文章编号：1000-8608(2016)05-0525-07

用于不平衡数据分类的模糊支持向量机算法

鞠 哲，曹隽喆，顾 宏*

（大连理工大学 控制科学与工程学院，辽宁 大连 116024）

摘要： 作为一种有效的机器学习技术，支持向量机已经被成功地应用于各个领域.然而当数据不平衡时，支持向量机会产生次优的分类模型；另一方面，支持向量机算法对数据集中的噪声点和野点非常敏感.为了克服以上不足，提出了一种新的用于不平衡数据分类的模糊支持向量机算法.该算法在设计样本的模糊隶属度函数时，不仅考虑训练样本到其类中心距离，而且考虑样本周围的紧密度.实验结果表明，所提模糊支持向量机算法可以有效地处理不平衡和噪声问题.

关键词： 支持向量机；模糊支持向量机；模糊隶属度；不平衡数据；分类

中图分类号： TP181 **文献标识码：** A **doi：** 10.7511/dllgxb201605013

0 引 言

支持向量机(support vector machine，SVM)是建立在统计学习中的VC维理论和结构风险最小化原则基础上的一种机器学习方法，能有效地处理小样本、高维数据、非线性等问题.作为一种常用的机器学习技术，SVM[1-2]已经被成功地应用到各种实际分类问题中.然而，真实的数据集常常含有噪声或野点.传统的SVM算法等同地对待所有训练样本并赋予它们统一的权值，因此SVM算法对训练样本中噪声和野点非常敏感[3].为了克服噪声数据对SVM分类结果的影响，Lin等[4]首次提出了模糊支持向量机(FSVM)算法.在FSVM算法中，训练集中不同的训练样本被赋予不同的模糊隶属度(即权值)来衡量样本对分类器的重要程度.Lin等[4]认为如果一个样本离其类中心越近，那么它就越有可能属于该类，需分配给该样本一个较高的模糊隶属度；相反，如果这个样本远离其类中心，那么它就越有可能是噪声或野点，需分配给该样本一个较小的模糊隶属度.FSVM算法是对传统SVM算法的一种改进，在一定程度上降低了SVM算法对噪声或野点的敏感度.然而，FSVM算法在设计模糊隶属度时只考虑了样本到其类中心的距离作为衡量样本的重要性指标.对于不规则分布的数据集，这种方法可能会将噪声样本当作正常样本进行训练，因而影响了算法的精度.随后，一些改进的FSVM算法被相继提出，如文献[5]利用核方法将FSVM算法在核空间中实现；文献[6]提出了一种基于样本间紧密度的FSVM算法；文献[7]提出了一种利用样本到其类内超平面来设计模糊隶属度的FSVM算法；文献[8]提出了一种基于类向心度的改进FSVM算法，算法在设计模糊隶属度时不仅考虑到样本到类中心的距离，而且考虑到样本之间的内在联系.

此外，尽管SVM算法在平衡数据集上具有良好的性能表现，但当数据不平衡时SVM算法的分类效果不佳.SVM算法偏向于保证多数类的分类精度，而在少数类上分类效果往往较差.目前，许多算法被用来处理基于不平衡数据的分类问题，以便提高SVM算法在不平衡数据集上的分类性能.从数据层面，将SVM算法和欠采样或过采样技术相结合来平衡正负类的样本比例[9].但采样方法实际上通过增加少数类样本或减少多数类样本来调整数据集的不平衡性，具有一定的

收稿日期： 2016-03-30； **修回日期：** 2016-04-11.
基金项目： 国家自然科学基金资助项目(61502074，U1560102)；高等学校博士学科点专项科研基金资助项目(20120041110008).
作者简介： 鞠 哲(1986-)，男，博士生，E-mail：juzhe1120@hotmail.com；顾 宏*(1961-)，男，教授，博士生导师，E-mail：guhong@dlut.edu.cn.

盲目性,结果的稳定性难以得到保证,而且不适用于不平衡比例较大的数据集.从算法层面,对SVM算法本身进行改进来处理不平衡数据分类.如文献[10]提出了一种基于实例重要性的不平衡SVM分类算法.算法首先将数据按其重要性排序,选择最重要的数据训练初始分类器,然后算法循环迭代.该算法在数据样本太大或太小时效果较差.文献[11]通过赋予正负类样本不同的惩罚因子(different error costs,DEC)来降低不平衡数据给SVM算法带来的影响.一个简单有效的处理方法是将惩罚因子直接设置为正负类数据的不平衡比率[12].文献[13]提出了一种改进近似支持向量机算法,该算法不仅赋予正负类样本以不同的惩罚因子,且在约束条件中增加新的参数,使分类面更具灵活性,提高了算法精度.

然而,现实中的数据集往往含有噪声,且正负类样本比例不均衡.虽然FSVM算法可以在一定程度上克服噪声数据对SVM算法的影响,但是这些算法不能有效地处理不平衡数据上的分类问题.而不平衡SVM算法虽然能处理不平衡数据分类,但是却容易受到数据集中噪声或野点的影响.对此,Batuwita等[14]将FSVM算法与DEC算法结合,提出了一种可以处理不平衡且含噪声数据分类的不平衡FSVM算法——FSVM-CIL.然而,FSVM-CIL中的FSVM-CIL$_{\text{lin}}^{\text{cen}}$、FSVM-CIL$_{\text{exp}}^{\text{cen}}$算法与FSVM算法一样都是基于样本到类中心距离来设计模糊隶属度.针对FSVM算法在设计模糊隶属度方面的不足,本文不仅考虑样本到类中心的距离,还利用样本周围的紧密度来设计模糊隶属度,并将改进的FSVM算法与DEC算法结合,提出一种新的用于不平衡数据分类的FSVM算法,以解决不平衡且含有噪声数据的分类问题.

1 SVM简介

对于二分类问题,SVM的基本思想就是在样本(核)空间寻找一个最优超平面,使得两类样本的分类间隔达到最大.给定训练集$(\boldsymbol{X},T)=\{(\boldsymbol{x}_i,t_i),i=1,2,\cdots,l\}$,其中$\boldsymbol{x}_i$为样本,$t_i$为样本$\boldsymbol{x}_i$的标签,$t_i\in\{-1,1\}$;引入非线性映射$\boldsymbol{\Phi}(\boldsymbol{x})$,将训练集映入高维空间$(\boldsymbol{\Phi}(\boldsymbol{X}),T)=\{(\boldsymbol{\Phi}(\boldsymbol{x}_i),t_i),i=1,2,\cdots,l\}$;选取适当的核函数$K(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{\Phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{y})$;引入松弛变量$\xi_i\geqslant 0$, $i=1,2,\cdots,l$.标准支持向量机的一般形式可以表示为

$$\begin{aligned}&\min_{\boldsymbol{\omega},\boldsymbol{\xi}}\ \frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{l}\xi_i^2\\&\text{s.t.}\ \ t_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i\\&\qquad\xi_i\geqslant 0,\ i=1,2,\cdots,l\end{aligned}\tag{1}$$

求解优化问题(1)的对偶问题:

$$\begin{aligned}&\min\ \ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i\\&\text{s.t.}\ \ \sum_{i=1}^{l}t_i\alpha_i=0\\&\qquad 0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,l\end{aligned}\tag{2}$$

假设对偶问题的解为α^*,则最优超平面的法向量$\boldsymbol{\omega}^*=\sum_{i=1}^{l}\alpha^* t_i\boldsymbol{\Phi}(\boldsymbol{x}_i)$.取某个$0<\alpha_j^*<C(j=1,2,\cdots,l)$对应的$\boldsymbol{x}_j$、$t_j$,可求得$b^*=t_j-\sum_{i=1}^{l}t_i\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_j)$.由此可得决策函数$f(\boldsymbol{x})=\sum_{i=1}^{l}t_i\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x})+b^*$.

2 用于不平衡数据分类的FSVM算法

传统的SVM算法认为,每一个样本的重要性是相同的,算法分配给每一个样本相同的权值.在实际应用中,数据往往是不平衡的,而且包含噪声.因此,一个合理的做法是根据样本不平衡性和样本重要性来分配不同的权值.对于二分类问题,给定训练集$(\boldsymbol{X},T)=\{(\boldsymbol{x}_i,t_i),i=1,2,\cdots,l\}$,其中$\boldsymbol{x}_i$为样本,$t_i$为样本$\boldsymbol{x}_i$的标签,$t_i\in\{-1,1\}$.不失一般性,假设前$p$个样本是正类样本(即$t_i=1$, $i=1,2,\cdots,p$),剩下后$l-p$个样本是负类样本(即$t_i=-1,i=p+1,p+2,\cdots,l$).不平衡FSVM的一般形式可以表示为

$$\begin{aligned}&\min_{\boldsymbol{\omega},\boldsymbol{\xi}}\ \frac{1}{2}\|\boldsymbol{\omega}\|^2+C^+\sum_{i=1}^{p}s_i^+\xi_i^2+C^-\sum_{i=p+1}^{l}s_i^-\xi_i^2\\&\text{s.t.}\ \ t_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i\\&\qquad\xi_i\geqslant 0,\ i=1,2,\cdots,l\end{aligned}\tag{3}$$

其中$\boldsymbol{\Phi}(\boldsymbol{x})$是非线性映射;$\xi_i(i=1,2,\cdots,l)$是松弛变量;$C^+$和$C^-$分别是正负类样本的惩罚因子,用来反映类间不平衡性;s_i^+和s_i^-是模糊隶属度函数,用来反映样本在其所属类中的重要性.可以看出,当$C^+=C^-$且$s_i^+=s_i^-=1$时,式(3)退化为传统的SVM算法;而当$C^+>C^-$且$s_i^+=s_i^-=1$

时，式(3)退化为 DEC 算法.

2.1　模糊隶属度设计

在文献[4]中模糊隶属度被定义为

$$s_i^+=1-\frac{d_i^{\text{cen}+}}{\max\limits_j(d_j^{\text{cen}+})+\delta};\ i=1,2,\cdots,p \tag{4}$$

$$s_i^-=1-\frac{d_i^{\text{cen}-}}{\max\limits_j(d_j^{\text{cen}-})+\delta};\ i=p+1,p+2,\cdots,l \tag{5}$$

其中 $d_i^{\text{cen}+}=\left\|\boldsymbol{x}_i-\frac{1}{p}\sum_{j=1}^{p}\boldsymbol{x}_j\right\|$，$d_i^{\text{cen}-}=\left\|\boldsymbol{x}_i-\frac{1}{l-p}\times\sum_{j=p+1}^{l}\boldsymbol{x}_j\right\|$；$\delta$ 是一个非常小的正数，用来保证模糊隶属度大于 0. 然而，这种方法只是将样本到其类中心的距离作为衡量样本的重要性的指标. 对于不规则分布的数据集，这种方法可能会将噪声样本当作正常样本进行训练，因而影响了算法的精度. 图 1 为带有一个噪声点的椭圆分布数据. 由图 1 可以看出，噪声点 x_5 到类中心的距离和正常样本点 x_1 到类中心的距离相等，按照 Lin 等[4]的方法，将会赋予该噪声点 x_5 和正常样本点 x_1 以相同的模糊隶属度，这显然是不符合实际情况的. 从图 1 可以明显看到，噪声点 x_5 的周围比较稀疏，而正常样本点 x_1 的周围比较稠密.

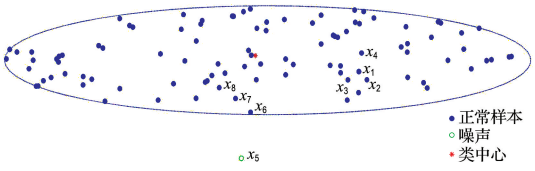

图 1　带有一个噪声点的椭圆分布数据示意图

Fig. 1　Sketch map for elliptical distribution data with a noise sample

基于上述观察，本文除了考虑样本到类中心的距离，还将样本周围的紧密度作为设计模糊隶属度的依据. 一个简单的 K-近邻准则用来衡量样本的紧密度. 如图 1 所示，噪声点 x_5 到其 3-近邻 $\{x_6,x_7,x_8\}$ 的平均距离要远远大于正常样本点 x_1 到其 3-近邻 $\{x_2,x_3,x_4\}$ 的平均距离. 因此，对于一个正样本 $\boldsymbol{x}_i$，它周围的紧密度可以定义为

$$D_i^+=\frac{1}{K}\sum_{\boldsymbol{x}_j\in N_K^+(\boldsymbol{x}_i)}\|\boldsymbol{x}_i-\boldsymbol{x}_j\| \tag{6}$$

其中 $N_K^+(\boldsymbol{x}_i)$ 记为 $\boldsymbol{x}_i$ 在正类样本集中的 K 个近邻组成的集合；类似地，对于一个负样本 $\boldsymbol{x}_i$，它周围的紧密度可以定义为

$$D_i^-=\frac{1}{K}\sum_{\boldsymbol{x}_j\in N_K^-(\boldsymbol{x}_i)}\|\boldsymbol{x}_i-\boldsymbol{x}_j\| \tag{7}$$

其中 $N_K^-(\boldsymbol{x}_i)$ 记为 $\boldsymbol{x}_i$ 在负类样本集中的 K 个近邻组成的集合. 直观地说，如果 D_i^+(D_i^-)越小，表明该样本周围越稠密，则它越有可能属于正(负)类；相反，如果 D_i^+(D_i^-)越大，表明该样本周围越稀疏，则它越有可能属于噪声或野点. 基于样本到类中心的距离(式(4)、(5))和样本周围的紧密度(式(6)、(7))，本文的模糊隶属度定义如下：

$$s_i^+=\left(1-\alpha\frac{d_i^{\text{cen}+}}{\max\limits_j(d_j^{\text{cen}+})+\delta}-(1-\alpha)\frac{D_i^+-\min\limits_j(D_j^+)}{\max\limits_j(D_j^+)-\min\limits_j(D_j^+)+\delta}\right)^m;\quad i=1,2,\cdots,p \tag{8}$$

$$s_i^-=\left(1-\alpha\frac{d_i^{\text{cen}-}}{\max\limits_j(d_j^{\text{cen}-})+\delta}-(1-\alpha)\frac{D_i^--\min\limits_j(D_j^-)}{\max\limits_j(D_j^-)-\min\limits_j(D_j^-)+\delta}\right)^m;\quad i=p+1,p+2,\cdots,l \tag{9}$$

其中 $\alpha\in[0,1]$，$m>0$，δ 是一个很小的正数来保证分母大于 0. 在本文中，δ 取值为 0.000 1，α 的选择范围是 $\{0,0.1,\cdots,1.0\}$，m 的选择范围是 $\{0.1,0.2,\cdots,1.0\}$. 此外，为了节省算法的训练时间，本文模糊隶属度中 K 统一设置为 10.

2.2　惩罚因子设置

一般来说，DEC 算法[11]通过赋给少数类以较大的权值、多数类以较小的权值，可以有效地减小不平衡性对 SVM 算法的影响. 然而 DEC 算法中，没有根据样本的重要性对类内样本加以区分，使得 DEC 算法很容易受到噪声或野点的影响. 而本文将模糊隶属度和惩罚因子结合起来，不仅可以较好地处理不平衡数据分类问题，而且可以克服噪声数据对 SVM 分类结果的影响. 根据文献[11-12]的结果，当 C^+/C^- 设置为多数类与少数类样本个数的比值时，SVM 算法可以得到较好的分类结果. 本文中惩罚因子 C^+ 设置为 $C(l-p)/p$，C^- 设置为 $C(C>0)$.

3　实验与结果分析

为了验证本文所提算法的有效性，将本文算法与 SVM、FSVM、DEC、FSVM-CIL$_{\text{lin}}^{\text{cen}}$ 和 FSVM-CIL$_{\text{exp}}^{\text{cen}}$ 算法进行比较. 从 UCI 数据库选取 8 个分

类数据集作为实验数据集，这些数据集的属性如表 1 所示. 这些数据集不平衡且不可避免地包含一些噪声或野点. 将上述算法分别在每一个 UCI 数据集上执行 10 折交叉验证. 为了保证结果的可靠性，每个数据集上进行 10 次 10 折交叉验证，取其 10 次均值作为最后结果. 实验中所有 SVM 算法均采用 RBF 核函数：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{\Phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}_j)=\exp(-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)$. 使用网格搜索法对惩罚因子 C 和核参数 γ 进行选择. C 的选择范围是 $\{2^0,2^1,\cdots,2^{15}\}$，$\gamma$ 的选择范围是 $\{2^{-15},2^{-14},\cdots,2^0\}$. LibSVM 软件包[15]用来训练本文中所有的 SVM 模型，LibSVM-weights-3.20 软件包用来训练本文中所有的 FSVM 模型. LibSVM-weights-3.20 软件包可以分配给每个样本不同的权值. 所有数值实验均在 2.7 GHz/2.0 GB 的 PC 机上利用 Matlab R2009 软件实现.

表 1 UCI 数据集及其相关属性

Tab.1 UCI datasets and related properties

数据集	正样本数	负样本数	样本总数	不平衡比	类别总数	正类所属类别
Ionosphere	126	225	351	1.78	2	1
Pima-Indians	268	500	768	1.87	2	1
Haberman	81	225	306	2.78	2	2
Transfusion	178	540	748	3.03	2	1
Ecoli	77	259	336	3.36	8	2
Glass	13	201	214	15.46	7	5
Yeast	51	1 433	1 484	28.10	10	5
Abanole	103	4 074	4 177	39.55	29	15

采用 3 个常用的分类器性能评价指标用来评价算法的表现，分别是敏感性 S_n、特异性 S_p 和几何平均值 G_m，定义如下：

$$S_n=\frac{T_p}{T_p+F_n} \tag{10}$$

$$S_p=\frac{T_n}{T_n+F_p} \tag{11}$$

$$G_m=\sqrt{S_nS_p} \tag{12}$$

其中 T_p、T_n、F_p 和 F_n 分别表示真阳性样本个数、真阴性样本个数、假阳性样本个数和假阴性样本个数. S_n 和 S_p 分别反映了分类器正确预测正负类样本的比率，显然 S_n 和 S_p 的值越大表明分类器性能越好. 然而 S_n 和 S_p 相互制约，尤其当数据不平衡时，很难同时用这两个指标反映分类器的性能. 于是引入它们的几何平均值 G_m 来反映分类器在不均衡数据上的性能. G_m 的值越大，表明分类器的综合性能越好.

本文算法与 SVM、FSVM、DEC、FSVM-CIL$_{\mathrm{lin}}^{\mathrm{cen}}$ 和 FSVM-CIL$_{\mathrm{exp}}^{\mathrm{cen}}$ 算法在上述 UCI 数据集上的比较结果由表 2 给出. 从表 2 可以看出，即使在数据不平衡比例较小的数据集（Ionosphere、Pima-Indians 和 Transfusion）上，FSVM 算法的表现也未必会优于传统的 SVM 算法. 这是因为数据分布有可能不是标准的球形分布，这样仅用样本到类中心的距离来设计模糊隶属度就会出现问题. 而且，由于 SVM、FSVM 算法没有考虑不平衡因素，在不平衡度较大的数据集（Yeast 和 Abanole）上，这两种算法表现很差. 虽然 DEC 算法考虑到数据的不平衡性，但没有考虑各类样本对其类的重要性，容易受到数据集中噪声或野点的影响，其表现结果远不如 FSVM-CIL$_{\mathrm{lin}}^{\mathrm{cen}}$、FSVM-CIL$_{\mathrm{exp}}^{\mathrm{cen}}$ 和本文提出的 FSVM 算法. FSVM-CIL$_{\mathrm{lin}}^{\mathrm{cen}}$、FSVM-CIL$_{\mathrm{exp}}^{\mathrm{cen}}$ 算法的模糊隶属度设计与 FSVM 算法相同，都是仅考虑样本到类中心的距离，因而在不规则分布的数据上，这些算法不能很好地反映样本的重要程度. 而本文的算法在设计模糊隶属度时，不仅考虑到样本到类中心的距离，而且考虑到样本周围的紧密度，因此本文所提算法要优于 FSVM-CIL$_{\mathrm{lin}}^{\mathrm{cen}}$ 和 FSVM-CIL$_{\mathrm{exp}}^{\mathrm{cen}}$ 算法. 由表 2 可以看到，在各种不平衡比例的 UCI 数据集上，本文提出的算法都取得了最大的 G_m 值. 这表明在带有噪声的不平衡数据集上，本文提出的算法表现都要优于 SVM、FSVM、DEC、FSVM-CIL$_{\mathrm{lin}}^{\mathrm{cen}}$ 和 FSVM-CIL$_{\mathrm{exp}}^{\mathrm{cen}}$ 算法. 然而，在分类精度提升的同时，本文所提算法需要优化的模型参数也随之增多. 与 SVM 算法相比，尽管本文算法在计算模糊隶属度时需要一部分额外的计算时间（模糊隶属度的复杂度为 $O(l)$，l 为样本数目），但从理论上说，本文所提算法的复杂度跟 SVM 算法的复杂度相当，都是 $O(l^3)$. 由于本文算法在设计模糊隶属度时引入了 3 个额外的参数：α、m 和 K，所以在 10 折交叉验证过程中需要更多的训练时间，特别是在训练集较大的样本上训练时间过长，这是本文算法需要改进的一个地方. 表 3 给出了这些算法在各个 UCI 数据集上的最优参数.

表 2　8 个 UCI 数据集上的分类结果比较

Tab. 2　Comparison of the classification results on eight UCI datasets

数据集	算法	S_n	S_p	G_m
Ionosphere	SVM	0.954 8±0.005 4	0.945 8±0.004 1	0.950 3±0.003 8
	FSVM	0.906 3±0.012 9	0.972 0±0.003 7	0.938 6±0.007 5
	DEC	0.954 8±0.005 4	0.949 3±0.003 1	0.952 0±0.002 8
	FSVM-CIL$_{lin}^{cen}$	0.954 8±0.005 4	0.947 6±0.004 6	0.951 1±0.003 4
	FSVM-CIL$_{exp}^{cen}$	0.957 1±0.004 1	0.948 4±0.003 1	0.952 8±0.002 8
	本文算法	0.958 7±0.003 3	0.947 1±0.003 3	0.952 9±0.002 4
Pima-Indians	SVM	0.564 2±0.005 4	0.805 6±0.007 9	0.674 1±0.006 3
	FSVM	0.560 1±0.015 9	0.808 2±0.007 6	0.672 7±0.009 0
	DEC	0.744 0±0.011 6	0.728 6±0.007 3	0.736 2±0.006 5
	FSVM-CIL$_{lin}^{cen}$	0.738 8±0.011 5	0.740 2±0.005 0	0.739 5±0.007 0
	FSVM-CIL$_{exp}^{cen}$	0.729 1±0.011 2	0.742 8±0.004 9	0.735 9±0.005 7
	本文算法	0.738 8±0.013 4	0.740 8±0.005 2	0.739 8±0.007 3
Haberman	SVM	0.351 9±0.024 9	0.793 3±0.020 1	0.528 0±0.020 1
	FSVM	0.364 2±0.030 9	0.768 4±0.013 5	0.528 5±0.020 8
	DEC	0.522 2±0.016 5	0.797 8±0.006 7	0.645 4±0.010 9
	FSVM-CIL$_{lin}^{cen}$	0.545 7±0.029 0	0.768 9±0.008 4	0.647 6±0.019 5
	FSVM-CIL$_{exp}^{cen}$	0.540 7±0.017 3	0.787 1±0.017 2	0.652 3±0.011 4
	本文算法	0.538 3±0.018 6	0.790 7±0.009 5	0.652 3±0.011 7
Transfusion	SVM	0.441 0±0.018 8	0.967 7±0.005 3	0.653 1±0.014 1
	FSVM	0.434 8±0.015 7	0.969 5±0.003 9	0.649 2±0.011 8
	DEC	0.812 9±0.011 3	0.796 7±0.002 7	0.804 7±0.005 6
	FSVM-CIL$_{lin}^{cen}$	0.811 8±0.007 1	0.799 5±0.003 2	0.805 6±0.003 9
	FSVM-CIL$_{exp}^{cen}$	0.811 8±0.009 3	0.798 9±0.002 9	0.805 3±0.004 7
	本文算法	0.815 2±0.011 7	0.797 9±0.003 5	0.806 5±0.005 3
Ecoli	SVM	0.809 1±0.018 4	0.939 0±0.004 0	0.871 6±0.010 5
	FSVM	0.835 1±0.017 4	0.927 0±0.007 4	0.879 8±0.012 3
	DEC	0.967 5±0.020 5	0.842 9±0.003 7	0.903 0±0.010 1
	FSVM-CIL$_{lin}^{cen}$	0.955 8±0.019 6	0.851 4±0.003 8	0.902 0±0.009 9
	FSVM-CIL$_{exp}^{cen}$	0.970 1±0.012 3	0.841 7±0.003 6	0.903 6±0.006 8
	本文算法	0.970 1±0.012 3	0.842 9±0.004 5	0.904 2±0.006 5
Glass	SVM	0.980 1±0.004 7	0.876 9±0.064 9	0.926 5±0.034 9
	FSVM	0.980 1±0.003 3	0.761 5±0.024 3	0.863 8±0.013 4
	DEC	0.971 1±0.003 1	0.923 1	0.946 8±0.001 5
	FSVM-CIL$_{lin}^{cen}$	0.947 3±0.005 3	0.923 1	0.935 1±0.002 6
	FSVM-CIL$_{exp}^{cen}$	0.973 6±0.005 8	0.923 1	0.948 0±0.002 8
	本文算法	0.974 6±0.004 9	0.923 1	0.948 5±0.002 4
Yeast	SVM	0.182 4±0.030 7	0.994 8±0.007 5	0.424 5±0.036 4
	FSVM	0.235 3±0.016 0	0.991 6±0.001 4	0.482 8±0.016 1
	DEC	0.837 3±0.009 5	0.860 9±0.002 7	0.849 0±0.005 7
	FSVM-CIL$_{lin}^{cen}$	0.819 6±0.008 3	0.879 7±0.001 7	0.849 1±0.004 5
	FSVM-CIL$_{exp}^{cen}$	0.839 2±0.008 3	0.860 4±0.002 0	0.849 8±0.004 8
	本文算法	0.823 5±0.001 6	0.877 7±0.002 7	0.850 2±0.001 3
Abanole	SVM	0	1.000 0	0
	FSVM	0	1.000 0	0
	DEC	0.790 3±0.023 9	0.715 0±0.002 2	0.751 6±0.011 2
	FSVM-CIL$_{lin}^{cen}$	0.695 1±0.026 0	0.758 6±0.002 2	0.726 1±0.012 9
	FSVM-CIL$_{exp}^{cen}$	0.803 9±0.017 6	0.715 6±0.001 7	0.758 4±0.007 9
	本文算法	0.802 9±0.021 5	0.716 9±0.001 6	0.758 6±0.009 7

表 3 各种算法在 8 个 UCI 数据集上的最优参数

Tab. 3 Optimal parameter of various algorithms on eight UCI datasets

数据集	算法	C	γ	β	α	m
Ionosphere	SVM	2^2	2^{-1}	—	—	—
	FSVM	2^3	2^{-2}	—	—	—
	DEC	2^3	2^{-1}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^3	2^{-1}	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^4	2^{-1}	0.7	—	—
	本文算法	2^3	2^{-1}	—	1.0	0.4
Pima-Indians	SVM	2^{10}	2^{-13}	—	—	—
	FSVM	2^{10}	2^{-13}	—	—	—
	DEC	2^0	2^{-11}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^1	2^{-11}	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^1	2^{-11}	0.1	—	—
	本文算法	2^1	2^{-11}	—	0.9	0.9
Haberman	SVM	2^{10}	2^{-5}	—	—	—
	FSVM	2^{15}	2^{-5}	—	—	—
	DEC	2^8	2^{-8}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^9	2^{-11}	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^8	2^{-12}	0.4	—	—
	本文算法	2^{11}	2^{-13}	—	0.1	0.7
Transfusion	SVM	2^{15}	2^0	—	—	—
	FSVM	2^{15}	2^0	—	—	—
	DEC	2^{15}	2^0	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^{15}	2^0	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^{15}	2^0	0.9	—	—
	本文算法	2^{15}	2^0	—	0.6	1.0
Ecoli	SVM	2^{13}	2^{-3}	—	—	—
	FSVM	2^{13}	2^{-2}	—	—	—
	DEC	2^5	2^{-5}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^6	2^{-2}	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^4	2^{-4}	0.1	—	—
	本文算法	2^3	2^{-1}	—	0.5	0.2
Glass	SVM	2^{10}	2^{-8}	—	—	—
	FSVM	2^8	2^{-6}	—	—	—
	DEC	2^{10}	2^{-7}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^{10}	2^{-9}	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^{11}	2^{-7}	0.5	—	—
	本文算法	2^{11}	2^{-7}	—	0.1	0.1
Yeast	SVM	2^{10}	2^0	—	—	—
	FSVM	2^{12}	2^0	—	—	—
	DEC	2^0	2^{-4}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^2	2^{-5}	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^4	2^{-8}	0.5	—	—
	本文算法	2^0	2^{-4}	—	0.6	1.0
Abanole	SVM	—	—	—	—	—
	FSVM	—	—	—	—	—
	DEC	2^3	2^{-1}	—	—	—
	FSVM-CIL$_{\text{lin}}^{\text{cen}}$	2^3	2^0	—	—	—
	FSVM-CIL$_{\text{exp}}^{\text{cen}}$	2^3	2^{-1}	0.1	—	—
	本文算法	2^2	2^0	—	0.1	0.1

4 结 语

本文提出了一种新的用于不平衡数据的FSVM算法.该算法不仅可以有效地降低不平衡数据对 SVM 造成的影响,而且可以降低数据中噪声和野点的干扰,进而提高分类器精度. UCI 数据集上的数值实验验证了该分类方法的有效性.然而需要指出的是,该算法在提升分类精度的同时,需要优化的参数也随之增多.下一步的研究工作是设计有效的参数选择策略来缩短算法的训练时间.

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]. Cambridge: Cambridge University Press, 2000.

[3] ZHANG Xue-gong. Using class-center vectors to build support vector machines [C] // **Proceedings of the 1999 IEEE Signal Processing Society Workshop**. Madison:IEEE, 1999:3-11.

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

[5] JIANG Xiu-feng, YI Zhang, LV Jian-cheng. Fuzzy SVM with a new fuzzy membership function [J]. **Neural Computing & Applications**, 2006, **15**(3): 268-276.

[6] 张 翔,肖小玲,徐光祐. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报, 2006, **17**(5): 951-958.
ZHANG Xiang, XIAO Xiao-ling, XU Guang-you. Fuzzy support vector machine based on affinity among samples [J]. **Journal of Software**, 2006, **17**(5):951-958. (in Chinese)

[7] 杜 喆,刘三阳,齐小刚. 一种新隶属度函数的模糊支持向量机[J]. 系统仿真学报, 2009, **21**(7): 1901-1903.
DU Zhe, LIU San-yang, QI Xiao-gang. Fuzzy support vector machine with new membership function [J]. **Journal of System Simulation**, 2009, **21**(7):1901-1903. (in Chinese)

[8] 许翠云,业 宁. 基于类向心度的模糊支持向量

机[J]. 计算机工程与科学，2014，**36**(8)：1623-1628.
XU Cui-yun，YE Ning. A novel fuzzy support vector machine based on the class centripetal degree [J]. **Computer Engineering and Science**，2014，**36**(8)：1623-1628. (in Chinese)

[9] Estabrooks A，Jo T，Japkowicz N. A multiple resampling method for learning from imbalanced data sets [J]. **Computational Intelligence**，2004，**20**(1)：18-36.

[10] 杨　扬，李善平. 基于实例重要性的SVM解不平衡数据分类[J]. 模式识别与人工智能，2009，**22**(6)：913-918.
YANG Yang，LI Shan-ping. Instance importance based SVM for solving imbalanced data classification [J]. **Pattern Recognition and Artificial Intelligence**，2009，**22**(6)：913-918. (in Chinese)

[11] Veropoulos K，Campbell C，Cristianini N. Controlling the sensitivity of support vector machines [C] // **International Joint Conference on AI**. Stockholm：IJCAI Press，1999：55-60.

[12] Batuwita R，Palade V. Class imbalance learning methods for support vector machines [M] // HE Hai-bo，MA Yun-qian，eds. **Imbalanced Learning：Foundations，Algorithms，and Applications**. Hoboken：Wiley-IEEE Press，2013：83-99.

[13] 刘　艳，钟　萍，陈　静，等. 用于处理不平衡样本的改进近似支持向量机新算法[J]. 计算机应用，2014，**34**(6)：1618-1621.
LIU Yan，ZHONG Ping，CHEN Jing，*et al*. Modified proximal support vector machine algorithm for dealing with unbalanced samples [J]. **Journal of Computer Applications**，2014，**34**(6)：1618-1621. (in Chinese)

[14] Batuwita R，Palade V. FSVM-CIL：Fuzzy support vector machines for class imbalance learning [J]. **IEEE Transactions on Fuzzy Systems**，2010，**18**(3)：558-571.

[15] Chang C C，Lin C J. LIBSVM：A library for support vector machines [J]. **ACM Transactions on Intelligent Systems and Technology**，2011，**2**(3)：27.

A fuzzy support vector machine algorithm for imbalanced data classification

JU　Zhe，　CAO　Jun-zhe，　GU　Hong*

(School of Control Science and Engineering，Dalian University of Technology，Dalian 116024，China)

Abstract：As an effective machine learning technique，support vector machine (SVM) has been successfully applied to various fields. However，when it comes to imbalanced datasets，SVM produces suboptimal classification models. On the other hand，the SVM algorithm is very sensitive to noise and outliers present in the datasets. To overcome the disadvantages of imbalanced and noisy training datasets，a novel fuzzy SVM algorithm for imbalanced data classification is proposed. When designing the fuzzy membership function，the proposed algorithm takes into account not only the distance between the training sample and its class center，but also the tightness around the training sample. Experimental results show that the proposed fuzzy SVM algorithm can effectively handle the imbalanced and noisy problem.

Key words：support vector machine (SVM)；fuzzy support vector machine；fuzzy membership；imbalanced data；classification