文章编号：0253-374X(2022)02-0153-07

DOI：10.11908/j.issn.0253-374x.21349

一种改进的嵌入式特征选择算法及应用

武小军，周文心，董永新

（同济大学 经济与管理学院，上海 200092）

摘要：针对非线性多分类问题，提出了一个改进的嵌入最小－最大值特征选择算法，并与支持向量机算法结合，提出了针对复杂的组合优化问题的启发式算法。为验证方法的有效性，在钢板缺陷识别工程数据集上进行了实验，表明所提出的方法具有较高的求解速度和预测准确度。

关键词：最小－最大值优化问题；特征选择；非线性多分类支持向量机

中图分类号：F253.3　　**文献标志码**：

A Novel Embedded Feature Selection Algorithm and Its Application

WU Xiaojun, *ZHOU Wenxin*, *DONG Yongxin*

(School of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: An improved embedded min-max feature selection algorithm was proposed for the nonlinear multi-label classification problem, and in combination with the support vector machine algorithm, a heuristic algorithm was proposed for the complex combinatorial optimization problem. The efficiency and accuracy of the proposed algorithm were verified after a series of experiments conducted on steel faults diagnosis dataset.

Key words: min-max optimization; feature selection algorithm; nonlinear multi-label support vector machine

近年来，人工智能技术的发展和应用呈现爆发式地增长，利用人工智能技术训练辅助决策模型，协助人类完成各种复杂任务已经成为重要的理论研究和实践应用前沿。在决策模型训练的过程中，大数据集中的特征选择是一个影响人工智能决策模型效果好坏的重要影响因素，良好的特征选择不仅能够提高模型预测的准确度，还能够提升模型训练的速度[1]。例如：Amoozegar和Minaei-Bidgoli[2]提出了改进多目标优化粒子群优化算法，并将该算法应用到有几百个特征的数据集上，结果表明该算法在训练速度上相较于其他算法有明显优势，且预测准确率也令人满意。总体来看，特征选择算法不仅可以应用于监督学习中的回归问题[3]和分类问题[4]，而且可以提高无监督模型预测的准确性[5]，因此大量学者在这个领域进行了深入地探索。

Chandrashekar和Sahin[6]总结认为特征选择算法主要有以下三种方法：过滤法(filter)[7]、嵌入法(embedded)[8]和包装法(wrapper)[9]。过滤法主要应用于数据预处理阶段，该方法将所有的特征按照其得分从高到低进行排序，排除得分较低的特征；嵌入法需要先使用某些机器学习的算法和模型进行训练，得到各个特征的权重，再根据权重来选取特征；包装法让算法自己决定使用哪些特征，是一种特征选择和模型训练同时进行的方法。

嵌入法是一种应用比较广泛的特征选择算法，有许多文献都将其应用到基于支持向量机算法(SVM)的拟合模型构建的研究中。例如：Jiménez-Cordero等[10]提出的嵌入最小－最大值特征选择算法通过搜索策略优化了核函数中的γ，将γ大于0.01的特征保留了下来，形成了最优的特征子集，数值实验表明，该文提出的算法对二分类问题比较有效。又如：Maldonado和López[11]提出了两种类型的罚函数，并基于牛顿法和线搜法提出了改进优化算法来对凹函数进行优化，其方法在绝大部分数据集中，结合SVM方法都较于使用所有特征时模型效果更好。

尽管嵌入特征选择算法已经受到很多学者的关注，但更多的是对二分类问题的研究，对基于多分类问题的最小－最大值特征选择算法的研究还略显不足。本文拟对此问题进行深入探讨，提出一个新的改进算法，并将其应用于钢板缺陷识别工程数据集，

收稿日期：2021-11-26

第一作者：武小军(1977—)，男，副教授，硕士生导师，管理学博士，主要研究方向为管理理论与工业工程。E-mail：wxjun9999@tongji.edu.cn

通信作者：周文心(1998—)，男，硕士生，主要研究方向为机器学习。E-mail：zhouwenxin@tongji.edu.cn

论文拓展介绍

以验证算法的有效性。

1 多分类支持向量机与改进嵌入最小–最大值特征选择算法

1.1 多分类支持向量机

首先简要阐述二分类SVM问题:给定标签分别为+1、−1的数据,通过对已知数据集进行训练,预测未知问题。对每组数据$i\in S$,S代表包含N个数字的集合$\{1,2,..,N\}$,输入特征x_i,$x_i\in R^m$,标签$y_i\in\{+1,-1\}$,y_i是x_i的类标记。已有大量文献详细介绍了SVM[12-14],故本文不再赘述模型的推导。非线性SVM的原始形式如下:

$$\min_{\omega,b,\varepsilon}\frac{1}{2}\|\omega\|_2^2+C\sum_{i\in S}\varepsilon_i \tag{1}$$
$$\text{s.t.}\begin{cases}y_i\left(\omega^{\mathrm{T}}\Phi\left(x_i\right)+b\right)\geqslant 1-\varepsilon_i\\ \varepsilon_i\geqslant 0\end{cases}$$

式中:C为正则项系数,通过非负松弛变量ε对误分类点进行惩罚。Φ是映射函数,将原数据集中的x_i映射到高维空间中使问题线性化。但在绝大部分情况下,Φ没有显式。基于此特点,学者考虑通过引入对偶问题和核函数求解(1)。已有许多文献系统地介绍了原问题的对偶问题[15],故本文不再赘述。原问题的对偶问题如下:

$$\min_{\lambda}\frac{1}{2}\sum_{i\in S}\sum_{j\in S}\lambda_i\lambda_j y_i y_j K(x_i,x_j)-\sum_{i\in S}\lambda_i \tag{2}$$
$$\text{s.t.}\begin{cases}\sum_{i\in S}\lambda_i y_i=0\\ 0\leqslant\lambda_i\leqslant C\end{cases}$$

式中:$\lambda=(\lambda_1,...,\lambda_N)^{\mathrm{T}}$是拉格朗日乘子向量。核技巧(kernel trick),即引入核函数$K(x_i,x_j)=\Phi(x_i)'\Phi(x_j)(R^m\times R^m\to R)$替代$\Phi$,其中$(\cdot)'$表示向量转置,且$K(x_i,x_j)$具有显式。(2)是凸规划问题,因此可以由数学规划软件求得全局最优解。

基于二分类SVM,学者提出多分类SVM,描述如下:给定含N个样本的训练集$X=\{(x_1,y_1),...,(x_N,y_N)\}$,其中,$x_i$是$k$维特征向量,$y_n\in\{1,2,...,M\}$,$n=1,...,N$。Hsu和Lin指出,解决多分类SVM主要有两种方法:one-against-one(一对一)方法和one-against-all(一对多)方法[16]。有学者指出,一对一方法更适合实际应用[16],同时也是LIBSVM库采用的方法[17]。因此,本文选择使用一对一方法求解多分类SVM问题。

一对一方法在每两个类之间均训练一个二分类SVM,故共需训练C_M^2个二分类SVM。对于第i类和第j类数据,训练一个SVM即求解以下二次规划问题:

$$\min_{\omega^{ij},b^{ij},\varepsilon^{ij}}\frac{1}{2}(\omega^{ij})'\omega^{ij}+C\sum_t\varepsilon_t^{ij} \tag{3}$$
$$\text{s.t.}\begin{cases}d_t^{ij}[(\omega^{ij})'\Phi(x_t)+b^{ij}]\geqslant 1-\varepsilon_t^{ij}\\ \varepsilon_t^{ij}\geqslant 0\end{cases}$$

其中,t是第i类和第j类并集的索引。定义$d_t^{ij}=1$,如果$y_t=i$;$d_t^{ij}=-1$,如果$y_t=j$。不加证明地给出式(3)的对偶形式如下

$$\min_{\lambda}\frac{1}{2}\sum_{u\in t}\sum_{v\in t}\lambda_u\lambda_v d_u^{ij}d_v^{ij}K(x_u,x_v)-\sum_{u\in t}\lambda_u \tag{4}$$
$$\text{s.t.}\begin{cases}\sum_{u\in t}\lambda_u d_u^{ij}=0\\ 0\leqslant\lambda_u\leqslant C\end{cases}$$

1.2 改进嵌入最小–最大值特征选择算法

基于Jiménez-Cordero等[10]提出的嵌入最小−最大值特征选择算法,Onel等[18]提出的特征选择算法,提出了一种适用于多分类任务的改进嵌入最小−最大值特征选择算法。引入0~1变量$z_u\in\{0,1\}$,$u\in\{1,2,...,N\}$,当选择特征i时,$z_i=1$,否则$z_i=0$。优化目标除极大化预测的准确率外,还极小化使用的特征数量,从而降低模型训练的成本。将问题(4)改写如下:

$$\min_z\bullet\min_{\lambda}\frac{1}{2}\sum_{u\in t}\sum_{v\in t}\lambda_u\lambda_v d_u^{ij}d_v^{ij}K(x_uz_u,x_vz_v)-\sum_{u\in t}\lambda_u \tag{5}$$
$$\text{s.t.}\begin{cases}\sum_{u\in t}\lambda_u d_u^{ij}=0\\ 0\leqslant\lambda_u\leqslant C\\ z_u\in\{0,1\}\end{cases}$$

为了在预测准确率和模型复杂度之间进行权衡,提出如下的多目标优化问题。

$$\min_z\left[C_2||z||+(1-C_2)\max_{\lambda}\sum_{u\in t}\lambda_u-\frac{1}{2}\sum_{u\in t}\sum_{v\in t}\lambda_u\lambda_v d_u^{ij}d_v^{ij}K\left(x_uz_u,x_vz_v\right)\right] \tag{6}$$
$$\text{s.t.}\begin{cases}\sum_{u\in t}\lambda_u d_u^{ij}=0\\ 0\leqslant\lambda_u\leqslant C\\ z_u\in\{0,1\}\end{cases}$$

其中,C_2为超参数,取值范围位于[0,1]之间。如果C_2趋近于0,则模型以极大化预测准确率为目标,模型的复杂度会上升;如果C_2趋近于1,模型将使用少

量的特征进行训练,会牺牲模型的准确率。

问题(6)的等价问题如下:

$$\min_{z}\left[C_2||z||+(1-C_2)\,\omega\right] \tag{7}$$

$$\text{s.t. } \omega\geqslant\max_{\lambda}\sum_{u\in t}\lambda_u-\frac{1}{2}\sum_{u\in t}\sum_{v\in t}\lambda_u\lambda_v d_u^{ij}d_v^{ij}K\left(x_uz_u,x_vz_v\right)$$

$$\text{s.t. }\sum_{u\in t}\lambda_u d_u^{ij}=0$$

$$0\leqslant\lambda_u\leqslant C$$

$$z_u\in\{0,1\}$$

问题(7)是一个由上层问题和一个下层问题叠加起来的一个双层优化问题。根据Mangasarian和Musicant[19]对拉格朗日对偶性的证明,这里不加证明地给出下层问题的拉格朗日函数。设下层问题的第一个约束对应的拉格朗日乘子为v,第i类和第j类数据一共有a个,约束$\lambda_u\geqslant0$对应的拉格朗日乘子为α_u^0,约束$\lambda_u\leqslant C$对应的拉格朗日乘子为α_u^C。下层问题的拉格朗日函数为

$$L(\lambda,v,\alpha^0,\alpha^C)=\sum_{u\epsilon t}\lambda_u-\frac{1}{2}\sum_{u\in t}\sum_{v\in t}\lambda_u\lambda_v d_u^{ij}d_v^{ij}K\left(x_uz_u,x_vz_v\right) \tag{8}$$

记diag(d^{ij})是一个a行a列,主对角线上是d^{ij}的各个元素,其余位置全是0的矩阵。记$G_Z=$diag(d^{ij})Kdiag(d^{ij}),其中K是核函数。则式(8)可以写为

$$L\left(\lambda,v,\alpha^0,\alpha^C\right)=e'\lambda-\frac{1}{2}\lambda'G_z\lambda-v(d^{ij})'\lambda+(\alpha^0)'\lambda-(\alpha^C)'(\lambda-Ce) \tag{9}$$

式中:e表示分量均为1,且和λ同阶的列向量。由Karush-Kuhn-Tucker,KKT条件可得

$$\frac{\partial L}{\partial\lambda}=e-G_z\lambda-vd^{ij}+\alpha^0-\alpha^C=0 \tag{10}$$

则问题(4)的对偶问题为

$$\min_{\lambda,v,\alpha^0,\alpha^C}e'\lambda-\frac{1}{2}\lambda'G_z\lambda-v(d^{ij})'\lambda+(\alpha^0)'\lambda-(\alpha^C)'(\lambda-Ce) \tag{11}$$

$$\text{s.t.}\begin{cases}\alpha^0\geqslant0,\alpha^C\geqslant0\end{cases}$$

故问题(7)等价于(12):

$$\min_{z}\left[C_2||z||+(1-C_2)\,\omega\right] \tag{12}$$

$$\text{s.t. }\min_{\lambda,v,\alpha^0,\alpha^C}e'\lambda-\frac{1}{2}\lambda'G_z\lambda-v\left(d^{ij}\right)'\lambda+(\alpha^0)'\lambda-(\alpha^C)'(\lambda-Ce)$$

$$\text{s.t. }e-G_z\lambda-vd^{ij}+\alpha^0-\alpha^C=0$$

$$\alpha^0\geqslant0,\alpha^C\geqslant0$$

式(12)目标函数第二项的目标是最小化w,w是(11)目标函数的下界,所以式(12)等价于

$$\min_{z,\lambda,v,\alpha^0,\alpha^C}C_2||z||+(1-C_2)[\min_{\lambda,v,\alpha^0,\alpha^C}e'\lambda-\frac{1}{2}\lambda'G_z\lambda-v\left(d^{ij}\right)'\lambda+(\alpha^0)'\lambda-(\alpha^C)'(\lambda-Ce)] \tag{13}$$

$$\text{s.t.}\begin{cases}e-G_z\lambda-vd^{ij}+\alpha^0-\alpha^C=0\\ \alpha^0\geqslant0\\ \alpha^C\geqslant0\\ 0\leqslant\lambda_u\leqslant C\\ Z_u\in\{0,1\}\end{cases}$$

1.3　搜索策略

本节提出的算法用于求解优化问题(13)。由于使用一对一方法,需要训练C_M^2个SVM,即C_M^2次外层循环。首先,为了避免过拟合,需要划分出训练集和测试集,分别用$\tilde{S}$和S_{test}表示,这个过程会重复N次,同时,需要确定超参数C_2。最优的C_2可以通过枚举不断调试。

其次,将$\tilde{S}$分为训练集和验证集,用S_{train}和S_{val}表示。对于C和γ进行网格搜索(grid search),在S_{train}上求解问题(4),并在S_{val}上计算模型的准确率。搜索完所有C和γ后,给出S_{val}上预测准确率最高的对应的C和γ,且γ是求解问题(11)的初始值。

接下来,使用C和γ在$\tilde{S}$上求解问题(11),返回λ,v,α^0,α^C的初始值。

获取到上述初始值后,在$\tilde{S}$上求解问题(13),得到z^*,λ^*,v^*,$(\alpha^0)^*$,$(\alpha^C)^*$。并计算在S_{test}上的模型准确率。

由于0~1变量的引入,在训练第i类和第j类数据时可以直接得到相应最优特征子序列。但由于总共要训练C_M^2个SVM,因此就不能直观地通过0~1变量的取值选取用于训练的特征。文献[20-22]指出,可以通过信息增益(Information gain, IG)选取最优特征子序列。考虑一个分类系统,以类别C为变量,其可能的取值有$C_1,C_2,\ldots,C_n$,每个类别出现的概率分别为$P(C_1),P(C_2),\ldots,P(C_n)$,其中$n$是类别总数。此时分类系统的信息熵为

$$H(C)=-\sum_{i=1}^{n}P(C_i)\log_2P(C_i) \tag{14}$$

假设T代表特征,记t代表T出现,则根据全概率公式可得

$$H(C|T)=P(t)H(C|t)+P(\bar{t})H(C|\bar{t}) \tag{15}$$

式中,$P(t)$代表特征T出现的概率,$P(\bar{t})$代表特征T不出现的概率。由此可以推出信息增益的公式如下:

$$IG(T)=H(C)-H(C|T)=H(C)-P(t)H(C|t)-P(\bar{t})H(C|\bar{t}) \quad (16)$$

通过设定阈值,可以判断应该选择哪些变量作为训练多分类SVM的特征。以上算法的伪代码在算法1中展示。

2 粒子群优化算法

粒子群优化算法(particle swarm optimization, PSO)是一种经典的启发式算法,使用无质量的例子来模拟鸟群里的鸟。粒子本身有两个属性:速度和位置。每个粒子在搜索空间中单独搜索最优解,记录其所获得的最优值,并将该值与粒子群里的其他粒子共享,经过比较后将最优粒子所获得的最优值作为整个粒子群本轮的全局最优值。而粒子群中所有非最优粒子根据自己找到的个体最优值和整个粒子群的全局最优值来调整其速度和位置。在求解非凸的子问题时,如果使用数学规划软件无法在规定的24h内返回最优解,则改为使用粒子群优化算法求解。

算法1 改进嵌入最小-最大值特征选择算法的搜索策略输入:$\tilde{S}$,S_{test}和超参数C_2

输出:最优特征子序列,S_{test}上的最大预测准确率

当$0\leqslant i\leqslant \text{a}-1$, $i+1\leqslant \text{j}\leqslant a-1$时

划分训练集 S_{train} 和验证集 S_{val}

对于所有方格中的(C,γ)

在S_{train}上求解问题(4)

在S_{val}上计算预测准确率

结束方格搜索

C^*, $\gamma^{\text{ini}}=S_{\text{val}}$上取到最大预测准确率的$(C,\gamma)$

在$\tilde{S}$上以C^*, γ^{ini}为初值求解问题(11),返回λ^{ini}, v^{ini}, $(\alpha^0)^{\text{ini}}$, $(\alpha^{\text{C}})^{\text{ini}}$

在$\tilde{S}$上以λ^{ini}, v^{ini}, $(\alpha^0)^{\text{ini}}$, $(\alpha^{\text{C}})^{\text{ini}}$为初值求解问题(13),返回$z^*$, λ^*, v^*, $(\alpha^0)^*$, $(\alpha^{\text{C}})^*$, 并计算S_{test}上的预测准确率

结束对于j的循环

结束对于i的循环

初始状态下,粒子群中所有粒子的均为随机,通过迭代找到最优解。每一轮迭代中,每个粒子根据其个体最优值(p_{best})和整个粒子群的全局最优值(g_{best})更新自身属性。在找到上述两个最优值后,每个粒子将通过以下两个公式来更新其位置和速度,即

$$x_i=x_i+v_i \quad (17)$$

$$v_i=v_i+c_{1\,\text{rand}}()(p_{\text{best}\,i}-x_i)+c_{2\,\text{rand}}()(g_{\text{best}\,i}-x_i) \quad (18)$$

式中:i是粒子群中粒子的个数,$i=1,2,...,N$;x_i是粒子的当前位置;v_i是粒子的速度;c_1、c_2是学习因子,通常均设为2;rand()是介于(0,1)之间的随机数。此外,式(19)也可用以更新粒子的速度

$$v_i=\omega v_i+c_{1\,\text{rand}}()(p_{\text{best}\,i}-x_i)+c_{2\,\text{rand}}()(g_{\text{best}\,i}-x_i) \quad (19)$$

式中:ω是超参数,代表惯性因子,其值非负。图(1)为PSO算法的流程。

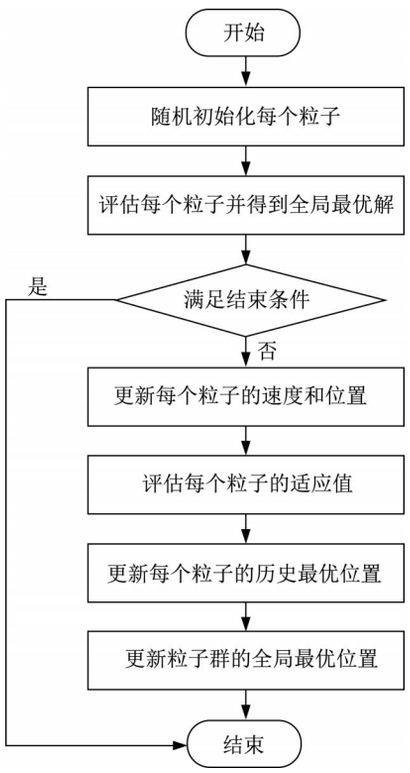

图1 PSO算法流程图

Fig.1 The flow chart of PSO algorithm

3 数值实验

3.1 钢板缺陷识别工程数据集介绍

钢板是一种重要的工业原材料,其质量很大程度上影响了产成品的质量。本文使用的钢板缺陷识别数据集来自于可以从UCI的网站http://archive.ics.uci.edu/ml下载。该数据集共包含27个特征,7个类别,1941组数据,其特征名称列于表1中。

这是一个多分类问题,共含有7个类别的标签。每个标签的名称及其数量如表2所示。

3.2 数据预处理

本数据集中,待分类的类别共7类,以独热编码(one-hot encoding)的形式存储。首先把类别转换成一一对应的形式,即将原数据集中的7列转换成1

表1　数据集中的特征名称

Tab.1　The name of features in dataset

编号	名称	编号	名称
1	X轴方向最小值	2	X轴方向最大值
3	Y轴方向最小值	4	Y轴方向最大值
5	像素面积	6	X轴方向外缘
7	Y轴方向外缘	8	亮度之和
9	亮度最小值	10	亮度最大值
11	传送带长度	12	A300钢铁型号
13	A400钢铁型号	14	钢板厚度
15	边缘指数	16	空白指数
17	面积指数	18	超出X轴范围指数
19	X轴边缘指数	20	Y轴边缘指数
21	超出全局范围指数	22	面积对数
23	X轴指数对数	24	Y轴指数对数
25	旋转指数	26	亮度指数
27	对面积进行sigmoid函数		

表2　标签的名称及其数量

Tab.2　The name of classes and its numbers

缺陷标号	缺陷名称	所占数目
1	油膜	158
2	Z型划痕	190
3	K型划痕	391
4	污渍	72
5	肮脏	55
6	碰撞	402
7	其他缺陷	673

列。其次,考察特征之间的相关性,排除存在共线性的特征。本文中,根据皮尔逊相关系数进行判断,排除相关系数绝对值大于0.95的特征。各个特征之间的相关系数如图2所示。经过判断,决定排除2、4、6、8、13号特征。

由于本文采用一对一方法,因此在训练每个

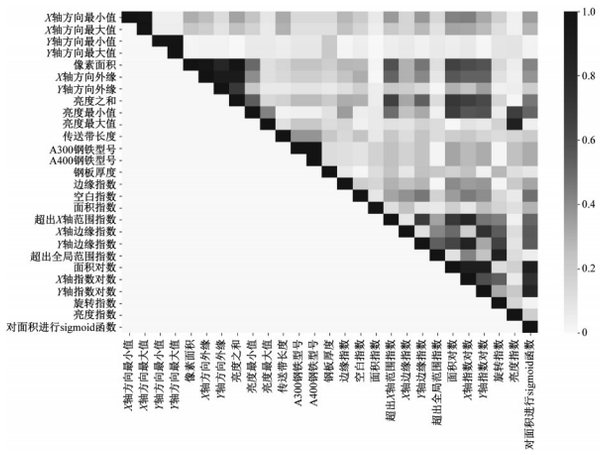

图2　各个特征之间的皮尔逊相关系数

Fig.2　The pearson correlation coefficients among different features

SVM时,都需要将$y_t=i$的标签设为+1,将$y_t=j$的标签设为−1。以上即为数据预处理部分。

3.3　实验设置

本文的所有程序均是在Windows 10环境下,使用Python 3.8编写。对于提出的优化问题,通过调用Gurobi 9.1.1或使用PSO算法进行求解;如果调用Gurobi求解的时间超过24h,即停止计算,选择启发式算法进行求解。

3.4　实验结果

按算法1,先解式(4),对C和γ进行方格搜索。假设$C=\{10^{-2},10^{-1},1,..,10,10^{2}\}$,$\gamma=\{10^{-2},10^{-1},1,...,10,10^{2}\}$。经过多次实验,得到当$C=5$,$\gamma=0.1$时,模型的平均得分最高。取其中一次的实验结果写入表3(保留两位小数)。

表3　C和γ进行方格搜索得到的模型预测准确率

Tab.3　The accuracy on test set using grid search with C and γ

C	γ												
	0.01	0.1	1	2	3	4	5	6	7	8	9	10	100
0.01	0.36	0.51	0.62	0.65	0.68	0.68	0.69	0.70	0.68	0.70	0.69	0.73	0.75
0.1	0.48	0.62	0.70	0.72	0.72	0.73	0.73	0.73	0.74	0.75	0.75	0.75	0.74
1	0.43	0.57	0.71	0.72	0.72	0.72	0.73	0.72	0.73	0.72	0.72	0.72	0.70
2	0.36	0.52	0.72	0.72	0.72	0.72	0.72	0.72	0.71	0.71	0.71	0.71	0.71
3	0.36	0.50	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
4	0.36	0.50	0.70	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.70
5	0.36	0.50	0.67	0.70	0.70	0.70	0.70	0.70	0.70	0.71	0.70	0.70	0.70
6	0.36	0.50	0.64	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
7	0.36	0.49	0.61	0.66	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
8	0.36	0.49	0.59	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.63
9	0.36	0.48	0.56	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
10	0.36	0.48	0.56	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
100	0.36	0.36	0.46	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48

如2.2节所述,超参数C_2的不同取值可以在模型复杂度和模型准确率之间进行权衡。选择C_2的所有可能取值为0.01,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9。由于设定超参数时尚未应用特征选择算法,意即固定$z=1$,取$C=5$,$\gamma=0.1$,求解问题(13)。根据文献[13]所述,如果固定γ,那么问题

(17)就是一个凸规划问题,可以调用Gurobi进行求解。在$\tilde{S}$上进行求解,在S_{test}上进行测试,得到的C_2与测试集上准确率如图3所示。

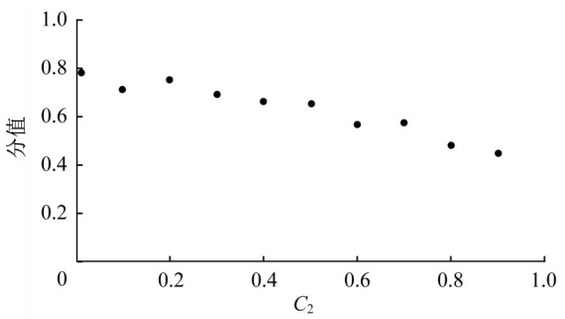

图3 C_2和模型准确率之间的关系

Fig.3 The relationship between C_2 and accuracy on test set

从图上可以看出,实验结果同先前假设基本一致,随着C_2不断增大,模型在测试集上的准确率不断降低,当$C_2=0.9$时,模型在测试集上的准确率降至44.60%。在该数据集中,选择超参数$C_2=0.2$,能够较好地平衡模型复杂度和模型预测准确率之间的关系。需要强调的是,超参数的选取同数据集本身相关。如文献[10]所述,在数值实验中,将特征选择算法应用到不同数据集时,所选取的C_2是不同的,因此作者建议超参数C_2应该由用户选取。

选取完C_2后,进入到算法1的流程中。由于0~1变量的引入,导致问题(11)和问题(13)的目标函数和约束条件均是非凸的。调用Gurobi解决问题(11)和问题(13)时,未能在规定的24h内求解出可行解,因此使用PSO算法对两个问题进行求解。

本数据集中,选择学习因子c_1、c_2等于2,惯性因子$\omega=0.3$,粒子个数为22,迭代次数为100次。先训练第i类和第j类数据之间的SVM,返回在该SVM情况下选择的特征。然后使用信息增益算法对返回的C_M^2个结果进行评估,从而选择合适的变量作为多分类SVM的最终特征。设定信息增益的阈值分别为0.2,0.5,0.8,1,大于该阈值的特征即被选中。被选中特征的数量和模型在测试集上的准确率如图(4)所示。

从图(4)可以看出,随着信息增益阈值不断增大,模型在测试集上的准确率不断变低。因为阈值变大,选取的特征数量会减少,从而导致有用信息的丢失。该数据集中,选取信息增益阈值等于0.5的变量作为训练的特征,这时选择的特征的编号是2,3,4,5,6,7,8,10,11,12,13,14,16,17,19,20。

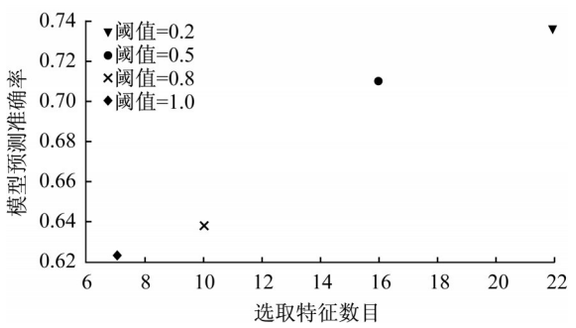

图4 信息增益阈值和模型预测准确率之间的关系

Fig.4 Relationship between threshold of information gain and accuracy on test set

最后将改进嵌入特征选择算法同未经特征选择的多分类SVM进行比较。作为基本假设,改进嵌入特征选择算法在预测准确度上不会优于基准算法且在可以接受的范围内,而训练的速度会快于基准算法,快的"程度"具体取决于用户选取的信息增益阈值。经过实验,基准算法训练SVM的时间是1.19s,在测试集上的准确率是73.8%;而使用改进嵌入特征选择算法训练SVM的时间是1.03s,在测试集上的准确率是71.2%。考虑到sklearn的库函数经过高度优化,且这个数据集的特征数目并不多,如若推广到更高维的数据集中,则提出的特征选择算法在多分类SVM问题中具有广泛的应用前景。

4 结论与不足

本文基于现有多分类SVM研究中的不足,提出了改进嵌入最小-最大值特征选择算法,在最小化特征数量的同时最大化模型预测的准确率。由于引入了0~1变量,该优化问题变成了组合优化问题,在限定时间内未能用数学规划软件求解得出全局最优解,因此选择使用启发式算法求解。

数值实验结果表明,本文提出的算法在该数据集上可以在牺牲可接受程度预测准确率下降的条件下换取模型训练时间的显著下降。

本文将SVM问题中的核函数限定为高斯核函数。然而,应用其他的核函数可能会得到不一样的结论。此外,提出的算法仅仅被应用在一个数据集中,可以考虑使用更多的经典数据集来验证其训练时间和训练精度。最后,未来的研究可以考虑放弃该模型中的0~1变量,转而使用其他评估方法来选取特征,因为这样可以充分利用到凸函数的性质,获得更令人满意的结果。

作者贡献声明:

武小军:提出选题,设计论文框架。

周文心:负责整理文献,算法构建,论文撰写与修订。

董永新:论文算法改进。

参考文献:

[1] KIRA K, RENDELL L A. A practical approach to feature selection[C]//Machine Learning Proceedings. San Francisco: Morgan Kaufmann, 1992: 249-256.

[2] AMOOZEGAR M, MINAEI-BIDGOLI B. Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism [J]. Expert Systems with Applications, 2018, 113: 499.

[3] CHEN Q, ZHANG M, XUE B. Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression [J]. IEEE Transactions on Evolutionary Computation, 2017, 21(5): 792.

[4] XIANG S, NIE F, MENG G, *et al*. Discriminative least squares regression for multiclass classification and feature selection [J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(11): 1738.

[5] CAI D, ZHANG C, HE X. Unsupervised feature selection for multi-cluster data [C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: [s.n.], 2010: 333-342.

[6] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods [J]. Computers & Electrical Engineering, 2014, 40(1): 16.

[7] LIU H, SETIONO R. A probabilistic approach to feature selection-a filter solution [C]//International Conference on Machine Learning. Bari: [s.n.], 1996: 319-327.

[8] LANGLEY P. Selection of relevant features in machine learning [C]//Proceedings of the AAAI Fall Symposium on Relevance. New Orleans: [s.n.], 1994, 184: 245-271.

[9] KOHAVI R, JOHN G H. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997, 97(1/2): 273.

[10] JIMÉNEZ-CORDERO A, MORALES J M, PINEDA S. A novel embedded min-max approach for feature selection in nonlinear support vector machine classification [J]. European Journal of Operational Research, 2021, 293(1): 24.

[11] MALDONADO S, LÓPEZ J. Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification[J]. Applied Soft Computing, 2018, 67: 94.

[12] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge University Press, 2000.

[13] KOTSIANTIS S B, ZAHARAKIS I D, PINTELAS P E. Machine learning: a review of classification and combining techniques [J]. Artificial Intelligence Review, 2006, 26(3): 159.

[14] 王乃芯. 多分类支持向量机的研究[D]. 上海:华东师范大学, 2020.
WANG Naixin. Research on multi-class classification support vector machines[D]. Shanghai: East China Normal University, 2020

[15] PLATT C J. Sequential minimal optimization: a fast algorithm for training support vector machines [R]. [S. l.]: Technical Report MSR-TR-98-14, 1998.

[16] HSU C W, LIN C J. A comparison of methods for multiclass support vector machines [J]. IEEE transactions on Neural Networks, 2002, 13(2): 415.

[17] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 1.

[18] ONEL M, KIESLICH C A, GUZMAN Y A, *et al*. Big data approach to batch process monitoring: Simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection [J]. Computers & Chemical Engineering, 2018, 115: 46.

[19] MANGASARIAN O L, MUSICANT D R. Lagrangian support vector machines [J]. Journal of Machine Learning Research, 2001(1): 161.

[20] AZHAGUSUNDARI B, THANAMANI A S. Feature selection based on information gain[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2013, 2(2): 18.

[21] LEE C, LEE G G. Information gain and divergence-based feature selection for machine learning-based text categorization [J]. Information Processing & Management, 2006, 42(1): 155.

[22] LEI S. A feature selection method based on information gain and genetic algorithm [C]//2012 International Conference on Computer Science and Electronics Engineering. [S.l.]: IEEE, 2012: 355-358.