI-RADS评估、年龄、形状等特征来预测患者乳腺肿块病变的严重程度。Wdbc数据集通过乳房肿块活检图像显示的细胞核的特征进行病情检验。House-vote数据集来自1984年美国国会投票记录数据库,包括美国众议院代表和国会议员对国会季刊年鉴上确定的16张关键选票的投票。Diabetes是机器学习数据库里的一个糖尿病数据集,主要用来预测人们是否患有糖尿病。Hepatitis是由卡内基梅隆大学的研究所捐赠的肝炎数据集,包含19个属性、155个实例数据,并且某些属性存在缺失值,可根据一系列医学指标来预测生命情况。

表4 UCI数据集信息

Table 4 Information of UCI data sets

Data set	Object	Attribute	Class
Ionosphere	351	34	2
Mammographic	961	5	2
Wdbc	569	31	2
House-vote	435	16	2
Diabetes	768	8	2
Hepatitis	155	19	2

3.2 结果对比

首先比较了已有方法与所提方法在噪声情况下的误差率变化。然后,实验给出了每一个特征子集参与分类的效果,并分别在分类准确率(Acc)、F_1-score和时间(s)3个指标上进行对比实验。为了保证实验效果的稳定性,所有的实验均重复5次,最后给出结果的平均值。

本文所提方法的实验主要包含两个阶段:第一阶段为属性约简阶段,来获取多个约简的特征子集;在此基础之上再进行第二阶段的自采样集成分类。自采样学习的机制有效地平滑了噪声的影响,为了说明在属性约简的情况下,最优特征不妨碍自采样学习方法的抗噪能力,图2给出两种方法的分类错误率随噪声水平增加的动态变化过程。

从图2(a)和(b)中可以看出,两种方法在6个UCI数据集上受噪声影响的误差率变化相似,特别是在Hepatitis和Ionosphere两个数据集上,基于属性约简的集成方法效果更优,这也进一步说明了冗余特征的淘汰不影响自采样学习的抗噪能力。

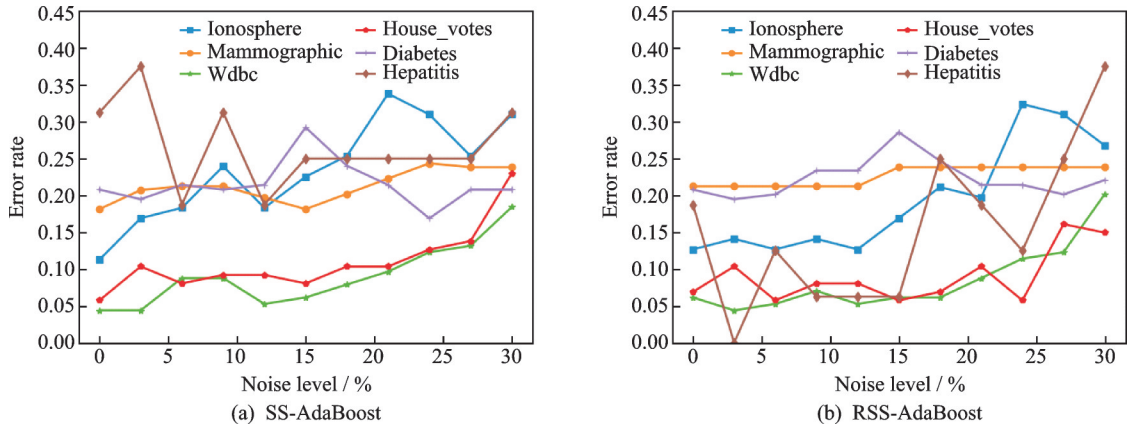

图2 SS-AdaBoost与RSS-AdaBoost方法错误率变化

Fig.2 Error rate variety of SS-AdaBoost and RSS-AdaBoost algorithms

利用蚁群算法优化的属性约简方法,能够得到一个信息表的多个特征约简空间,在参与分类的过程中这些特征集是近似等效的,每一个属性约简集合都是一个最优的特征子集,实验所得的多个属性约简集合结果如表5所示。

表5 属性约简子集

Table 5 Attribute reduction subsets

Data set	Attribute reduction subset
Ionosphere	{3,4,5,7,11,13,16,18,27},{3,4,5,7,13,16,18,21,27},{3,4,5,7,13,16,18,23,27},{3,4,7,13,16,18,21,23,27},{3,4,7,11,13,16,18,25,27},{3,4,7,11,13,16,18,27,29},{3,4,7,13,16,18,21,23,27},{3,4,7,13,16,18,21,27,29},{3,4,7,13,16,18,25,27,29}
Mammographic	{1,3},{1,2}
Wdbc	{2,3,5,6,8,9,14,18,19,23,24,25},{2,4,5,8,9,14,18,19,21,23,24,25},{2,4,8,9,10,14,18,19,21,23,24,25},{2,5,6,8,9,14,15,18,19,23,24,25}
House-vote	{1,2,3,4,5,7,8,9,11},{1,2,3,4,7,8,9,10,11},{1,2,3,5,6,7,8,9,11},{1,2,3,6,7,8,9,10,11}
Diabetes	{1,2,3,4,5,6}
Hepatitis	{3,6,9,15,17,18,19},{3,6,10,15,17,18,19},{6,9,14,15,17,18,19},{6,9,11,15,17,18,19},{6,10,11,15,17,18,19},{6,11,14,15,17,18,19},{6,12,14,15,17,18,19}

在获取到每个数据集的属性约简集合后,为证明所用的蚁群优化方法在获取多个最优属性子集上的可靠性,把任意一个特征子集作为数据集的特征输入进行实验。图3表示了由不同的特征集合参与分类得到的分类效果。

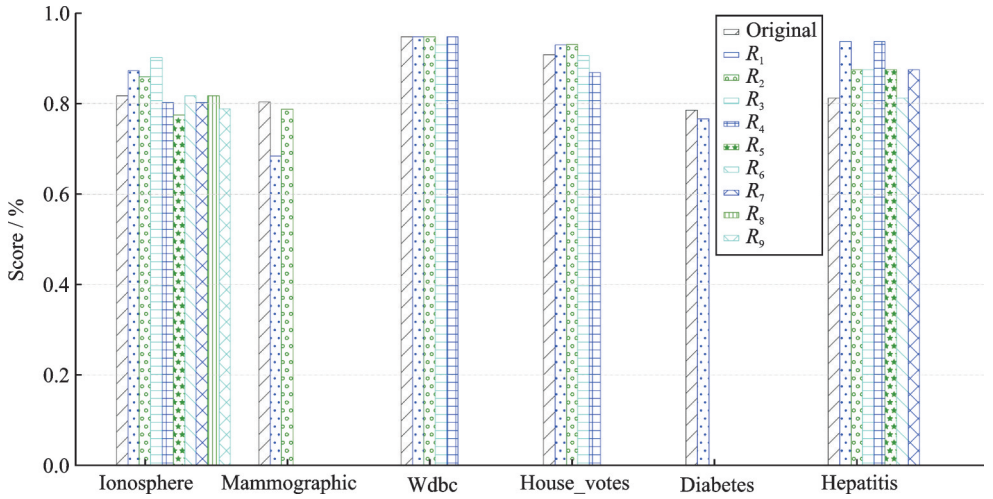

图3 原始属性与各约简属性分类效果比较

Fig. 3 Comparison of classification effects between original attributes and reduced attributes

在图3中,Original代表原始属性集,R_1,R_2,…,R_9分别表示属性约简集合,因为不同数据集获取的特征子集个数不同,图中不同数据集的条形R值可能不同。从图中可以看出:在数据集Ionosphere、Wdbc和House-vote上,大部分经过约简的属性集都具有和未处理时相似的分类精度;在数据集Hepatitis上,所有约简后的属性集分类结果比未处理时要好,这可能是因为在属性约简的过程中过滤掉了样本的干扰特征,避免了集成学习在迭代训练分类器的过程中去拟合这些干扰特征带来的误分类损失。

在比较AdaBoost、SS-AdaBoost与所提RSS-AdaBoost方法在最终实验效果上的差异时,为避免每个约简的特征子集分类性能高低不等,此处采用所有属性约简子集在每个评价指标上的均值作为本文方法的最终结果。实验中所用到的参数,如:采样比参数μ,平衡参数α和时效参数δ的取值按照已有

SS-AdaBoost方法中的取值原则,$\alpha\in[0,0.5]$,$\delta\in[0.9,1]$。除此之外,为了体现3种方法对噪声数据具有鲁棒性,本文中的对比实验均在噪声数据为10%左右进行,同时为保证实验效果的稳定性,所有实验重复执行5次,取平均值进行比较。实验效果如表6所示,在数据集Ionosphere、House-vote和Hepatitis上,本文所提方法具有更好的准确率和F_1-score;虽然在其他3个数据集上的Acc和F_1-score不是最优的,但差别不是很大,比如在Wdbc数据集上与SS-AdaBoost方法的Acc和F_1-score仅差0.004。在时间消耗上,本文方法在6个UCI数据集上均具有最小的计算时间,说明了以任意一个约简的特征子集作为集成分类的特征输入能在一定程度上减少分类器的内存消耗和计算时间,从而进一步验证了基于属性约简的自采样集成分类模型的有效性。

表6 实验结果

Table 6 Experimental results

Data set	AdaBoost			SS-AdaBoost			RSS-AdaBoost		
	Acc	F_1-score	Time/s	Acc	F_1-score	Time/s	Acc	F_1-score	Time/s
Ionosphere	0.802 8	0.798 4	2.24	0.816 9	0.811 4	2.96	0.826 3	0.818 6	2.02
Mammographic	0.777 2	0.777 1	1.53	0.803 1	0.803 2	1.71	0.735 8	0.735 8	1.48
Wdbc	0.921 1	0.920 4	2.15	0.947 4	0.947 1	3.36	0.943 0	0.942 7	1.70
House-vote	0.896 6	0.897 2	1.49	0.908 0	0.908 8	1.78	0.909 1	0.909 4	1.26
Diabetes	0.779 2	0.774 6	1.68	0.785 7	0.781 9	2.02	0.766 2	0.759 9	1.56
Hepatitis	0.687 5	0.719 2	1.08	0.812 5	0.797 2	1.41	0.883 9	0.855 2	0.85

4 结束语

随着大数据和人工智能时代的快速发展,对数据进行有效的分类已成为关键。集成学习是机器学习中的一个热点问题,在数据挖掘和机器学习领域有着巨大的应用潜力。为了提高集成分类模型的泛化能力和效率,本文在粗糙集属性约简的研究基础上提出了一种基于属性约简的自采样集成分类方法。该方法应用蚁群优化和属性约简相结合的策略去得到多个最优的属性约简集,以任意一个约简的特征集作为集成分类的特征输入能在一定程度上减少分类器的内存消耗和计算时间;然后再结合以样本的学习结果和学习速度为约束条件的自采样方法,迭代训练每个基分类器。在6个UCI真实数据集上,将本文RSS-AdaBoost方法与已有分类方法进行了实验对比分析,结果进一步验证了本文所提方法不仅减小了集成学习的规模,而且在分类效率和性能上有一定效果。

参考文献:

[1] SAINI R, GHOSH S K. Ensemble classifiers in remote sensing: A review[C]//Proceedings of 2017 International Conference on Computing, Communication and Automation (ICCCA). [S.l.]: IEEE, 2017: 1148-1152.

[2] GOMES H M, BARDDAL J P, ENEMBRECK F, et al. A survey on ensemble learning for data stream classification[J]. ACM Computing Surveys (CSUR), 2017, 50(2): 1-36.

[3] GALICIA A, TALAVERA-LLAMES R, TRONCOSO A, et al. Multi-step forecasting for big data time series based on ensemble learning[J]. Knowledge-Based Systems, 2019, 163: 830-841.

[4] SUK H I, LEE S W, SHEN D, et al. Deep ensemble learning of sparse regression models for brain disease diagnosis[J]. Medical Image Analysis, 2017, 37: 101-113.

[5] ZHANG Chongsheng, LIU Changchang, ZHANG Xiangliang, et al. An up-to-date comparison of state-of-the-art classification algorithms[J]. Expert Systems with Applications, 2017, 82: 128-150.

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

[7] 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): 1229-1246.
WANG Guoyin, YAO Yiyu, YU Hong. A survey on rough set theory and applications[J]. Journal of Computer Science, 2009, 32(7): 1229-1246.

[8] 张文修. 粗糙集理论与方法[M]. 北京:科学出版社, 2001.
ZHANG Wenxiu. Rough set theory and approach[M]. Beijing: Science Press, 2001.

[9] 于洪, 杨大春. 基于蚁群优化的多个属性约简的求解方法[J]. 模式识别与人工智能, 2011, 24(2): 176-184.
YU Hong, YANG Dachun. Approach to solving attribute reductions with ant colony optimization[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(2): 176-184.

[10] WANG Guoyin, ZHAO Jun, AN Jiujiang, et al. Theoretical study on attribute reduction of rough set theory: Comparison of algebra and information views[C]//Proceedings of IEEE International Conference on Cognitive Informatics. [S.l.]: IEEE, 2004: 148-155.

[11] LIU Xiaoshuang, LUO Senlin, PAN Limin. Robust boosting via self-sampling[J]. Knowledge-Based Systems, 2020, 193: 105424.

[12] HU Xiaohua, CERCONE N. Learning in relational databases: Rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338.

[13] 范敏, 刘文奇, 朱兴东. 基于粗集可辨识矩阵的属性约简算法[J]. 计算机工程与应用, 2004(13): 79-80.
FAN Min, LIU Wenqi, ZHU Xingdong. An approach for attribute reduction based on discernibility matrix of rough set[J]. Computer Engineering and Applications, 2004(13): 79-80.

[14] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.
WANG Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy[J]. Journal of Computer Science, 2002, 25(7): 759-766.

[15] JAMES L. Foundations of constraint satisfaction[J]. Journal of the Operational Research Society, 1995, 46(5): 666-667.

[16] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[C]//Proceedings of Intelligent Decision Support. Dordrecht: Springer, 1992: 331-362.

[17] SABZEVARI M, MARTÍNEZ-MUÑOZ G, SUÁREZ A. Vote-boosting ensembles[J]. Pattern Recognition, 2018, 83: 119-133.

[18] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[C]//Proceedings of 13th International Conference on Machine Learning. Bari: Internatinal Machine Learning Society (IMLS), 1996: 148-156.

[19] PI Te, LI Xi, ZHANG Zhongfei, et al. Self-paced boost learning for classification[C]//Proceedings of 25th International Joint Conference on Artificial Intelligence. New York: AAAI Press, 2016: 1932-1938.

[20] WANG Kaidong, WANG Yao, ZHAO Qian, et al. SPLBoost: An improved robust boosting algorithm based on self-paced learning[J]. IEEE Transactions on Cybernetics, 2019, 51(3): 1556-1570.

[21] KE Liangjun, FENG Zuren, REN Zhigang. An efficient ant colony optimization approach to attribute reduction in rough set theory[J]. Pattern Recognition Letters, 2008, 29(9): 1351-1357.

作者简介:

李朋飞(1993-),男,硕士研究生,研究方向:智能信息处理, E-mail: lipengfei.sun@qq.com。

于洪(1972-),**通信作者**,女,教授,研究方向:粗糙集、粒计算、三支决策、智能信息处理、Web智能、数据挖掘, E-mail: yuhong@cqupt.edu.cn。

(编辑:刘彦东)