鲁东大学学报(自然科学版)
Journal of Ludong University(Natural Science Edition) 2021,37(3):220—227

边界自适应三角模糊非线性优化支持向量分类器

王 琰[1a],李秀芳[1b],张志旺[2],周 莉[1a]

(1.鲁东大学 a.信息与电气工程学院;b.科学技术处,山东 烟台 264039;2.南京财经大学 信息工程学院,南京 210023)

摘要: 为了提高对存在噪声的大规模数据集的分类效果,本文提出了一种边界自适应三角模糊非线性优化支持向量分类器 BAT-FNOSVC。该分类器在支持向量分类器 SVC 的基础上引入边界自适应三角模糊隶属函数以更好地解决噪声带来的干扰问题,同时在模型中构造模糊列核矩阵及稀疏化函数,提高了算法的可解释性。对含噪数据集的实验结果表明,与采用三角形模糊隶属函数的稀疏非线性优化分类器 TFNOSVC、SVC、1-范数支持向量分类器 L1SVC 及最小二乘支持向量分类器 LSSVC 相比,BAT-FNOSVC 的准确率有明显提高,说明 BAT-FNOSVC 算法对有噪声的数据集具有较好的分类效果。

关键词: 特征选择;核方法;非线性规划支持向量分类器

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673-8020(2021)03-0220-08

支持向量机(support vector machine,SVM)是一种建立在统计学习理论基础上的有监督的机器学习方法,在一定程度上克服了维数灾难和过学习等传统困难,近年来在农产品损失预估[1]、硬件故障检测[2]及系统工程[3]等方面有着广泛的应用。SVM 遵从结构风险最小化原理,通过求解一个凸二次规划问题得到最优解,其泛化能力由分隔间隔所决定,该分隔间隔由 L2 范数导出的距离表示[4]。然而,随着数据的维度越来越高,标准 SVM 由于其模型设计机制的制约,存在算法计算复杂性高、抗噪能力较差等问题,在应用中受到很多限制。为进一步提高 SVM 对含噪数据集的分类准确率,2002 年 Lin 等[5]构造了模糊支持向量机(fuzzy support vector machine,FSVM)模型,将模糊理论与 SVM 算法相结合,提出了利用模糊隶属度对样本的重要性进行评价的思路。FSVM 算法可以一定程度上减少噪声等异常点对分类的影响,提高了 SVM 算法的分类准确率[6—7],其重点在于构造合适的隶属函数,将直接影响 FSVM 模型性能的好坏。本文提出一种边界自适应三角模糊非线性支持向量分类器(boundary adaptive triangular fuzzy nonlinear optimization support vector classifier,BAT-FNOSVC),该分类器给出一种边界自适应的三角模糊隶属函数,利用该函数计算输入样本点的模糊隶属度以去除噪声点,提高算法的分类精度;并且对权向量采用稀疏化函数,在分类的同时对数据集进行特征选择,提高模型的可解释性。

本文的结构安排如下:第一章介绍支持向量分类器及三角模糊隶属函数的相关理论知识,第二章介绍边界自适应三角模糊非线性优化支持向量分类器模型,第三章对 BAT-FNOSVC 模型进行实验验证与结果分析,第四章给出结论。

1 相关知识

1.1 支持向量分类器

支持向量分类器(support vector classifier,SVC)的目的是寻找具有最大间隔的分离超平面来区分正样本和负样本。对于线性可分的情况,在输入空间构造分离超平面进行分类;对于近似线性可分及非线性可分的情况,SVC 采用核函数方法将输入空间的样本映射到高维空间,在高维空间中构造具有最大软间隔的分离超平面以对样本进行分类。具体来说,假设有一组训练数据 T

收稿日期:2021-03-11;修回日期:2021-04-25
基金项目:国家自然科学基金重大研究计划项目(91538201);国家自然科学基金青年科学基金项目(61304052)

第一作者简介:王琰(1996—),女,山东德州人,硕士硕士生,研究方向为知识发现与智能决策支持系统。E-mail:1787274000@qq.com

通信作者简介:周莉(1966—),女,山东烟台人,教授,硕士研究生导师,博士,研究方向为信息融合。E-mail:zxm2zl@126.com

$=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$，其中 $\boldsymbol{x}_i\in\boldsymbol{X}$ 为样本输入，$y_i\in\{-1,+1\}$ 为类标签。SVC 的分离超平面定义为

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x})+b,\boldsymbol{w}\in\boldsymbol{F},b\in\mathbf{R},\quad(1)$$

其中 $\boldsymbol{w}$ 表示各个变量的权重，b 表示偏差。

这里 $\varphi(\cdot):\boldsymbol{X}\rightarrow\boldsymbol{F}$ 表示从输入空间 $\boldsymbol{X}$ 到高维特征空间 $\boldsymbol{F}$ 的非线性映射。为了构造这个最优超平面，SVC 的原始优化问题表示为

$$\begin{aligned}&\min_{\boldsymbol{w},b,\boldsymbol{\xi}}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i\\&\text{s.t.}\,y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b]\geqslant 1-\xi_i,\\&\xi_i\geqslant 0,i=1,2,\cdots,n,\end{aligned}\quad(2)$$

其中：$\boldsymbol{\xi}=(\xi_1,\xi_2,\cdots,\xi_n)^{\mathrm{T}}$，$\xi_i$ 是松弛变量；C 为输入参数，表示对错样本的惩罚常量。将模型(2)转化为相应的对偶问题并化简，得到如下模型：

$$\begin{aligned}&\min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i\\&\text{s.t.}\sum_{i=1}^{n}\alpha_i y_i=0,\\&0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,n,\end{aligned}\quad(3)$$

其中，α_i 表示拉格朗日乘法因子，$\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi^{\mathrm{T}}(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)$ 为核函数。由此，在低维空间中近似线性可分或非线性可分的问题被转换为高维空间中线性可分的问题，其相应的决策平面定义为

$$f(x)=\sum_{i=1}^{n}y_i\alpha_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)+b\ 。\quad(4)$$

1.2　基于类中心的三角形模糊隶属函数

考虑到 SVC 对数据集中的噪声和异常值非常敏感，为解决该问题，文献[8—10]在模型中引入适当的模糊隶属函数。模糊隶属函数的作用是对不同的数据点赋予不同的权重，即对一些具有代表性的点给予较高的模糊隶属函数值，而对噪声数据、异常点等给予较低的模糊隶属函数值[9]。对于正态分布的数据集来说，较为经典的构造方法为基于样本点与其所属的类中心距离构造模糊隶属函数[11]，包括基于类中心的均值模糊隶属函数、基于类中心的三角模糊隶属函数、基于类中心的梯形模糊隶属函数等。其中，相较于均值模糊隶属函数，三角形模糊隶属函数对数据拟合性更好，分类准确率更高；相较于梯形模糊隶属函数，三角模糊隶属函数的复杂度相对较低。因此，本文在三角模糊隶属函数的基础上进行改进。常见的三角模糊隶属函数以均值为中心，以三角形拟合数据分布[12]，近似高斯分布，如图 1 所示。其中：横坐标表示输入样本点，纵坐标表示样本点对应的模糊隶属值，红、绿、蓝 3 条曲线代表 3 种不同的高斯分布，黑色曲线表示三角形拟合曲线。

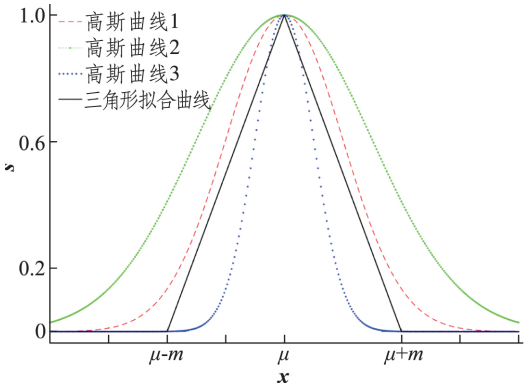

图 1　标准三角形模糊隶属函数曲线

Fig.1 Standard triangular fuzzy membership function curves

设 x_i 表示第 i 个输入点，μ 为输入点的均值，则三角形模糊隶属函数计算公式如下：

$$s_i=\begin{cases}0,|x_i-\mu|>m_i,\\1-\dfrac{x_i-\mu}{m_i},|x_i-\mu|<m_i,\end{cases}\quad(5)$$

其中 m_i 为第 i 个输入点的模糊子集边界，用来确认第 i 个输入点隶属度的必要条件，一般取 $m_i=2\sigma$，σ 表示第 i 个输入点的方差。

相对应的模糊隶属向量 $\boldsymbol{s}$（$\boldsymbol{s}\in\mathbf{R}^n$）为

$$\boldsymbol{s}=(s_1,s_2,\cdots,s_n)^{\mathrm{T}}。\quad(6)$$

2　边界自适应三角模糊非线性优化支持向量分类器 BAT-FNOSVC

本章首先针对原有模糊隶属函数构造的不足，提出一种边界自适应三角模糊隶属函数；其次将三角模糊隶属函数与列核矩阵相结合构造模糊列核矩阵；最后给出 BAT-FNOSVC 模型及相应算法。

2.1　边界自适应三角模糊隶属函数

如 1.2 节所述，常用的三角形模糊隶属函数通常是对数据点进行划分，根据隶属度的大小将数据点划分为噪声点和有用样本点，并且一般情况下三角模糊隶属函数将区分噪声的边界点设定为定值（一般选择在 2σ 处），然而实际应用中数

据集的概率分布不一定都在均值左右两倍方差的范围内。如图1中三角拟合曲线对高斯分布2拟合效果较好,但对高斯分布1和高斯分布3的拟合效果较差。由于不同的数据集有不同的概率密度分布,因此三角形模糊隶属函数应根据不同数据集的样本分布灵活设定其边界值。本文针对样本特征提出边界自适应三角形模糊隶属函数。该隶属函数根据统计学中的3σ原则构造,对数据集的样本特征计算模糊隶属度,在三角形边界点处引入方差倍数τ,使得对于不同概率分布的数据集,三角模糊隶属函数可以选择不同的边界点,以更好地适应数据分布,提高模型的分类准确率。边界自适应三角模糊隶属函数的构造思路如下所述。

令输入样本$\boldsymbol{x}_i$的第j个特征的特征值x_{ij}与特征均值μ_j之间的距离为d_{ij},即

$$d_{ij}=|x_{ij}-\mu_j|,\quad i=1,2,\cdots,n,\ j=1,2,\cdots,d。\tag{7}$$

根据距离公式(7),以均值μ_j为中心,距其左右$3\sigma_j$范围为界,当d_{ij}超出$3\sigma_j$时视为噪声点或异常值,设其模糊隶属度为0;当$\mu_j-3\sigma_j\leqslant d_{ij}\leqslant\mu_j+3\sigma_j$时,根据分段函数计算该特征相对应的模糊隶属度。根据上述思路,对每个输入特征x_{ij},其对应边界自适应三角模糊隶属函数计算公式为

$$s_{ij}=\begin{cases}1-\dfrac{d_{ij}}{\tau\sigma_j+\zeta},0\leqslant d_{ij}<\tau\sigma_j,\\0,\text{其他},\end{cases}\tag{8}$$

其中:s_{ij}表示输入样本x_i的第j个特征的模糊隶属度值;σ_j表示整体样本的第j个特征的方差;ζ表示极小的常数,用于防止分母为0;$\tau\in[0,3]$表示方差倍数,根据不同数据集的概率分布取不同数值。使用方差倍数可以使模糊隶属函数的分类边界点灵活变动,更好地拟合数据样本的分布。

假设输入样本$\boldsymbol{x}_i$的不同特征$f_j(j=1,2,\cdots,d)$之间相互独立同分布,因此对于每个输入点$\boldsymbol{x}_i$的模糊隶属度s_i为

$$s_i=\prod_{j=1}^{d}s_{ij},\ i=1,2,\cdots,n。\tag{9}$$

由此得到相对应的边界自适应三角模糊隶属向量$\boldsymbol{s}(\boldsymbol{s}\in\mathbf{R}^n)$为

$$\boldsymbol{s}=[s_1,s_2,\cdots,s_n]^{\mathrm{T}}。\tag{10}$$

2.2 模糊列核矩阵

在实际应用中,针对线性不可分或近似线性可分问题,传统方法为:对输入点$\boldsymbol{x}$使用适当的基本函数$\varphi(\boldsymbol{x})$,将样本点从原始输入空间映射到高维空间,使数据集在新的空间中线性可分离。因此,对于原始输入空间中的输入点$\boldsymbol{x}_i$和$\boldsymbol{x}_j$,通过采用单核函数$\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=[\varphi^{\mathrm{T}}(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)]$代替二者的点积$\varphi^{\mathrm{T}}(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)$的方法,实现数据点的线性可分离。然而上述单核函数的引入,一定程度上削弱了SVC模型的可解释性和稀疏性[13]。为了克服上述弊端,Nazarpour等[14]提出了通过组合具有不同特征或变量的多个核函数进行多核学习的方法。基于此,本文考虑对不同特征核及相应核权重组合线性函数,提出模糊列核矩阵,其构造过程如下。

给定具有模糊隶属度s的数据集$T'=\{(\boldsymbol{x}_1,y_1,s_1),(\boldsymbol{x}_2,y_2,s_2),\cdots,(\boldsymbol{x}_n,y_n,s_n)\}$,每个输入点$\boldsymbol{x}_i\in\mathbf{R}^n$属于两个类$y_i\in\{-1,+1\}(i=1,2,\cdots,n)$。对于线性可分的情况,输入点$\boldsymbol{x}_i$的第$j$个特征可表示为

$$\boldsymbol{f}_j=[x_{1j},x_{2j},\cdots,x_{nj}]^{\mathrm{T}},j=1,2,\cdots,d,\tag{11}$$

其中d为特征维数,n为样本数量。

由2.1节边界自适应三角模糊隶属函数可知,模糊列核矩阵中的特征核权重为s,定义第j个特征的模糊列核向量

$$\widetilde{\boldsymbol{U}}_j(\widetilde{\boldsymbol{U}}_j\in\mathbf{R}^n,j=1,2,\cdots,d)$$

为第j个特征的线性核$\boldsymbol{f}_j\boldsymbol{f}_j^{\mathrm{T}}$与其对应的特征核权重$s$的乘积,即

$$\widetilde{\boldsymbol{U}}_j=(\boldsymbol{f}_j\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{s}=[(x_{1j},\cdots,x_{nj})^{\mathrm{T}}(x_{1j},\cdots,x_{nj})]\boldsymbol{s}=\begin{bmatrix}x_{1j}x_{1j}&\cdots&x_{1j}x_{nj}\\\vdots&\ddots&\vdots\\x_{nj}x_{1j}&\cdots&x_{nj}x_{nj}\end{bmatrix}\begin{bmatrix}s_1\\\vdots\\s_n\end{bmatrix}=$$

$$\left[\sum_{l=1}^{n}s_lx_{lj}x_{1j},\cdots,\sum_{l=1}^{n}s_lx_{lj}x_{nj}\right]^{\mathrm{T}},j=1,2,\cdots,d。\tag{12}$$

对于非线性可分的情况,输入点$\boldsymbol{x}_i$的第j个特征可定义为

$$\boldsymbol{f}_j=[\varphi(x_{1j}),\varphi(x_{2j}),\cdots,\varphi(x_{nj})]^{\mathrm{T}},\quad j=1,2,\cdots,d。\tag{13}$$

通过使用核函数$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{y})=[\varphi(\boldsymbol{x})^{\mathrm{T}}\varphi(y)]$代替原始输入空间中输入点$\boldsymbol{x}$和$y_i$的点积$\varphi(\boldsymbol{x})^{\mathrm{T}}\varphi(y)$,则与第$j$个特征相对应的模糊列核向量$\widetilde{\boldsymbol{U}}_j(\widetilde{\boldsymbol{U}}_j\in\mathbf{R}^n,j=1,2,\cdots,d)$可描述为

$$\tilde{\boldsymbol{U}}_j = (\boldsymbol{f}_j \boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{s} = \{[\varphi(x_{1j}),\cdots,\varphi(x_{nj})]^{\mathrm{T}}[\varphi(x_{1j}),\cdots,\varphi(x_{nj})]\}\boldsymbol{s} = \begin{bmatrix} K(x_{1j},x_{1j}) & \cdots & K(x_{1j},x_{nj}) \\ \vdots & \ddots & \vdots \\ K(x_{nj},x_{1j}) & \cdots & K(x_{nj},x_{nj}) \end{bmatrix}\begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} = \left[\sum_{l=1}^{n} s_l K(x_{lj},x_{1j}),\cdots,\sum_{l=1}^{n} s_l K(x_{lj},x_{nj})\right]^{\mathrm{T}}$$

$$j = 1,2,\cdots,\mathrm{d}, \tag{14}$$

因此,对应上述两种情况的模糊列核矩阵 $\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{V}} \in \mathbf{R}^{n\times d})$ 为

$$\tilde{\boldsymbol{V}} = [\tilde{\boldsymbol{U}}_1,\tilde{\boldsymbol{U}}_2,\cdots,\tilde{\boldsymbol{U}}_d]。 \tag{15}$$

通过模糊列核矩阵,模型可以根据不同特征的核权重来获得每个特征对分类的贡献或重要性,对实际应用的分类结果给出合理的解释。此外,还可以对数据集进行特征选择或降维,以提高模型的分类效率。

常用的核函数包括线性核函数、RBF 核函数及多项式核函数[15],本文采用 RBF 核函数。对于任何两个输入点 $\boldsymbol{x}_i$ 和 $y_i(i = 1,2,\cdots,n)$,第 j 个特征的 RBF 核定义为

$$K(x_{ij},y_{ij}) = \exp\left[-\frac{(x_{ij}-y_{ij})^2}{2\sigma^2}\right](\sigma > 0)。 \tag{16}$$

2.3　边界自适应三角模糊非线性优化支持向量分类器

SVC 的泛化能力是由分隔间隔所决定的,该分隔间隔由 L2 范数导出的距离表示[4]。L2 范数正则化虽然可以解决过拟合问题,但默认样本的所有特征的权重为 1,无法进行特征选择; 同时 L2 范数正则化倾向于使用所有的特征,即每个特征的权重基本不会为 0,因此模型的稀疏性较差。Tibshirani 等[16—17]提出最小绝对收缩和选择算子(least absolute shrinkage and selection operator,LASSO)思想,即 L1 范数正则化,可以将不显著的变量系数压缩至 0,即将不重要的特征的权重压缩为 0,实现特征选择的目的。受 LASSO 思想的启发,本文采用权向量 $\boldsymbol{w}(\boldsymbol{w} \in \mathbf{R}^d)$ 的稀疏函数 $\|\boldsymbol{w}\|_1$(称为 L1 范数正则化),同时为了进一步加大模型的“惩罚”力度,增强模型的约简能力,本文对误差向量 ξ 进行 L2 范数正则化,由此得到新的分类模型

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} F\|\boldsymbol{w}\|_1 + \sum_{i=1}^{n}\xi_i^2$$

$$\text{s.t. } y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b] + \xi_i \geqslant 1,$$

$$i = 1,2,\cdots,n。 \tag{17}$$

由 1.1 节 SVC 模型的描述,模型(2)中的映射函数 $\varphi(\cdot)$ 可以选择合适的核函数,将样本从原始输入空间映射到高维空间。因此,在模型(17)中,将权向量 $\boldsymbol{w}(\boldsymbol{w} \in \mathbf{R}^d)$ 与 2.2 节构造的模糊列核矩阵 $\tilde{\boldsymbol{V}}$ 相结合,得到新模型

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} F\|\boldsymbol{w}\|_1 + \sum_{i=1}^{n}\xi_i^2$$

$$\text{s.t.} y_i(\tilde{\boldsymbol{V}}_i\boldsymbol{w} + b) \geqslant 1 - \xi_i,$$

$$i = 1,2,\cdots,n, \tag{18}$$

其中 $\tilde{\boldsymbol{V}}_i = \left[\sum_{l=1}^{n} s_l K(x_{l1},x_{i1}),\cdots,\sum_{l=1}^{n} s_l K(x_{ld},x_{id})\right]$,$\tilde{\boldsymbol{V}}_i$ 是 2.2 节定义的矩阵 $\tilde{\boldsymbol{V}}$ 中的第 i 行; F 为稀疏因子,用于平衡惩罚函数与损失函数。通过模糊列核矩阵,可以提高模型的特征选择能力。

由于模型(18)中的目标函数包含绝对值,即 $\|\boldsymbol{w}\|_1 = \sum_{j=1}^{d}|w_j|$,为计算方便,对模型(18)引入新的向量 $\boldsymbol{t}(\boldsymbol{t} \in \mathbf{R}^d)$,令 $|w_j| \leqslant t_j(j = 1,2,\cdots,d)$,得到 BAT-FNOSVC 模型

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{t}} F\sum_{j=1}^{d} t_j + \sum_{i=1}^{n}\xi_i^2$$

$$\text{s.t.} y_i(\tilde{\boldsymbol{V}}_i\boldsymbol{w} + b) \geqslant 1 - \xi_i,$$

$$-t_j \leqslant w_j \leqslant t_j, t_j \geqslant 0,$$

$$i = 1,2,\cdots,n, j = 1,2,\cdots,d。 \tag{19}$$

因此,对于来自独立测试集的新输入点 $\boldsymbol{x}$,其类标签 y 由决策函数计算得到。决策函数定义为

$$f(\boldsymbol{x}) = \mathrm{sign}(\tilde{\boldsymbol{V}}_x\boldsymbol{w} + b), \tag{20}$$

其中对应输入点的 $\tilde{\boldsymbol{V}}_x$ 由公式(15)计算。

3　系统工程数据集实验

本章使用 BAT-FNOSVC 模型和其他分类器对系统工程领域的实际数据集进行分类实验,并对实验结果进行对比分析。

3.1　数据集

本文选用来自 UCI 平台的 5 个实际数据集[18]。每个数据集都包括若干个条件属性和 1 个用于区分样本的类属性。其中,类属性中的+1 表示正类,记为正实例; 而-1 表示负类,记为负实例。表 1 展示了不同数据集的信息,包括数据集名称、正负实例数量、样本总数和特征数。

表1 数据集基本信息

Tab.1 Basic information of datasets

数据集名称	正实例数量	负实例数量	样本总数	特征数
Ionosphere	225	126	351	34
Waveform	1692	1653	3345	40
MasterA	35	52	87	36
Sonar	97	111	208	60
Sensorless	5319	5319	10 638	49

3.2 实验设计及评估指标

本文所有的实验都是在MATLAB 2014平台上进行,处理器为Intel(R) Core(TM) i5-4590 CPU @ 3.30 GHz,运行内存为8.00 GB。选用引入三角模糊隶属函数的非线性优化支持向量分类器(triangular fuzzy nonlinear optimization support vector classifier,TFNOSVC)、SVC、1-范数支持向量分类器(1-norm support vector classifier,L1SVC)及最小二乘支持向量分类器(least squares support vector classifier,LSSVC)作为BAT-FNOSVC的对比模型。对于训练集均采用分层的5折交叉验证方法,通过网格搜索方法从预先定义的参数集合中选择分类器模型的最佳参数。对于TFNOSVC及BAT-FNOSVC,定义权向量w的惩罚因子为F,F的取值范围为{0.01,0.02,0.05,0.1,0.2,0.25,0.5,1,2};令实验中使用的RBF核的带宽为σ,σ的取值范围为{0.01,0.02,0.05,0.062 5,0.1,0.2,0.25,0.5,1}。

本文在实验中使用5种精度指标来评估上述分类器的分类性能,包括:总体准确率、第一类准确率、第二类准确率、F-measure及约简率,这些性能分别定义为:

1) 总体准确率(total accuracy,记为TA)

$$TA = \frac{TP + TN}{TP + FN + FP + TN}; \tag{21}$$

2) 第一类准确率(type I accuracy,表示对正实例的识别准确率,记为T_1A)

$$T_1A = \frac{TP}{TP + FN}; \tag{22}$$

3) 第二类准确率(type II accuracy,表示对负实例的识别准确率,记为T_2A)

$$T_2A = \frac{TN}{FP + TN}; \tag{23}$$

4) F-measure(召回率和精确率的混合度量,记为F_1)

$$F_1 = \frac{2TP}{2TP + FN + FP}; \tag{24}$$

5) 约简率(reduction rate,表示衡量模型特征选择的能力,记为R_r)

$$R_r = \frac{|X_o| - |X_r|}{|X_o|} \times 100\%; \tag{25}$$

其中:TP表示被正确识别的正例的数量,TN表示被正确识别的负例的数量,FN表示正例被错误识别为负例的数量,FP表示负例被错误识别为正例的数量,X_o表示原始特征集,X_r表示被约简的特征组成的集合。F_1是召回率和精确率调和以后的平均度量指标,这是一个考虑敏感度和准确度影响的评价标准;约简率代表模型去除冗余特征的能力,约简率越高,模型去除冗余特征的能力越好。

3.3 模型分类性能的比较分析

本文将BAT-FNOSVC模型和其他分类器分别在5个实际应用的数据集进行实验,所得实验结果如图2~5所示。

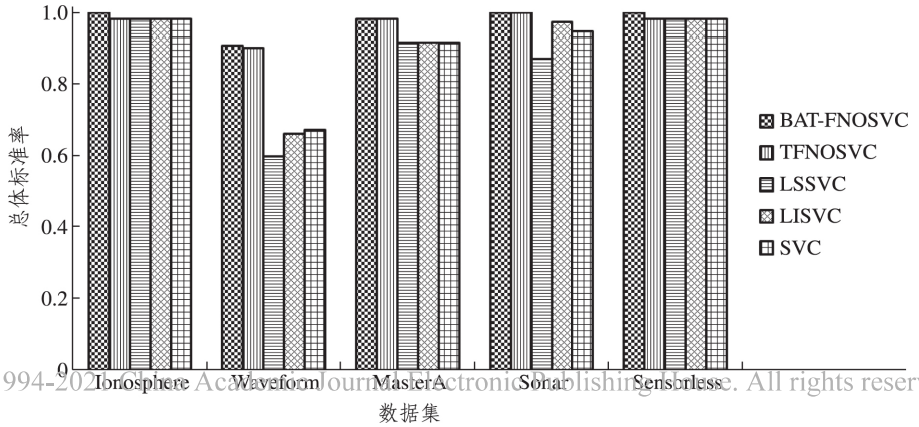

图2 不同分类器在5个数据集上的总体准确率比较

Fig.2 The total accuracy comparison of different classifiers on 5 datasets

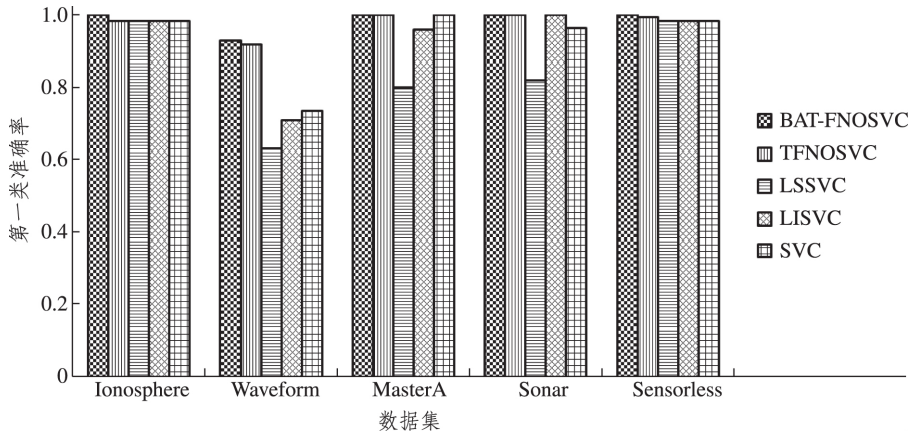

图 3　不同分类器在 5 个数据集上的第一类准确率比较

Fig.3 The type I accuracy comparison of different classifiers on 5 datasets

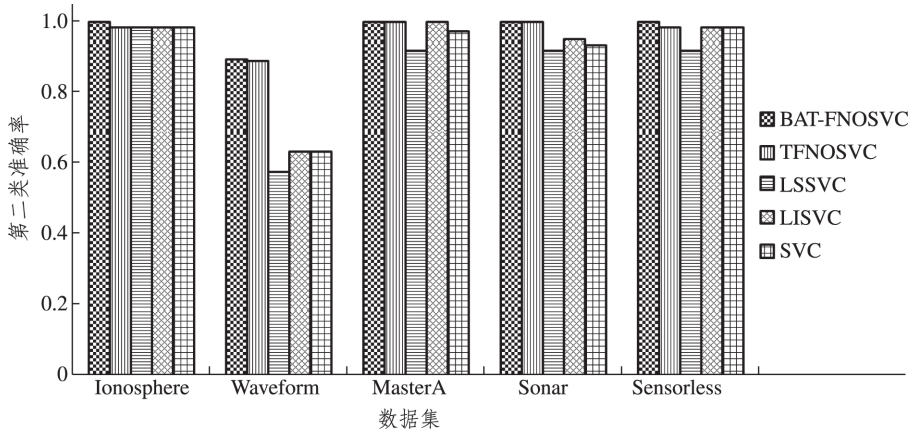

图 4　不同分类器在 5 个数据集上的第二类准确率比较

Fig.4 The type II accuracy comparison of different classifiers on 5 datasets

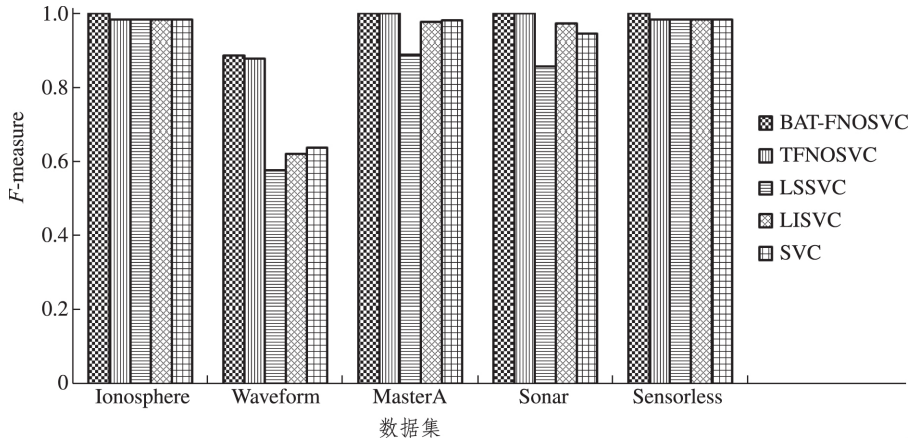

图 5　不同分类器在 5 个数据集上的 F-measure 比较

Fig.5 The F-measure comparison of different classifiers on 5 datasets

从图 2~5 中可以看出,在 5 个实际数据集的实验结果中,本文提出的 BAT-FNOSVC 的分类性能明显优于其他 4 种模型。其中,在图 2 的总体准确率中,BAT-FNOSVC 算法分类的总体准确率在 98%以上,较 TFNOSVC 提高了 0.12%,较 SVC 提高了 2.5%,较 LSSVC 提高了 4.1%,较 LISVC 提高了 2.7%; 尤其在 Waveform 数据集中,BAT-FNOSVC 准确率几乎达到 100%,表明边界自适应

三角模糊隶属函数确实能够更好地适应数据的分布,进一步去除噪声等干扰,提高模型的分类准确率。在图3的第一类准确率和图4的第二类准确率中,BAT-FNOSVC能达到88%以上,尤其在数据集Ionosphere中能达到100%,相较于TFNOSVC的准确率有明显提高,说明采用边界自适应三角模糊隶属函数的BAT-FNOSVC对不同数据集的正例及负例的识别准确率表现均良好,而其他4种分类器表现相对较差。图5的F-measure也表明了BAT-FNOSVC对数据集的综合处理能力较强,在实际数据集上有较好的分类效果。

3.4 特征选择与重要性分析

本文提出的BAT-FNOSVC模型在5个实际数据集实现了特征选择,结果见表2。

表2 BAT-FNOSVC模型在5个数据集上的特征选择结果

Tab.2 Feature selection results of BAT-FNOSVC model on 5 datasets

数据集	原始特征数	所选特征数	约简率/%	重要的特征(权重)
Ionosphere	34	2	94.12	F16(-0.100);F1(-0.050)
Waveform	40	3	92.50	F9(0.004);F23(-0.003);F17(0.003)
MasterA	36	2	94.44	F1(-0.250);F3(0.250);
Sonar	60	5	91.67	F3(0.050);F4(0.050);F6(-0.050);F11(0.050);F12(0.050)
Sensorless	49	3	93.88	F22(-0.010);F24(-0.010);F25(-0.010)

表2展示了BAT-FNOSVC模型在5个实际数据集上的特征选择及重要性分析的结果。如表2所示,采用边界自适应三角模糊隶属函数及多核学习的BAT-FNOSVC模型对数据集的约简率总体在90%以上,在数据集MasterA甚至达到94.44%,说明BAT-FNOSVC模型可以从每个数据集中识别并提取出较少的重要特征,同时自动去除其他不重要或不相关的特征,表明BAT-FNOSVC模型具有良好的特征约简能力。此外,经过5折交叉验证后,BAT-FNOSVC模型可以从最佳分类结果中选择最优特征并给出特征的重要性程度,如表2的第5列所示,权重越大表明该特征对分类的重要性程度越高(正数表示正相关,负数表示负相关)。对于实际应用中的每个数据集,可以根据特征的重要性对分类值进行原因分析,从而得到可追溯和可解释的结果。

4 结论

本文提出了一种边界值适应三角模糊非线性优化支持向量分类器BAT-FNOSVC,不仅扩展了SVC的功能,提高了对含噪数据集进行分类的整体性能,还增强了模型分类的可解释性。BAT-FNOSVC模型的特点是对输入点计算边界自适应三角模糊隶属函数,根据不同数据集的概率分布灵活设置分界点,以更好地适应数据集的变化,去除噪声和异常值;同时,将边界自适应三角模糊隶属函数结合多核学习,构造模糊列核矩阵,对实际应用中的分类问题给出合理的解释,提高分类效率。其次,通过采用权向量w的L1范数正则化得到各个特征对分类的贡献或重要性,使模型更稀疏,提高了模型的可解释性;再次,该模型采用误差向量ξ的L2范数正则化,加大惩罚项,使得模型的去冗余效果更好。最后,在5个实际数据集对BAT-FNOSVC模型进行了实验,实验结果和对比分析表明,BAT-FNOSVC模型在TFNOSVC模型的基础上进一步提高了分类精度,对数据集的适用性更广泛,具有较好的应用前景。

参考文献:

[1] 毛媛媛,张东,华小草,等.基于支持向量机的生鲜农产品风险损失预估[J].现代农业研究,2020,49(1):47-50.

[2] 江文建,姜斌,廖鹤,等.基于ILLE和SVM的卫星执行机构系统故障检测与定位[J].航天控制,2019,37(3):18-24.

[3] 王爽.基于时频分析和模糊函数的LPI雷达波形识别算法研究[D].哈尔滨:哈尔滨工程大学,2019.

[4] 吕洪林.基于L1范数支持向量分类机的算法扰动设计[J].平顶山学院学报,2019,34(5):37-39.

[5] LIN C F,WANG S D.Fuzzy support vector machines[J].IEEE Transactions on Neural Networks,2002,13(2):464-471.

[6] 邱云志,汪廷华,余武清.模糊支持向量机研究综述[J].赣南师范大学学报,2020,41(3):26-32.

[7] 王甜甜.加权模糊支持向量机及其应用研究[D].杭州:浙江工业大学,2017.

[8] BATUWITA R, PALADE V. Efficient resampling methods for training support vector machines with imbalanced datasets [C] // International Joint Conference on Neural Networks, 2010.
[9] AN W, LIANG M. Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises [J]. Neurocomputing, 2013, 110(13): 101-110.
[10] HEIKAMP K, BAJORATH J. Support vector machines for drug discovery [J]. Expert Opinion on Drug Discovery, 2014, 9(1): 93-104.
[11] 王宇凡.未确知信息分析的模糊支持向量机优化研究 [D].西安:西北工业大学, 2014.
[12] 罗周全,左红艳,王益伟.人—机—环境系统安全性的模糊熵评价方法 [J].模糊系统与数学, 2011(6): 169-174.
[13] ZHANG Z W, GAO G X, TIAN Y J. Multi-kernel multi-criteria optimization classifier with fuzzification and penalty factors for predicting biological activity [J]. Knowledge-Based Systems, 2015, 89: 301-313.
[14] NAZARPOUR A, ADIBI P. Two-stage multiple kernel learning for supervised dimensionality reduction [J]. Pattern Recognition, 2015, 48(5): 1854-1862.
[15] 吴晓萍,赵学靖,乔辉,等.基于 LASSO-SVM 的软件缺陷预测模型研究 [J]. 计算机应用研究, 2013, 30(9): 2748-2751.
[16] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society, Series B: Statistical Methodological, 1996, 58(1): 267-288.
[17] TIBSHIRANI R, SAUDERS M, ROSSET S, et al. Sparsity and smoothness via the fused lasso [J]. Journal of the Royal Statistical Society, Series B: Statistical Methodological, 2005, 67(1): 91-108.
[18] UCI Machine Learning Repository. Data Sets [DB/OL]. (2018-09-24) [2020-12-23]. http://archive.ics.uci.edu/ml/.

Boundary Adaptive Triangular Fuzzy Nonlinear Optimization Support Vector Classifier

WANG Yan[1a], LI Xiufang[1b], ZHANG Zhiwang[2], ZHOU Li[1a]

(1.a.School of Information and Electrical Engineering; b.Division of Science and Technology, Ludong University, Yantai 264039, China;
2.College Information Engineering, Nanjing University of Finance & Economics, Nanjing 210023, China)

Abstract: In order to improve the classification effect of large-scale noisy datasets, a boundary adaptive triangular fuzzy nonlinear optimization support vector classifier BAT-FNOSVC was proposed. Based on the support vector classifier SVC, the boundary adaptive triangular fuzzy membership function was introduced to better solve the interference problem caused by noise, and at the same time a fuzzy column kernel matrix and a sparse function were constructed in the model, which improves the interpretability of the algorithm. The experimental results on noisy data sets show that the accuracy of BAT-FNOSVC is significantly improved compared with triangular fuzzy nonlinear optimization support vector classifier TFNOSVC, SVC, 1-norm support vector classifier L1SVC and least squares support vector classifier LSSVC, indicating that the BAT-FNOSVC has the better classification effect on noisy datasets.

Keywords: fuzzy set; feature selection; kernel method; non-linear programming support vector classifier

(责任编辑　李秀芳)