文章编号：1000-8608(2023)04-0427-06

基于两步特征加权的模糊支持向量机算法

鞠 哲*，宋一明

（沈阳航空航天大学 理学院，辽宁 沈阳 110136）

摘要：提出一种基于两步特征加权的模糊支持向量机算法.首先，利用信息增益算法获取样本的特征权重.然后，计算最大权重的特征与其他特征间的斯皮尔曼相关系数，并将二者相乘后再与原有的特征权重相加，得到新的特征权重，减少弱相关和不相关特征对分类造成的影响.最后，在设计样本模糊隶属度时，不仅考虑样本与类中心的距离，还引入了样本间的亲和度，并将二者进行融合，以此减弱样本分布不均对分类精度的影响.在UCI数据集上的实验表明，与现有流行的几种模糊支持向量机算法相比，所提算法在准确率和 F_1 值上得到了提升.

关键词：模糊支持向量机；特征加权；信息增益；隶属度函数

中图分类号：TP18 **文献标识码：**A **doi：**10.7511/dllgxb202304013

0 引 言

支持向量机(support vector machine，SVM)是有着坚实理论基础的统计学习方法[1-3]，其学习策略是间隔最大化，旨在找到一个最优超平面将不同类样本尽可能地分隔开.SVM算法可以有效处理样本维度高、数量少、非线性等问题，目前已实际应用于各个领域.然而标准SVM算法对于噪声和野点敏感，导致最终生成的分类超平面次优，并且当数据集中正负类样本数量不均衡时，分类超平面也会向少数类偏移.为了克服上述问题，Lin等[4]提出了模糊支持向量机(fuzzy support vector machine，FSVM)，针对不同的样本点给定不同的模糊隶属度，使得不同样本对分类超平面的建立有着不同的贡献，一定程度上降低了噪声对SVM的影响.Lin等[4]认为样本越靠近类中心，属于该类的可能性越大，赋予较高的权重；反之若距离类中心越远，则赋予较低的权重，将其视为噪声，一定程度上降低了噪声对SVM的影响.文献[5]提出了对不同类样本赋予不同的惩罚因子，加大对少类样本的惩罚，可以有效解决数据不平衡导致的SVM算法失效问题.文献[6]将样本的不确定性和样本与类中心的距离相结合，提出了一种基于信息熵的改进FSVM算法，对非平衡数据集有着更高的分类精度.文献[7]加入了参数来调整分类超平面与样本的距离，有效改善了样本分布不均导致分类精度下降的问题.文献[8]在设计隶属度函数时，不仅考虑了样本与类中心的距离，还考虑了样本之间的紧密度.文献[9]对核函数做出了修正，提出了基于中心核对齐的模糊支持向量机.

上述算法均未考虑样本特征权重对分类超平面的影响，目前已有学者将特征加权方法引入模糊隶属度设计.文献[10]提出了特征加权支持向量机算法，避免了弱相关或不相关特征对分类超平面的干扰.邱云志等[11]在文献[9-10]的基础上，提出了双重特征加权的模糊支持向量机，考虑了特征加权对核函数的影响.左喻灏等[12]提出了Relief-F特征加权的FSVM算法，结合了样本权重和特征权重，提高了分类效率.然而，现有基于特征加权的模糊支持向量机算法在特征权重的获取上只计算了特征间的信息增益，未将重要和次重要的特征与不相关特征之间的特征权重差值放大，导致弱相关和不相关特征对分类还存在一定程度的干扰，最终使得分类效果不理想，并且在隶

收稿日期：2022-08-24； **修回日期：**2023-05-31.
基金项目：辽宁省自然科学基金资助项目(2019-BS-187)；辽宁省教育厅系列项目-青年科技人才“育苗”项目(JYT19027).
作者简介：鞠 哲*(1986—)，男，副教授，硕士生导师，E-mail：juzhe@sau.edu.cn.

属度函数设计上只考虑了样本与类中心的距离，无法缓解样本内部分布不均导致分类精度下降的问题. 为此，本文提出基于两步特征加权的模糊支持向量机算法. 首先，利用信息增益算法获取样本的特征权重. 然后，选择信息增益最大的特征，计算其与剩余特征的斯皮尔曼相关系数，将最大的特征权重与其他特征的相关系数相乘并加到其他特征原有的权重上，得到新的特征权重. 将得到的特征权重应用到隶属度函数距离的计算中，同时考虑样本的亲和度，通过样本内部的分布情况对隶属度函数做进一步修正.

1 模糊支持向量机简介

SVM 的思想是在样本空间或核空间中，使不同类样本的间隔尽可能大，并获取间隔最大时的分类超平面[1-3]. FSVM 是在 SVM 模型的基础上，给每个样本添加一个隶属度，用来表示不同样本对分类超平面的不同重要程度[4]. 对于一个训练集 $S=\{(\boldsymbol{x}_i,y_i,s_i)\}_{i=1}^{N}$，$\boldsymbol{x}_i\in\boldsymbol{R}^n$ 为训练样本；$y_i\in\{+1,-1\}$，为训练样本的标签，+1 为正类，−1 为负类；$s_i\in[0,1]$，为模糊隶属度，表示样本 $\boldsymbol{x}_i$ 属于类 y_i 的权重. FSVM 模型为

$$\begin{aligned}&\min_{\boldsymbol{\omega},b,\xi}\ \frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{N}s_i\xi_i\\&\text{s.t.}\ \ y_i(\boldsymbol{\omega}\cdot\varphi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,\ i=1,2,\cdots,N\\&\qquad\xi_i\geqslant 0,i=1,2,\cdots,N\end{aligned}\tag{1}$$

求解方式可见文献[11]，最终得到分类的决策函数为

$$f(\boldsymbol{x})=\operatorname{sgn}(\boldsymbol{\omega}\cdot\varphi(\boldsymbol{x})+b)=\operatorname{sgn}\Big(\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x}\cdot\boldsymbol{x}_i)+b\Big)\tag{2}$$

其中 $K(\boldsymbol{x}\cdot\boldsymbol{x}_i)=\varphi(\boldsymbol{x})\cdot\varphi(\boldsymbol{x}_i)$，为核函数，目的是将样本通过非线性映射 $\varphi(\boldsymbol{x})$ 映入高维空间.

2 基于两步特征加权的隶属度函数设计

本文提出的算法首先对特征进行两步加权，再使用特征加权距离计算样本的间距以及亲和度，得到每个样本的隶属度. 算法步骤如下：

步骤 1 进行特征加权 通过式(3)、(4)计算出所有特征的信息增益 $G(k)$. 通过式(5)计算出特征之间的斯皮尔曼相关系数 $c_r(i,k)$，i 和 k 为样本特征，如 $c_r(3,5)$ 表示第 3 个特征与第 5 个特征的相关系数. 找到信息增益最大的特征，位置记为 M，然后根据其与剩余特征的相关系数，以式(6)赋予最终的特征权重 $w(k)$. 由于斯皮尔曼相关系数的定义，绝对值大于 0.4 可以认为具有一定相关性，故找到绝对值大于 0.4 的特征，赋予其新的特征权重，小于 0.4 的特征不做处理. 将已得到 $w(k)$ 的特征忽略，对未赋予 $w(k)$ 的特征重复上述过程，直至全部特征都被赋予新的 $w(k)$. 此步骤的流程图如图 1 所示.

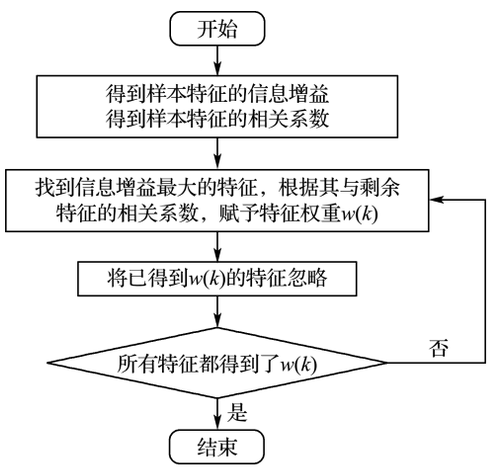

图 1 步骤 1 流程图

Fig. 1 Flowchart of Step 1

特征加权表示给样本的特征赋予相应的权重. 特征的信息增益表示特征对样本集合不确定性的减少程度，为信息熵与条件熵之差. 其中信息熵用来度量样本集合的不确定性，条件熵为特征给定条件下样本集合的不确定性. 具体计算方法如下：设 D 为数据集，$|D|$ 为数据集中的样本个数，D 中有 h 个类别标签 K_i（$i=1,2,\cdots,h$），$|K_{i,D}|$ 为 D 中标签为 K_i 的样本个数. D 的信息熵为

$$E_{\text{nt}}(D)=-\sum_{i=1}^{h}\frac{|K_{i,D}|}{|D|}\log_2\frac{|K_{i,D}|}{|D|}\tag{3}$$

若特征 B 有 v 个取值 B_j（$j=1,2,\cdots,v$）、D_j 为 D 中特征 B 上取值为 B_j 的数据集. 特征 B 对 D 进行划分得到的信息增益 $G(D,B)$ 为

$$G(D,B)=E_{\text{nt}}(D)-\sum_{j=1}^{v}\frac{|D_j|}{|D|}E_{\text{nt}}(D_j)\tag{4}$$

斯皮尔曼相关系数用来衡量两个变量之间的相关性大小，越趋于 0 表示两个变量之间的相关

性越低.斯皮尔曼相关系数在计算上采用取值等级而非取值本身,可以大大降低错误和极端数据对结果的影响.其计算公式为

$$\rho=\frac{\sum_{i=1}^{N}(R_i-\overline{R})(S_i-\overline{S})}{\left[\sum_{i=1}^{N}(R_i-\overline{R})^2\sum_{i=1}^{N}(S_i-\overline{S})^2\right]^2} \tag{5}$$

其中 R_i 和 S_i 表示样本 i 取值等级,$\overline{R}$ 和 $\overline{S}$ 表示 R 和 S 的平均等级,N 为样本个数.

$$w(k)=\begin{cases}G(k)+G(k); & k=M\\ G(k)+\max(G)\cdot|c_r(M,k)|; & k\neq M\end{cases} \tag{6}$$

步骤 2　设计模糊隶属度函数　本文在距离的计算上均使用特征加权距离,方法如式(7)所示,其中 l 表示特征的个数.通过模糊 C 均值算法得到样本的正负类中心 $\boldsymbol{x}_{\text{cen}}^{+}$,$\boldsymbol{x}_{\text{cen}}^{-}$ 以及 $d_i^{\text{cen}+}=d(\boldsymbol{x}_i,\boldsymbol{x}_{\text{cen}}^{+})$,$d_i^{\text{cen}-}=d(\boldsymbol{x}_i,\boldsymbol{x}_{\text{cen}}^{-})$,以式(8)赋予特征加权隶属度 $s_1(\boldsymbol{x}_i)$.然后考虑样本亲和度 $a(\boldsymbol{x}_i)$,根据特征加权距离进行紧密度和分散度计算,并将样本亲和度归一化,以此保证亲和度与 $s_1(\boldsymbol{x}_i)$ 在同等数量级上,得到隶属度 $s_2(\boldsymbol{x}_i)=-a(\boldsymbol{x}_i)$,计算得到最终的隶属度函数:$s(\boldsymbol{x}_i)=s_1(\boldsymbol{x}_i)+s_2(\boldsymbol{x}_i)$.再将 $s(\boldsymbol{x}_i)$归一化,防止隶属度为负.

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\sum_{k=1}^{l}w(k)(x_i^k-x_j^k)^2} \tag{7}$$

$$s_1(\boldsymbol{x}_i)=\begin{cases}s_1^{+}(\boldsymbol{x}_i)=1-\dfrac{d_i^{\text{cen}+}}{\max\limits_i(d_i^{\text{cen}+})+\delta};\\ \qquad i=1,2,\cdots,p\\ s_1^{-}(\boldsymbol{x}_i)=1-\dfrac{d_i^{\text{cen}-}}{\max\limits_i(d_i^{\text{cen}-})+\delta};\\ \qquad i=p+1,p+2,\cdots,N\end{cases} \tag{8}$$

此步骤使用了模糊 C 均值算法.假定对数据集 S 进行分类,每个点 $\boldsymbol{x}_i$ 属于第 j 个聚类中心 $\boldsymbol{c}_j$ 的隶属度为 μ_{ij},表达式为

$$J=\sum_{i=1}^{N}\sum_{j=1}^{H}\mu_{ij}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}_j\|^2 \tag{9}$$

约束条件为 $\sum_{j=1}^{H}\mu_{ij}=1,i=1,2,\cdots,N$.式中:$N$ 与 H 分别表示样本个数与聚类中心数,m 为隶属度因子,$\|\boldsymbol{x}_i-\boldsymbol{c}_j\|^2$ 表示 $\boldsymbol{x}_i$ 到聚类中心 $\boldsymbol{c}_j$ 的欧氏距离.要求 J 的值越小越好,通过反复求导计算,得到 μ_{ij} 与 $\boldsymbol{c}_j$ 的迭代公式为

$$\mu_{ij}=\frac{1}{\sum_{q=1}^{H}(\|\boldsymbol{x}_i-\boldsymbol{c}_j\|/\|\boldsymbol{x}_i-\boldsymbol{c}_q\|)^{2/(m-1)}} \tag{10}$$

$$\boldsymbol{c}_j=\sum_{i=1}^{N}(\mu_{ij}^{m}\cdot\boldsymbol{x}_i)\Big/\sum_{i=1}^{N}\mu_{ij}^{m} \tag{11}$$

本文样本亲和度为每个样本对数据集的影响程度,样本对数据集的影响由样本的分散度及紧密度体现[13].

样本分散度:删除每个样本前后,样本间距离标准差的变化比率.

$$U(\boldsymbol{x}_i,D)=\frac{t_{\text{std}}(D/\{\boldsymbol{x}_i\})}{t_{\text{std}}(D)};\ i=1,2,\cdots,N \tag{12}$$

样本紧密度:删除每个样本前后,样本均值的变化比率.

$$T(\boldsymbol{x}_i,D)=\frac{m_{\text{mean}}(D/\{\boldsymbol{x}_i\})}{m_{\text{mean}}(D)};\ i=1,2,\cdots,N \tag{13}$$

样本亲和度:删除每个样本前后,样本分散度与样本紧密度之比.

$$a(\boldsymbol{x}_i,D)=\frac{U(\boldsymbol{x}_i,D)}{T(\boldsymbol{x}_i,D)} \tag{14}$$

其中 t_{std} 与 m_{mean} 分别表示样本的距离标准差与均值.由上述可知,当样本分散度低、紧密度高时,样本对数据集的影响就越大,样本的亲和度就越小.

3　实验与结果分析

实验在 2.90 Hz/4.0 GB 的计算机上使用 Matlab 2021a 中的 libsvm 工具包实现.使用 UCI 数据集中的 8 个二分类数据集,数据集名称及相关信息见表 1.

表 1　UCI 数据集特征

Tab.1　UCI dataset characteristics

数据集名称	正类样本数	负类样本数	特征维数
Ionosphere	126	225	34
Hepatitis	32	123	19
Breast Cancer (BC)	201	85	9
Climate Model Simulation Crashes (CMSC)	46	494	18
Australian Credit Approval (ACA)	307	383	15
Stalog Heart (SH)	150	120	13
Vertebral Column (VC)	210	100	6
Breast Cancer Wisconsin (BCW)	444	239	9

核函数选择通用性较好的 RBF 核函数 $K(\boldsymbol{x}_i\cdot$

$\boldsymbol{x}_j)=\exp(-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)$. 为了更好地与现有算法对比，本文采用文献[11]中训练集和测试集的划分方式，将数据打乱顺序后以 7∶3 的比例分配训练集和测试集，使用网格搜索的方式选择最优参数，$C=\{2^{-5},2^{-4},\cdots,2^{15}\}$，$\gamma=\{2^{-15},2^{-14},\cdots,2^{3}\}$. 为防止数据集正负样本不平衡对分类精度的影响，本文在参数 C 的设定上使用文献[5]的方式，对不同类样本以不同的惩罚项 $C^+=C^-(N-p)/p$，其中 C^+、C^- 分别为少类样本与多类样本的惩罚项，p 表示少类样本个数，$N-p$ 为多类样本个数.

本文的评价指标为准确率(A_{cc})和 F_1 值. 准确率(A_{cc})表示分类正确的样本与总样本数量之比：

$$A_{cc}=\frac{T_p+T_n}{T_p+F_p+F_n+T_n} \tag{15}$$

F_1 值为精准率(P_{re})和召回率(R_{ec})的调和平均：

$$P_{re}=\frac{T_p}{T_p+F_p} \tag{16}$$

$$R_{ec}=\frac{T_p}{T_p+F_n} \tag{17}$$

$$F_1=\frac{2P_{re}\times R_{ec}}{P_{re}+R_{ec}} \tag{18}$$

式中：T_p 表示被正确分类的正类样本个数，T_n 表示被正确分类的负类样本个数，F_p 表示被误分类成正类样本的个数，F_n 表示被误分类成负类样本的个数.

本文算法与标准 SVM[1]、FSVM[4]、FWSVM[10]、改进算法 CKA-FSVM[9]，以及基于特征加权的 FWFSVM[12]、DFW-FSVM[11] 进行比较，结果见表2. 为了更好地与本文算法对

表 2 本文算法与其他算法在 UCI 数据集上的比较结果

Tab. 2 Comparison results of the proposed algorithm with other algorithms in UCI dataset

数据集	算法	准确率	F_1 值	数据集	算法	准确率	F_1 值
Ionosphere	SVM[1]	91.51	86.96	ACA	SVM[1]	76.81	75.00
	FSVM[4]	93.40	90.14		FSVM[4]	71.50	67.04
	FWSVM[10]	95.28	92.11		FWSVM[10]	82.13	82.30
	FWFSVM[12]	94.34	92.11		FWFSVM[12]	75.85	74.23
	CKA-FSVM[9]	92.45	88.89		CKA-FSVM[9]	68.12	69.44
	DFW-FSVM[11]	95.28	93.15		DFW-FSVM[11]	84.06	84.36
	本文算法	**96.30**	**94.69**		本文算法	**85.05**	**86.44**
Hepatitis	SVM[1]	80.85	40.00	SH	SVM[1]	79.01	81.32
	FSVM[4]	80.85	57.14		FSVM[4]	74.07	76.92
	FWSVM[10]	76.60	47.62		FWSVM[10]	80.25	82.98
	FWFSVM[12]	82.98	55.56		FWFSVM[12]	72.84	76.09
	CKA-FSVM[9]	85.11	36.36		CKA-FSVM[9]	76.54	78.65
	DFW-FSVM[11]	82.98	60.00		DFW-FSVM[11]	81.48	84.21
	本文算法	**86.67**	**62.50**		本文算法	**85.19**	**84.51**
BC	SVM[1]	68.60	78.74	VC	SVM[1]	82.80	88.06
	FSVM[4]	74.42	83.33		FSVM[4]	86.02	90.65
	FWSVM[10]	72.09	81.54		FWSVM[10]	87.10	90.91
	FWFSVM[12]	75.58	84.21		FWFSVM[12]	87.10	91.30
	CKA-FSVM[9]	69.77	82.19		CKA-FSVM[9]	86.02	90.08
	DFW-FSVM[11]	76.74	85.07		DFW-FSVM[11]	89.25	92.42
	本文算法	**77.27**	**85.51**		本文算法	**90.32**	**92.68**
CMSC	SVM[1]	92.59	45.45	BCW	SVM[1]	95.54	96.60
	FSVM[4]	93.21	42.11		FSVM[4]	95.05	96.24
	FWSVM[10]	93.83	44.44		FWSVM[10]	96.04	96.97
	FWFSVM[12]	94.44	52.63		FWFSVM[12]	95.54	96.60
	CKA-FSVM[9]	94.44	66.67		CKA-FSVM[9]	95.05	96.24
	DFW-FSVM[11]	**95.68**	**66.67**		DFW-FSVM[11]	96.53	97.36
	本文算法	95.28	66.67		本文算法	**97.20**	**97.79**

比，将 FWFSVM[12] 中 Relief-F 算法进行特征加权的方式替换为信息增益算法.

本文算法与 FWSVM 和 FWFSVM 相比，准确率与 F_1 值全部得到了提升，说明进行两步特征加权的方式可以最大限度地放大重要和次重要特征与弱相关和不相关特征在权重上的差值，有效避免了后者对分类的干扰，加强了相对重要特征对分类的贡献，训练出了分类性能良好的模型. 在 ACA 和 SH 数据集上，FSVM 算法的准确率和 F_1 值要低于 SVM 算法，原因是只考虑样本与类中心距离的隶属度函数会因数据集的不规则分布导致分类精度下降. 本文算法在考虑了样本亲和度后，衡量了每个样本的存在对数据集的影响，利用样本内部的分布情况对隶属度函数做出了适当修正，减小了仅使用样本与类中心距离作为隶属度函数时对数据集几何形状的依赖，在数据集非球形分布时的分类精度也获得了提升，降低了噪声和野点对分类超平面的干扰，并且本文使用了聚类的方式获得类中心，相比于求平均值计算出的类中心，虽然前者在计算上有一定的耗时，但其含有数据集中更多的样本信息，有助于获取更准确的样本隶属度.

从表 2 的结果上看，本文算法在 7 个 UCI 数据集(除 CMSC 数据集)上的准确率和 F_1 值有 0.5%～4.0%的提升. 其中在 SH 数据集上的准确率提升最大，在 Hepatitis 数据集上的 F_1 值提升最大，说明本文所提出的基于两步特征加权思想有效地提高了算法的泛化性. 但是针对某些数据集该算法也会存在特征加权失效的情况. 例如，在 CMSC 数据集上的准确率并未得到提升，低于 DFW-FSVM 算法，原因是在此数据集上，特征的信息增益接近，并且相关系数都趋于 0，导致提出的两步特征加权方法失效，特征加权步骤近似退化为 DFW-FSVM 的计算方式，并且此数据集正负类样本比例高度不平衡，这也对算法的分类精度产生一定影响，但在设计隶属度函数时考虑了样本的亲和度，使得准确率相比于 FWSVM 与 FWFSVM 算法还存在一定的提升. 另外，本文提出的特征权重计算及隶属度函数设计虽然相比于 FSVM 有着额外的耗时，但是算法的复杂度并未增加，有一定的推广价值.

4　结　语

本文设计的两步特征加权方法充分放大了重要特征与弱相关或不相关特征在权重上的差值，有效防止了后者对分类的影响，并且根据样本内部的分布情况对隶属度函数进行进一步修正，使得每个样本都具有相对合理的隶属度，降低了噪声和野点对分类超平面的干扰. 但在不平衡数据集下，本文并未在算法层面提出新的计算方法，下一步的研究目标为设计出新的针对不平衡数据集的 FSVM 算法.

参考文献：

[1] VAPNIK V N. **The Nature of Statistical Learning Theory** [M]. New York: Springer, 1995.

[2] CRISTIANINI N, SHAWE-TAYLOR J. **An Introduction to Support Machines and Other Kernel-based Learning Methods** [M]. Cambridge: Cambridge University Press, 2000.

[3] 李　航. 统计学习方法 [M]. 北京：清华大学出版社，2012.
LI Hang. **Statistical Learning Methods** [M]. Beijing: Tsinghua University Press, 2012. (in Chinese)

[4] LIN Chunfu, WANG Shengde. Fuzzy support vector machines [J]. **IEEE Transactions on Neural Networks**, 2002, **13**(2): 464-471.

[5] VEROPOULOS K, CAMPBELL I C G, CRISTIANINI N. Controlling the sensitivity of support vector machines [C]// **Proceedings of the International Joint Conference on Artificial Intelligence**. Stockholm: IJCAI Press, 1999: 55-60.

[6] 魏　鑫，张雪英，李凤莲，等. 面向非平衡数据集分类的改进模糊支持向量机 [J]. 计算机工程与设计，2019，**40**(11)：3124-3129，3199.
WEI Xin, ZHANG Xueying, LI Fenglian, *et al*. Improved fuzzy support vector machine for classification of imbalanced datasets [J]. **Computer Engineering and Design**, 2019, **40**(11): 3124-3129, 3199. (in Chinese)

[7] 李村合，姜　宇，李　帅. 基于不等距超平面距离的模糊支持向量机 [J]. 计算机系统应用，2020，**29**(10)：185-191.
LI Cunhe, JIANG Yu, LI Shuai. Fuzzy support vector machine algorithm based on inequality hyper-

plane distance [J]. **Computer Systems and Applications**, 2020, **29**(10): 185-191. (in Chinese)

[8] 鞠 哲, 曹隽喆, 顾 宏. 用于不平衡数据分类的模糊支持向量机算法 [J]. 大连理工大学学报, 2016, **56**(5): 525-531.
JU Zhe, CAO Junzhe, GU Hong. A fuzzy support vector machine algorithm for imbalanced data classification [J]. **Journal of Dalian University of Technology**, 2016, **56**(5): 525-531. (in Chinese)

[9] WANG Tinghua, QIU Yunzhi, HUA Jialin. Centered kernel alignment inspired fuzzy support vector machine [J]. **Fuzzy Sets and Systems**, 2020, **394**: 110-123.

[10] 汪廷华, 田盛丰, 黄厚宽. 特征加权支持向量机 [J]. 电子与信息学报, 2009, **31**(3): 514-518.
WANG Tinghua, TIAN Shengfeng, HUANG Houkuan. Feature weighted support vector machine [J]. **Journal of Electronics and Information Technology**, 2009, **31**(3): 514-518. (in Chinese)

[11] 邱云志, 汪廷华, 戴小路. 双重特征加权模糊支持向量机 [J]. 计算机应用, 2022, **42**(3): 683-687.
QIU Yunzhi, WANG Tinghua, DAI Xiaolu. Doubly feature-weighted fuzzy support vector machine [J]. **Journal of Computer Applications**, 2022, **42**(3): 683-687. (in Chinese)

[12] 左喻灏, 贾连印, 游进国, 等. 基于 Relief-F 特征加权的模糊支持向量机的分类算法 [J]. 化工自动化及仪表, 2019, **46**(10): 834-838, 864.
ZUO Yuhao, JIA Lianyin, YOU Jinguo, *et al*. Classification algorithm based on Relief-F feature weighting fuzzy support vector machine [J]. **Control and Instruments in Chemical Industry**, 2019, **46**(10): 834-838, 864. (in Chinese)

[13] 李 娟, 王宇平. 结合紧密度和分散度的近邻亲和相似度函数 [J]. 西安电子科技大学学报, 2014, **41**(3): 123-130.
LI Juan, WANG Yuping. New nearest neighbor affinity similarity function based on separation and compactness between samples [J]. **Journal of Xidian University**, 2014, **41**(3): 123-130. (in Chinese)

Fuzzy support vector machine algorithm based on two-step feature weighting

JU Zhe*, SONG Yiming

(College of Science, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: A fuzzy support vector machine algorithm based on two-step feature weighting is proposed. Firstly, the information gain algorithm is used to obtain the feature weights of the samples. Then, the Spearman correlation coefficients between the feature with the maximum weight and other features are calculated, and the corresponding Spearman correlation coefficients are multiplied by the maximum feature weight. Then the results are added with the original feature weights to get the new feature weights, so as to reduce the impact of weakly correlated features and irrelevant features on classification. Finally, when designing the fuzzy membership of samples, not only the distance between samples and class center is considered, but also the affinity between samples is introduced. And the distance and the affinity are fused so as to reduce the influence of uneven distribution of samples on classification accuracy. Experiments on UCI dataset show that compared with several popular fuzzy support vector machine algorithms, the proposed algorithm is improved in accuracy and F_1 value.

Key words: fuzzy support vector machine; feature weighting; information gain; membership function