第63卷第2期 大连理工大学学报 Vol. 63, No. 2
2023年3月 Journal of Dalian University of Technology Mar. 2023

文章编号：1000-8608(2023)02-0199-06

# 一种基于动态类中心模型选择的模糊支持向量机

宋一明，鞠　哲*

（沈阳航空航天大学 理学院，辽宁 沈阳　110136）

**摘要：**模糊支持向量机的核心思想是赋予样本模糊隶属度，给每个样本以不同的权重，从而克服标准支持向量机对噪声和异常点敏感的问题.现有的模糊支持向量机算法通常以样本与类中心距离为基础，给每个样本赋予一个固定的隶属度，没有根据样本分布对隶属度做进一步修正.提出了一种新的动态方式赋予样本隶属度，利用萤火虫算法不断地更新样本中心的位置和隶属度函数，同时利用粒子群算法优化模糊支持向量机参数.在UCI数据集上的实验结果表明，该算法可以有效减少噪声和野点对超平面的影响，分类性能要优于几类常用的模糊支持向量机算法.

**关键词：**模糊支持向量机；隶属度函数；分类；粒子群算法；萤火虫算法

**中图分类号：**TP18　　**文献标识码：**A　　**doi：**10.7511/dllgxb202302011

## 0　引　言

支持向量机(support vector machine，SVM)是一种统计学习算法，常用于处理二分类问题，其学习策略是间隔最大化[1-3]，针对小样本、高维度和非线性等问题有着很好的分类性能和泛化能力.但是当数据集中存在噪声和孤立点时，获得的超平面往往不理想，影响分类效果.为了解决这一问题，Lin等[4]首次提出了模糊支持向量机(fuzzy support vector machine，FSVM)，该方法通过给每个样本赋予不同的隶属度，使每个样本点对超平面的贡献程度不同，降低噪声和离群点的干扰，从而获得更高的分类精度.文献[5]提出了基于不等距超平面距离的模糊支持向量机，通过加入参数来控制超平面与样本间的距离，改善了由于样本分布不均和噪声导致分类精度下降的问题.文献[6]提出了一种新的划分指数最大化聚类算法，为FSVM获得更合理、更稳健的模糊隶属度.该算法先使用聚类算法获得合适的数据中心，再修改聚类的边界以获得新的隶属度，成功降低了算法的计算复杂度.文献[7]提出了基于异类类内超平面的模糊支持向量机，其方法是根据样本到异类类内超平面的距离构造隶属度函数，更加重视距离异类点较近范围的样本，降低了隶属度函数对样本几何形状的依赖.左喻灏等[8]提出了Relief-F特征加权的FSVM算法，该算法计算了各特征权重并删除权重较小的特征后，基于加权欧氏距离构造隶属度函数，提高了分类的效率.文献[9]给正负样本以不同罚因子(different error costs，DEC)，在算法层面减少不平衡数据给SVM算法带来的影响.文献[10]结合了文献[9]的思想，在隶属度函数的设计上不仅考虑了样本与其类中心的距离，还考虑了样本的紧密度.邱云志等[11]提出了一种双重特征加权的模糊支持向量机，通过信息增益的方式计算样本的特征权重，再基于特征权重计算样本与类中心的欧氏距离，构造出加权隶属度函数和核函数，有效地防止了分类结果被弱相关的特征干扰.文献[12]根据凸壳的定义在核空间中得到数据的几何特征，将正常类数据包含在最小超球内，并使噪声与超球的间隔最大化，此算法对数据内部噪声不敏感，降低了训练时间.

现有的大部分算法都是直接赋予样本一个固定的隶属度，没有考虑在训练后根据测试结果修

**收稿日期：**2022-04-11；　**修回日期：**2023-01-31.
**基金项目：**辽宁省自然科学基金资助项目(2019-BS-187)；辽宁省教育厅项目(JYT19027).
**作者简介：**宋一明(1995-)，男，硕士生，E-mail：793259031@qq.com；鞠　哲*(1986-)，男，副教授，硕士生导师，E-mail：juzhe@sau.edu.cn.

改隶属度函数并再次进行训练的动态模型以及模糊支持向量机参数确定的问题. 本文受文献[4, 13]的启发,以几何球形为基础,利用萤火虫算法不断地更新样本中心的位置和隶属度函数,同时利用粒子群算法优化模糊支持向量机参数,以获得更好的分类效果.

## 1 模糊支持向量机简介

模糊支持向量机进行训练时,需要在原有公式上添加每个样本的隶属度,得到一个带权重的分类误差项. 实际操作方式:对于一个训练集 $S=\{(\boldsymbol{x}_i,y_i,s_i)\}_{i=1}^{N}$,其中 $\boldsymbol{x}_i\in\boldsymbol{R}^n$ 为训练样本;$y_i\in\{+1,-1\}$为训练样本的标签,+1 代表正类,−1 代表负类;$s_i\in[0,1]$为模糊隶属度,是该训练样本 $\boldsymbol{x}_i$ 属于类 $y_i$ 的权重. FSVM 模型为

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

求解此问题,构造拉格朗日函数:

$$\begin{aligned}L(\boldsymbol{\omega},b,\alpha,\xi,\mu)=&\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{N}s_i\xi_i+\sum_{i=1}^{N}\alpha_i[1-\\&\xi_i-y_i(\boldsymbol{\omega}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)+b)]-\sum_{i=1}^{N}\mu_i\xi_i\end{aligned}\tag{2}$$

其中 $\alpha_i\geqslant 0,\mu_i\geqslant 0$ 为拉格朗日乘子,对 $\boldsymbol{\omega}$、$b$、$\xi$ 的偏导为 0,有

$$\begin{aligned}&\boldsymbol{\omega}-\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{\varphi}(\boldsymbol{x}_i)=\boldsymbol{0}\\&\sum_{i=1}^{N}\alpha_i y_i=0\\&s_iC-\alpha_i-\mu_i=0\end{aligned}\tag{3}$$

将式(3)代入式(2),得到对偶问题:

$$\begin{aligned}&\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{i=1}^{N}\alpha_i\\&\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0\\&\qquad \alpha_i\geqslant 0;\ i=1,2,\cdots,N\end{aligned}\tag{4}$$

其中 $K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)=\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_i)\cdot\boldsymbol{\varphi}(\boldsymbol{x}_j)$为核函数,将数据映射到新的空间,使其变为线性可分问题. 分类的决策函数为

$$\begin{aligned}f(\boldsymbol{x})=&\operatorname{sgn}(\boldsymbol{\omega}^*\cdot\boldsymbol{\varphi}(\boldsymbol{x})+b^*)=\\&\operatorname{sgn}\Big(\sum_{i=1}^{N}\alpha_i^* y_i K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+b^*\Big)\end{aligned}\tag{5}$$

其中 sgn(·)为符号函数.

本文选用具有良好性能的径向基核函数 $K(\boldsymbol{x},\boldsymbol{z})=\exp(-\gamma\cdot\|\boldsymbol{x}-\boldsymbol{z}\|^2)$进行实验.

## 2 隶属度设计

合适的隶属度函数能使算法有着更好的分类性能. Lin 等[4]是以类中心为基础赋予隶属度,其方法为先找到类的几何中心 $\boldsymbol{x}_+$、$\boldsymbol{x}_-$,再求出距离类中心最远样本的距离作为半径 $r_+=\max\|\boldsymbol{x}_i-\boldsymbol{x}_+\|$,$r_-=\max\|\boldsymbol{x}_i-\boldsymbol{x}_-\|$,以式(6)赋予隶属度,$\delta$ 为提前给出的很小的数. 但此时给定的隶属度是固定的,并未考虑在同一样本下,赋予其不同的隶属度是否会得到更高的分类精度.

$$s_i=\begin{cases}1-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_+\|}{r_++\delta}; & y_i=+1\\[2ex] 1-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_-\|}{r_-+\delta}; & y_i=-1\end{cases}\tag{6}$$

文献[13]将经验风险作为粒子群算法的适应度函数,并未考虑分类器的综合性能,即在得到较高正类样本分类精度的同时,也得到较高负类样本的分类精度.

本文以上述方法为基础,利用萤火虫算法和粒子群算法,提出一种新的模糊支持向量机算法. 算法通过实时更改类几何中心位置和隶属度函数中的变量 $R_0$,并同时搜索最优的参数 $C$ 和 $\gamma$,来对样本赋予隶属度,训练后得到最优超平面. 算法的流程如图 1 所示.

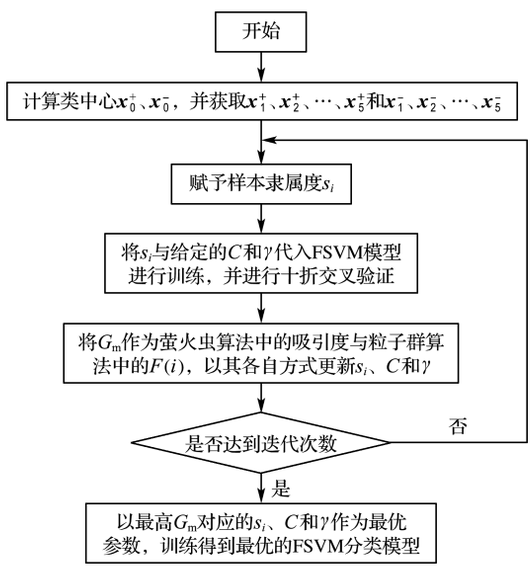


图 1 算法流程图

Fig. 1 Flowchart of the algorithm

算法的详细步骤如下：

(1)确定正、负样本类中心. 通过模糊 $C$ 均值算法，获得初始的正、负类样本的类中心 $\boldsymbol{x}_0^+$、$\boldsymbol{x}_0^-$，然后找到距离中心最近的点 $\boldsymbol{x}_1^+$、$\boldsymbol{x}_2^+$、…、$\boldsymbol{x}_5^+$ 和 $\boldsymbol{x}_1^-$、$\boldsymbol{x}_2^-$、…、$\boldsymbol{x}_5^-$.

此步骤的思路：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 $\mu_{ij}$，表达式为

$$J=\sum_{i=1}^{N}\sum_{j=1}^{H}\mu_{ij}^m\|\boldsymbol{x}_i-\boldsymbol{c}_j\|^2 \tag{7}$$

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

$$\mu_{ij}=1\Big/\sum_{k=1}^{H}(\|\boldsymbol{x}_i-\boldsymbol{c}_j\|/\|\boldsymbol{x}_i-\boldsymbol{c}_k\|)^{2/(m-1)} \tag{8}$$

$$\boldsymbol{c}_j=\sum_{i=1}^{N}\mu_{ij}^m\cdot\boldsymbol{x}_i\Big/\sum_{i=1}^{N}\mu_{ij}^m \tag{9}$$

(2)设计模糊隶属度函数. 寻找距离正、负类中心最远的点 $\boldsymbol{x}_{0\max}^+$、$\boldsymbol{x}_{0\max}^-$ 及其距离 $R_{\max}^+$、$R_{\max}^-$，并以此距离的 8/10 作为初始半径 $R_0$（$R_0=8R_{\max}/10$），距离大于 $R_0$ 的点，将其隶属度设置为 0，小于 $R_0$ 的点，以式(10)赋予隶属度. $R_i$ 为第 $i$ 个点与其中心点的距离.

$$s_i=\begin{cases}1-R_i/R_0^+; & y_i=+1\\ 1-R_i/R_0^-; & y_i=-1\end{cases} \tag{10}$$

(3)利用粒子群算法，搜索最优参数 $C$ 和 $\gamma$. 给定 FSVM 中参数 $C$ 与 $\gamma$ 的候选区间，随机给定 20 个点，作为粒子群算法的初始点，每个点 $\boldsymbol{X}_i$ 是二维向量（$\boldsymbol{X}_i=(C\quad\gamma)$），相当于给定了 20 个不同的 $C$ 与 $\gamma$，在此 $\boldsymbol{X}_i$ 上进行迭代计算，将第(2)步设定的隶属度 $s_i$ 与 $\boldsymbol{X}_i$ 代入进行训练并做十折交叉验证，找到最好 $G_{\mathrm{m}}$ 下的 $\boldsymbol{X}_i$ 为最佳点.

此步骤的计算使用了粒子群算法，思路为个体在空间中搜索最优解，并与全体空间得到的结果进行交换. 初始点随机获得，迭代公式为

$$v_i=v_i+c_1d_{\mathrm{rand}}(p_{\mathrm{best}}-z_i)+c_2d_{\mathrm{rand}}(g_{\mathrm{best}}-z_i) \tag{11}$$

$$z_i=z_i+v_i \tag{12}$$

其中 $v_i$ 为个体速度，$c_i$ 为学习因子，$d_{\mathrm{rand}}$ 为[0,1]随机数，$p_{\mathrm{best}}$ 为局部最优，$g_{\mathrm{best}}$ 为全局最优. 流程：①设置初值 $z_i$、$v_i$；②计算个体适应度函数 $F(i)$；③将 $F(i)$ 与个体极值 $p_{\mathrm{best}}(i)$ 比较，如果 $F(i)>p_{\mathrm{best}}(i)$，那么 $F(i)$ 将 $p_{\mathrm{best}}(i)$ 替换；④将计算的个体的 $F(i)$ 与全局极值 $g_{\mathrm{best}}$ 比较，如果 $F(i)>g_{\mathrm{best}}$，那么 $F(i)$ 将 $g_{\mathrm{best}}$ 替换；⑤更新个体的 $z_i$、$v_i$；⑥达到最大迭代次数或误差小于给定精度 $\varepsilon$，停止，获得最好的 $g_{\mathrm{best}}$，否则回到②继续计算.

式(13)为变异项，在每轮迭代中生成一个随机数，如果大于给定的变异系数，则进行下式计算，目的是以一定概率跳出当前位置，防止陷入局部极小值.

$$z_i=(z_{\max}-z_{\min})d_{\mathrm{rand}}+z_{\min} \tag{13}$$

(4)利用萤火虫算法更新类中心. 使用萤火虫算法，将以不同的 $\boldsymbol{x}_i$ 为中心获得的 $G_{\mathrm{m}}$ 作为吸引度，最大 $G_{\mathrm{m}}$ 下的 $\boldsymbol{x}_i$ 与 $R_i$ 作为最大吸引度的点. 其余点根据不同的吸引度进行移动，得到新的 $\boldsymbol{x}_{1\mathrm{new}}^+$、$\boldsymbol{x}_{2\mathrm{new}}^+$、…、$\boldsymbol{x}_{5\mathrm{new}}^+$ 和 $\boldsymbol{x}_{1\mathrm{new}}^-$、$\boldsymbol{x}_{2\mathrm{new}}^-$、…、$\boldsymbol{x}_{5\mathrm{new}}^-$，$R_0$ 也以同样的方式改变. 待点更新结束后，先以第(2)步的方法进行隶属度赋予，再代入第(3)步进行计算. 如此往复，直到运算达到次数要求.

此步骤使用了萤火虫算法，方法是亮度高的个体会将其他亮度低于自己的个体吸引过来. 个体间的吸引度为 $\beta(r)=\beta_0\exp(-\eta r_{ij})$，其中 $\eta$ 为吸引度系数，$r_{ij}$ 为个体间距离，$\beta_0$ 为最大吸引度. 最优迭代公式为

$$l_i(t+1)=l_i(t)+\beta(l_i(t)-l_j(t))+\theta(d_{\mathrm{rand}}-1/2) \tag{14}$$

其中 $l_i$、$l_j$ 表示个体的位置，$\theta$ 为步长因子. 具体流程：①设置初始参数；②随机设置初始点，计算个体的最大亮度；③计算群体中个体的相对亮度和吸引度，以亮度决定移动方向；④更新个体位置，最好的个体随机移动；⑤更新全部个体位置后，重新计算个体的亮度. 满足最大迭代次数或精度要求后，输出最优极值点，如未满足，则返回③继续迭代.

算法迭代结束后，以最高的 $G_{\mathrm{m}}$ 对应的 $s_i$、$C$ 和 $\gamma$ 作为最优参数，训练得到最优的 FSVM 分类模型.

## 3　实验结果与分析

数值实验在 2.9 GHz/4.0 GB 的计算机上使用 Matlab 2021 中的 libsvm 工具包实现. 本文使用 UCI 数据集中有关物理、生物、医学和工业生产等类型的数据进行训练和测试. 参数 $C$ 的候选区间为[1,2 000]，$\gamma$ 的候选区间为[$1\times10^{-6}$,1].

当数据集不平衡比例过大时采用文献[9]的方式，给多数类样本以小的惩罚，给少数类样本以大的惩罚，尽可能防止超平面向少数类偏移. 算法模型中的式(1)更新为

$$\begin{aligned}\min_{\boldsymbol{\omega},b,\boldsymbol{\xi}}\quad & \frac{1}{2}\|\boldsymbol{\omega}\|^2+C^+\sum_{i=1}^{p}s_i^+\xi_i+C^-\sum_{i=p+1}^{N}s_i^-\xi_i\\ \text{s.t.}\quad & y_i(\boldsymbol{\omega}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i;\ i=1,2,\cdots,N\\ & \xi_i\geqslant 0;\ i=1,2,\cdots,N\end{aligned}\tag{15}$$

其中 $p$ 表示少数类样本个数，$C$ 的设定为 $C^+=C^-(N-p)/p$，$N-p$ 则为多数类样本个数.

各个数据集的特征见表 1.

表 1 UCI 数据集特征

Tab. 1 UCI dataset characteristics

| 数据集 | 特征维数 | 训练样本数 | 数据所在领域 |
|---|---|---|---|
| Ionosphere | 34 | 350 | 电离层 |
| India | 8 | 768 | 印第安人糖尿病 |
| Haberman | 3 | 306 | 哈伯曼癌症 |
| Ecoli | 7 | 336 | 大肠杆菌 |
| Glass | 9 | 213 | 玻璃类型 |
| Yeast | 8 | 1 484 | 酵母 |
| Abalone | 7 | 4 177 | 鲍鱼 |

选择不同的初始半径 $R_0$ 时算法分类结果如图 2 所示，当 $R_0$ 高于 $7R_{\max}/10$ 时有着更高的分类精度，并且高于 $7R_{\max}/10$ 后得到的分类结果相差不大，低于 $7R_{\max}/10$ 时的结果不理想. 原因一是将过多的样本无差别剔除，减少了大量样本后，降低了数据集内含有的信息；二是初始 $G_m$ 过低会导致萤火虫算法和粒子群算法的搜索效率降低，较难向更高的 $G_m$ 移动，从而难以得到更高的分类精度. 故本文选取分类精度相对较高的 $R_0=8R_{\max}/10$ 对数据集进行实验.

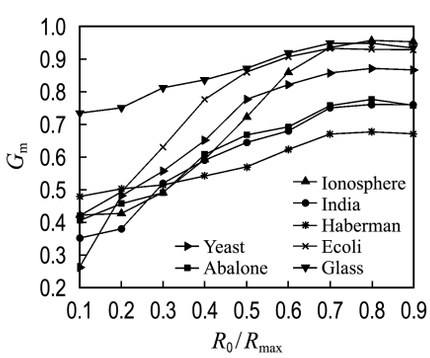


图 2 不同 $R_0$ 时得到的结果

Fig. 2 Results obtained at different $R_0$

将本文的结果与已有的常见算法进行对比，使用的性能评价指标分别为敏感性 $S_n$、特异性 $S_p$ 和几何平均值 $G_m$，定义为

$$S_n=T_p/(T_p+F_n)\tag{16}$$

$$S_p=T_n/(T_n+F_p)\tag{17}$$

$$G_m=\sqrt{S_nS_p}\tag{18}$$

其中 $T_p$、$F_p$、$F_n$、$T_n$ 的含义如表 2 所示. $S_n$ 与 $S_p$ 越高表示分类准确率越高，$G_m$ 则用来反映分类器的综合性能.

表 2 混淆矩阵内容

Tab. 2 Confusion matrix contents

| 预测值 | 真实值 | |
|---|---|---|
| | 正类 | 负类 |
| 正类 | $T_p$ | $F_p$ |
| 负类 | $F_n$ | $T_n$ |

从表 3 的结果可以看出，并非所有标准的 FSVM 分类效果都优于 SVM（如数据集 Ionosphere、India），原因之一是隶属度函数设计得不够好，或者是以固定的方式赋予隶属度时，并未尝试测试对隶属度函数做进一步修正后，算法能否得到更优的分类结果. 首先，本文在数据类中心的寻找上并未使用传统意义上求平均值的方法，而是采用了聚类的方式，这样可以获取并使用数据中具有较高代表性的样本，有助于在下一步隶属度函数迭代中获得较高的初始分类精度. 其次，通过比较每次迭代得到的分类精度来更新隶属度函数并进行训练和测试. 从最终结果来看，本文算法分类精度不仅超过标准的 FSVM 与 SVM，在一些数据集上的表现已超越现有的常用算法，使得每个样本对最终的决策超平面有着相对合理的影响，有效降低了噪声和异常值对最优超平面的干扰. 当数据的正、负类样本数量差距过大时，标准的 SVM 与 FSVM 已经难以处理，其得到的 $G_m$ 低于 0.5，算法的分类效果甚至不如不需要任何计算的随机猜测法，而且在 Abalone 数据集下少数类的分类正确率为 0，已经无法达到预期的分类目的. 所以本文在大比例不平衡数据集下运用文献[9]的结论，在算法层面进行改进，之后再代入本文所提出的动态模型进行训练，从分类效果可以发现其已经跳出了标准 SVM 算法存在的缺陷，并且 $G_m$ 与现有算法相比也得到了提升. 在 Glass 数据集上，由于数据是绝对不平衡的，即少数类样本数量非常少，本文并未合成新的

表 3　本文算法与其他算法在 UCI 数据集下的比较结果

Tab. 3　Comparison results of this algorithm with other algorithms in UCI dataset

| 数据集名称及其不平衡比 | 算法 | $S_n$ | $S_p$ | $G_m$ |
|---|---|---|---|---|
| Ionosphere 不平衡比 (1∶1.8) | SVM | 0.954 8 | 0.945 8 | 0.950 3 |
| | FSVM | 0.906 3 | 0.972 0 | 0.938 6 |
| | DEC-SVM | 0.954 8 | 0.949 3 | 0.952 0 |
| | 文献[7]算法 | 0.930 9 | 0.952 4 | 0.941 6 |
| | 文献[10]算法 | 0.958 7 | 0.947 1 | 0.952 9 |
| | 文献[11]算法 | 0.949 5 | 0.956 1 | 0.952 8 |
| | 本文算法 | 0.957 4 | 0.974 7 | 0.966 0 |
| India 不平衡比 (1∶1.9) | SVM | 0.564 2 | 0.805 6 | 0.674 1 |
| | FSVM | 0.560 1 | 0.808 2 | 0.672 7 |
| | DEC-SVM | 0.744 0 | 0.728 6 | 0.736 2 |
| | 文献[7]算法 | 0.777 8 | 0.708 3 | 0.742 2 |
| | 文献[10]算法 | 0.739 9 | 0.740 8 | 0.739 8 |
| | 文献[11]算法 | 0.774 6 | 0.727 5 | 0.750 7 |
| | 本文算法 | 0.753 2 | 0.752 5 | 0.752 9 |
| Haberman 不平衡比 (1∶2.6) | SVM | 0.351 9 | 0.793 3 | 0.528 0 |
| | FSVM | 0.364 2 | 0.768 4 | 0.528 5 |
| | DEC-SVM | 0.522 2 | 0.797 8 | 0.645 4 |
| | 文献[7]算法 | 0.638 9 | 0.661 8 | 0.650 2 |
| | 文献[10]算法 | 0.538 3 | 0.790 7 | 0.652 3 |
| | 文献[11]算法 | 0.583 3 | 0.733 0 | 0.653 9 |
| | 本文算法 | 0.555 6 | 0.791 1 | 0.662 9 |
| Ecoli 不平衡比 (1∶3.3) | SVM | 0.809 1 | 0.939 0 | 0.871 6 |
| | FSVM | 0.835 1 | 0.927 0 | 0.879 8 |
| | DEC-SVM | 0.967 5 | 0.842 9 | 0.903 0 |
| | 文献[7]算法 | 0.953 8 | 0.838 0 | 0.894 0 |
| | 文献[10]算法 | 0.970 1 | 0.842 9 | 0.904 2 |
| | 文献[11]算法 | 0.942 9 | 0.864 4 | 0.902 8 |
| | 本文算法 | 0.959 2 | 0.870 4 | 0.913 7 |
| Glass 不平衡比 (1∶15) | SVM | 0.980 1 | 0.876 9 | 0.926 5 |
| | FSVM | 0.980 1 | 0.761 5 | 0.863 8 |
| | DEC-SVM | 0.971 1 | 0.923 1 | 0.946 8 |
| | 文献[7]算法 | 0.956 0 | 0.923 1 | 0.939 4 |
| | 文献[10]算法 | 0.967 2 | 0.923 1 | 0.944 9 |
| | 文献[11]算法 | 0.960 2 | 0.923 1 | 0.941 5 |
| | 本文算法 | 0.975 0 | 0.923 1 | 0.948 7 |
| Yeast 不平衡比 (1∶28) | SVM | 0.182 4 | 0.994 8 | 0.424 5 |
| | FSVM | 0.235 3 | 0.991 6 | 0.482 8 |
| | DEC-SVM | 0.837 3 | 0.860 9 | 0.849 0 |
| | 文献[7]算法 | 0.704 5 | 0.909 1 | 0.800 3 |
| | 文献[10]算法 | 0.823 5 | 0.877 7 | 0.850 2 |
| | 文献[11]算法 | 0.851 1 | 0.862 0 | 0.856 5 |
| | 本文算法 | 0.837 2 | 0.888 7 | 0.862 6 |
| Abalone 不平衡比 (1∶41) | SVM | 0 | 1.000 0 | 0 |
| | FSVM | 0 | 1.000 0 | 0 |
| | DEC-SVM | 0.790 3 | 0.715 0 | 0.751 6 |
| | 文献[7]算法 | 0.583 8 | 0.888 9 | 0.720 4 |
| | 文献[10]算法 | 0.802 9 | 0.716 9 | 0.758 6 |
| | 文献[11]算法 | 0.808 5 | 0.710 0 | 0.757 7 |
| | 本文算法 | 0.767 4 | 0.752 9 | 0.760 1 |

少数类样本使其变为相对不平衡数据集，以至于最终分类精度并未获得明显提升. 当数据集更偏向于球形分布时，本文算法会得到相对更高的分类精度. 但本文算法复杂度较高，在验证过程中耗时较大，并且具有随机搜索算法的缺点，结果可能会陷入局部极小值无法跳出，导致得到的参数不是最优解，无法获得更高的分类精度，当分类精度过低时就需要重新训练，再次大量耗时. 为了解决这一问题，在粒子群算法更新参数后进行变异操作，即在迭代中生成一个随机数，若此数大于设定好的变异系数，则通过式(13)计算出新的参数，以一定概率跳出局部极小值，从而在新的区间内寻找最优解.

## 4　结　语

针对现有的大部分算法都是直接赋予样本一个固定的隶属度，并未考虑对隶属度函数做进一步修正的问题，本文提出了通过实时更改类中心位置和隶属度函数的方法来获得更加合适的隶属度，并同时优化模糊支持向量机参数，有效降低了噪声和异常点的影响，提升了分类精度. 今后将尝试在本文算法中使用改进的粒子群算法，而且在不影响分类精度的前提下降低算法复杂度，还将研究如何在处理不平衡数据集问题上提出新的数据预处理方式.

## 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 Vector Machines and Other Kernel-based Learning Methods** [M]. 1st ed. 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] 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)

[6] WU Zhenning, ZHANG Huaguang, LIU Jinhai. A fuzzy support vector machine algorithm for classification based on a novel PIM fuzzy clustering method [J]. **Neurocomputing**, 2014, **125**: 119-124.

[7] 陈继强，余志鹏，张 峰，等. 基于异类类内超平面的模糊支持向量机及其应用 [J]. 河北工程大学学报：自然科学版，2020，**37**(4)：99-104.
CHEN Jiqiang, YU Zhipeng, ZHANG Feng, *et al*. Fuzzy support vector machine based on heterogeneous class internal hyperplane and its application [J]. **Journal of Hebei University of Engineering: Natural Science Edition**, 2020, **37**(4): 99-104. (in Chinese)

[8] 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)

[9] 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.

[10] 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)

[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. Double feature-weighted fuzzy support vector machine [J]. **Journal of Computer Applications**, 2022, **42**(3): 683-687. (in Chinese)

[12] 周国华，卢剑炜，顾晓清，等. 基于简约凸壳的一类模糊支持向量机 [J]. 电子学报，2019，**47**(8)：1708-1716.
ZHOU Guohua, LU Jianwei, GU Xiaoqing, *et al*. One-class fuzzy support vector machines based on reduced convex hull [J]. **Acta Electronica Sinica**, 2019, **47**(8): 1708-1716. (in Chinese)

[13] PAN Lei, LUO Yi. Parameters selection of support vector machine using an improved PSO algorithm [C]// **Proceedings - 2010 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2010**. Nanjing: IEEE, 2010: 196-199.

# Fuzzy support vector machine based on dynamic class-center model selection

SONG Yiming, JU Zhe*

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

**Abstract**: The core idea of fuzzy support vector machine is to give the fuzzy membership to the samples and give different weights to each sample, so as to overcome the problems that the standard support vector machine is sensitive to noise and outliers. The existing fuzzy support vector machine algorithms usually assign a fixed membership to each sample based on the distance between the sample and the class-center, without further modifying the membership according to the sample distribution. A new dynamic mode is proposed to assign membership to the sample. The firefly algorithm is used to update the position and membership function of the sample center constantly, and the particle swarm optimization algorithm is used to optimize the parameters of fuzzy support vector machine. Experimental results on UCI dataset show that the proposed algorithm can effectively reduce the influence of noise and wild points on the hyperplane, and the classification performance is better than that of several common fuzzy support vector machine algorithms.

**Key words**: fuzzy support vector machine; membership function; classification; particle swarm optimization algorithm; firefly algorithm