面向不均衡分类的隶属度加权模糊支持向量机

杨志民[1],王甜甜[2],邵元海[1]
YANG Zhimin[1], WANG Tiantian[2], SHAO Yuanhai[1]

1.浙江工业大学 之江学院,杭州 310024
2.浙江工业大学 理学院,杭州 310023
1.Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, China
2.College of Science, Zhejiang University of Technology, Hangzhou 310023, China

YANG Zhimin, WANG Tiantian, SHAO Yuanhai. Weighted fuzzy support vector machine faced on fuzzy membership of imbalanced classification. Computer Engineering and Applications, 2018, 54(2):68-75.

Abstract: In view of the classification of imbalanced data set, a weighted fuzzy support vector machine is proposed, making use of the balanced adjustment factor and the fuzzy membership based on the features of samples. Firstly, it trains the classification hyperplane by traditional support vector machine and gets the fuzzy membership of every sample to be considered as the contribution rate of every sample to eliminate the error caused by noises and outliers and subtract the number of samples in a certain extent. Subsequently, it computes the balanced adjustment factor to alleviate the migration of hyperplane. Ultimately, experiments on a number of real-world data sets even including the data sets are imbalanced show that the proposed weighted fuzzy support vector machine algorithm is scalable and outperforms the existing fuzzy support vector machine as well as the typical support vector machine counterparts.

Key words: fuzzy support vector machine; weighted fuzzy support vector machine; classification hyperplane; fuzzy membership; balanced adjustment factor

摘 要:针对不均衡分类问题,提出了一种基于隶属度加权的模糊支持向量机模型。使用传统支持向量机对样本进行训练,并通过样本点与所得分类超平面之间的距离构造模糊隶属度,这不仅能够消除噪点和野值点的影响,而且可以在一定程度上约减样本;利用正负类的平均隶属度和样本数量求得平衡调节因子,消除数据不平衡时造成的分类超平面的偏移现象;通过实验结果验证了该算法的可行性和有效性。实验结果表明,该算法能有效提高分类精度,特别是对不平衡数据效果更加明显,在训练速度和分类性能上比传统支持向量机和模糊支持向量机有进一步的提升。

关键词:模糊支持向量机;加权模糊支持向量机;分类超平面;模糊隶属度;平衡调节因子
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1609-0112

1 引言

支持向量机(Support Vector Machine,SVM)[1-2]是由Vapnik等人于1995年首先提出的用于解决二分类问题的数据挖掘(Data Mining)[3]方法,它是建立在统计学理论的VC维理论[4]与结构风险最小化(Structural Risk Minimization,SRM)[2,4]原则基础上的,根据有限的样本信息在模型的复杂度和学习能力之间寻求最佳折中,以获取最好的推广能力。SVM在解决小样本、非线性及高维模式识别等问题时表现了良好的泛化能力,已被推广到多分类、回归预测及聚类等[5-7]大量实际问题当中,并取得了很好的结果。

然而,支持向量机在处理不同数据集时存在的问题也随之出现,主要有以下两种:(1)传统支持向量机对训练集中噪点和野值点特别敏感[8]。如Inoue等[9]主要针

基金项目:国家自然科学基金(No.10926198);浙江省自然科学基金(No.LY16A010020)。
作者简介:杨志民(1957—),男,教授,博士生导师,主要研究方向为数据挖掘与支持向量机、不确定性信息数学处理等;王甜甜(1991—),女,硕士研究生,主要研究方向为数据挖掘与支持向量机,E-mail:17816873919@163.com;邵元海(1983—),男,硕士生导师,主要研究方向为数据挖掘与支持向量机、最优化等。
收稿日期:2016-09-07 **修回日期:**2017-02-24 **文章编号:**1002-8331(2018)02-0068-08
CNKI **网络优先出版:**2017-02-28, http://kns.cnki.net/kcms/detail/11.2127.TP.20170228.1845.026.html

对在多分类中“一对一”和“一对多”支持向量机存在分类盲区而提出一种模糊支持向量机模型；Lin Chunfu等[10]对支持向量机中二次规划的惩罚参数添加模糊隶属度，构造模糊支持向量机模型，提高了支持向量分类机的扩展性；杨志民等[11]引入模糊事件理论，将训练样本点的输出转化为三角模糊数，并以可能性测度为基础，把模糊分类问题转化为模糊机会约束规划问题，较为充分地在数学本质上建立了模糊支持向量机模型等。(2)传统支持向量机处理不平衡数据集时由于分类超平面的偏移而致使分类性能大大下降[12-13]。为此，研究者提出一系列针对不平衡数据分类问题的方法，主要体现在训练数据集准备和算法改造方面。从不平衡数据集入手，通过对不平衡数据集进行过采样(如SMOTE[14]，SNOCC[15])或欠采样(如EUSBoost[16])等重采样技术[17]先对数据集进行预处理，减小数据集的不平衡比例，这些方法是独立于现有的分类器的；从算法改进方面入手，Veropolos K等[18]提出为正负类松弛变量给出不同惩罚因子的方法(Different Error Costs，DEC)，将正类或负类样本点数目或者数据集的不平衡比直接作为平衡因子来调整数据集的不平衡，该方法忽略了样本点分布对分类超平面的影响，在实际处理过程中可能会剔除两类的边界点，从而失去一些有用的信息；Benjamin等[19]将Boosting方法与不同惩罚系数方法相结合处理不平衡问题，当存在高不平衡类重叠分布时，该方法在一定程度上减少了多数类的分类误差，但并不彻底；袁兴梅等[20]提出一种基于代价敏感的结构化支持向量机集成分类器模型，通过训练样本的聚类得到初始权值，再运用AdaBoost策略对各样本权重进行动态调整，该方法分类结果易受聚类算法的影响等。

本文从解决上述两个问题的方向出发，综合考虑正负类样本数目、样本分布情况及与分类超平面的距离等因素，以杨志民等[11]提出的模糊支持向量机(Fuzzy Support Vector Machine，FSVM)为基础，提出了一种不均面向不均衡分类的隶属度加权模糊支持向量机(Weighted Fuzzy Support Vector Machine Faced on Fuzzy Membership of Imbalanced Classification，IFM-WFSVM)模型。首先，利用各样本点到以传统支持向量机训练所得的分类超平面的距离构造隶属度，为样本点赋予不同的权值以消除噪点和野值点对分类超平面的影响，并在一定程度上起到了简约样本的作用；然后，利用正负类的平均隶属度和样本数量求得平衡调节因子，赋予正负类样本以不同的惩罚参数，修正了因数据不平衡造成的分类误差；在本文的最后，通过人工数据集和UCI数据集两部分实验验证了该算法在提升分类精度上的有效性。

2　模糊支持向量机

模糊支持向量机可增强支持向量机的抗噪性。其主要思想是，引入三角模糊数构造出新的训练集——模糊训练集，以可能性测度为基础将分类问题模糊化，并对每个训练样本点赋予不同隶属度，这样，通过对噪点或野值点赋予很小的权值，以消除或减少噪点或野值点的影响。

给定输入空间的训练样本集 $S=\{(x_i,y_i)\}_i^n(x_i\in R^d)$，其中，$y_i\in\{-1,1\}$，$i=1,2,\cdots,n$。本文使用一种特定的三角模糊数[21]作为正负类标签“1”和“－1”的扩充。在此定义样本点为正类的隶属度为 $\delta_+(0.5<\delta_+\leqslant 1)$，样本点为负类的隶属度为 $\delta_-(0.5<\delta_-\leqslant 1)$，所对应的输出为三角模糊数：

$$\tilde{y}=\begin{cases}\left(\dfrac{2(\delta_+)^2+\delta_+-2}{\delta_+},2\delta_+-1,\dfrac{2(\delta_+)^2-3\delta_++2}{\delta_+}\right),\\ \quad 0.5<\delta_+\leqslant 1\\ \left(\dfrac{2(\delta_-)^2-3\delta_-+2}{-\delta_-},1-2\delta_-,\dfrac{2(\delta_-)^2+\delta_--2}{-\delta_-}\right),\\ \quad 0.5<\delta_-\leqslant 1\end{cases}\tag{1}$$

并为了下面说明方便，做一定的排序处理，从而得到如下形式的模糊训练集：

$$S=\{(x_1,\tilde{y}_1),\cdots,(x_p,\tilde{y}_p),(x_{p+1},\tilde{y}_{p+1}),\cdots,(x_n,\tilde{y}_n)\}\tag{2}$$

其中，$(x_i,\tilde{y}_i)(i=1,2,\cdots,p)$ 为模糊正类点，$(x_j,\tilde{y}_j)(j=p+1,p+2,\cdots,n)$为模糊负类点(注：为简单起见，在此忽略 $\delta_+=0.5$ 和 $\delta_-=0.5$ 的情况，因为此时对应的三角模糊数 $\tilde{y}=(-2,0,2)$ 不提供正负类信息)。

引入适当的映射 $\varphi:x_i\to\varphi(x_i)$，将样本 x_i 映射到高维特征空间中。选取适当的核函数，使得 $K(x_i,x_j)=\varphi(x_i)\varphi(x_j)$。引入松弛变量 $\zeta_i(i=1,2,\cdots,n)$ 及惩罚因子 C，对于某一置信水平 $\lambda(0<\lambda\leqslant 1)$，在最小化经验风险的原则下，利用Tikhonov正则化框架理论，得到模糊支持向量机的问题模型：

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\zeta_i\\&\text{s.t. } Pos\{\tilde{y}_i((w\cdot\varphi(x_i))+b)+\zeta_i\geqslant 1\}\geqslant\lambda\\&\quad\zeta_i\geqslant 0,i=1,2,\cdots,n\end{aligned}\tag{3}$$

其中，$Pos\{A\}$ 为模糊事件 A 的可能性测度[22]。

模糊支持向量机在实际分类问题[23-24]当中也得到应用。该算法在支持向量机的基础上，根据训练样本在训练过程中所起的作用的不同，为每个训练样本赋予不同的隶属度，将含有重要意义的样本点正确分类，并且忽略噪点和野值点的影响，提高了分类性能。而在实际处理不平衡数据集的过程中，经常会将少数类样本点误认为是噪点或野值点[25]并出现分类超平面的偏移等问题。

3　面向不均衡分类的隶属度加权模糊支持向量机(IFM-WFSVM)

3.1　IFM-WFSVM模型

为了解决上述问题，对数据集的不均衡所造成的影

响进行相应的补偿,考虑将隶属度 δ_i(为了表示上的方便而将 δ_+ 与 δ_- 合并)作为权重应用到FSVM模型的目标函数中,并对模糊正类点和负类点分别赋予不同的惩罚参数 C_+ 和 C_-,得到IFM-WFSVM模型。其优化问题如下:

$$\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\delta_i\zeta_i+C_-\sum_{j=p+1}^{n}\delta_j\zeta_j$$

$$\text{s.t. } Pos\{\tilde{y}_i((w\cdot\varphi(x_i))+b)+\zeta_i\geqslant 1\}\geqslant\lambda$$

$$\zeta_i\geqslant 0, i=1,2,\cdots,n \tag{4}$$

在置信水平 $\lambda(0<\lambda\leqslant 1)$ 下,式(4)的清晰等价规划[26](即与其等价的普通规划)如下所示:

$$\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\delta_i\zeta_i+C_-\sum_{j=p+1}^{n}\delta_j\zeta_j$$

$$\text{s.t. } ((1-\lambda)r_{i3}+\lambda r_{i2})((w\cdot\varphi(x_i))+b)+\zeta_i\geqslant 1$$

$$((1-\lambda)r_{j1}+\lambda r_{j2})((w\cdot\varphi(x_j))+b)+\zeta_j\geqslant 1$$

$$i=1,2,\cdots,p;j=p+1,p+2,\cdots,n$$

$$\zeta_k\geqslant 0,k=1,2,\cdots,l \tag{5}$$

其中,$(1-\lambda)r_{i3}+\lambda r_{i2}(i=1,2,\cdots,p)$ 为模糊正类点输出 $\tilde{y}_i$ 的 λ-水平截集右端点;$(1-\lambda)r_{j1}+\lambda r_{j2}(j=p+1,p+2,\cdots,n)$为模糊负类点输出 $\tilde{y}_j$ 的 λ-水平截集左端点。式(5)为凸二次规划,不存在局部最优的问题。

构造Lagrange函数,求得其对偶问题:

$$\min\ \frac{1}{2}(A+2B+D)-\left(\sum_{i=1}^{p}\alpha_i+\sum_{j=p+1}^{n}\beta_j\right)$$

$$\text{s.t. }\sum_{i=1}^{p}\alpha_i((1-\lambda)r_{i3}+\lambda r_{i2})+\sum_{j=p+1}^{n}\beta_j((1-\lambda)r_{j1}+\lambda r_{j2})=0$$

$$0\leqslant\alpha_i\leqslant C_+(\delta_+)_i,i=1,2,\cdots,p$$

$$0\leqslant\beta_i\leqslant C_-(\delta_-)_j,j=p+1,p+2,\cdots,n \tag{6}$$

其中,α_i 与 β_i 为Lagrange乘子;$\delta_i(i=1,2,\cdots,n)$ 为模糊隶属度;

$$A=\sum_{i=1}^{p}\sum_{t=1}^{p}\alpha_i\alpha_t((1-\lambda)r_{i3}+\lambda r_{i2})((1-\lambda)r_{t3}+\lambda r_{t2})K(x_i,x_t)$$

$$B=\sum_{i=1}^{p}\sum_{t=p+1}^{n}\alpha_i\beta_t((1-\lambda)r_{i3}+\lambda r_{i2})((1-\lambda)r_{t1}+\lambda r_{t2})K(x_i,x_t)$$

$$D=\sum_{i=p+1}^{n}\sum_{t=p+1}^{n}\beta_i\beta_t((1-\lambda)r_{i1}+\lambda r_{i2})((1-\lambda)r_{t1}+\lambda r_{t2})K(x_i,x_t)$$

规划式(6)为凸二次规划,其最优解为:

$$(\alpha^*,\beta^*)^{\mathrm{T}}=(\alpha_1^*,\cdots,\alpha_p^*,\beta_{p+1}^*,\cdots,\beta_n^*)^{\mathrm{T}} \tag{7}$$

求得超平面的法向量为:

$$w^*=\sum_{i=1}^{p}\alpha_i^*((1-\lambda)r_{i3}+\lambda r_{i2})\varphi(x_i)+\sum_{j=p+1}^{n}\beta_j^*((1-\lambda)r_{j1}+\lambda r_{j2})\varphi(x_j) \tag{8}$$

选取某个 $0<\alpha_i^*<C_+(\delta_+)_i(i=1,2,\cdots,p)$ 或 $0<\beta_j^*<C_-(\delta_-)_j(j=p+1,p+2,\cdots,n)$ 所对应的 $(x_i,\tilde{y}_i)$,求得:

$$b^*=((1-\lambda)r_{s3}+\lambda r_{s2})-[\sum_{i=1}^{p}\alpha_i^*((1-\lambda)r_{i3}+\lambda r_{i2})K(x_i,x_i)+\sum_{j=p+1}^{n}\beta_j^*((1-\lambda)r_{j1}+\lambda r_{j2})K(x_j,x_s)]$$

或

$$b^*=((1-\lambda)r_{q1}+\lambda r_{q2})-[\sum_{i=1}^{p}\alpha_i^*((1-\lambda)r_{i3}+\lambda r_{i2})K(x_i,x_q)+\sum_{j=p+1}^{n}\beta_j^*((1-\lambda)r_{j1}+\lambda r_{j2})K(x_j,x_q)] \tag{9}$$

最终,得到决策函数为:

$$f(x)=\mathrm{sgn}((w^*\cdot\varphi(x))+b^*) \tag{10}$$

3.2 模糊隶属度

不难看出,模糊隶属函数对一个算法来说起着至关重要的作用,且目前尚未见确定的构造方法。本文中,考虑到样本的分布情况,利用样本点到超平面的距离来作为贡献率的度量,构造一种隶属函数。

在实际操作中,首先对原始数据集利用SVM进行初步训练,得到分类超平面 $((w^*\cdot\varphi(x))+b^*=0$,然后计算样本点 x_i 到该超平面的距离 d_i,并设样本点距超平面的最大距离为 $d_{\max}^+$ 和 $d_{\max}^-$,至此,定义模糊隶属度 δ_i:

$$\delta_i=\begin{cases}1-\dfrac{d_i}{d_{\max}^++\varepsilon},i=1,2,\cdots,p\\[2ex]1-\dfrac{d_j}{d_{\max}^-+\varepsilon},j=p+1,p+2,\cdots,n\end{cases} \tag{11}$$

其中,ε 为一任意小的正数。最后,将所求隶属度作用于SVM,重新构造分类超平面,并更新隶属度 δ,直到相邻两次迭代所得结果相差小于给定的误差阈值 ε_0 或者迭代次数达到设定的迭代次数阈值 k-max 时,停止迭代,用停止迭代时的结果作为所需要的模糊隶属度参与IFM-WFSVM的计算。在本文中,取 $\varepsilon_0=0.01$,并根据实际实验经验设置 k-max 为5。

从本文的定义可以看出,距超平面近的包括支持向量的样本点保留了下来作为新的训练集且被赋予较高的隶属度,而距超平面远的(包括噪点和野值点)非支持向量从训练集中剔除,这样可以减少野值点和噪点对实验结果的影响,从而保证了学习精度。

3.3 平衡调节因子

数据的不平衡问题主要体现在正负类样本点数量上和样本点在空间分布上的不平衡。当正负类样本数目接近时,错分率主要取决于样本分布上的差异;当样本数目差距较大时,决定错分率的主要因素是样本数目的差异。本文综合考虑样本数量和分布情况,提出一种

新的惩罚参数的计算方法。

正负样本数量可以直接得出；而对于样本分布情况，由于隶属度是根据各样本点与超平面的距离得来的，可以反映出样本点在超平面周围的分布情况，选定以隶属度来表示样本分布情况。要使得正类和负类训练数据的误差相对均衡，最好应该要满足下面的条件：

$$C_+^2\sum_{i=1}^{p}(\delta_+)_i^2=C_-^2\sum_{j=p+1}^{n}(\delta_-)_j^2 \tag{12}$$

对式(12)进行处理，引入正负类样本点数目 n^+、n^- 和正负类样本点隶属度的平均值 $\bar{\delta}_+$、$\bar{\delta}_-$，得：

$$C_+^2\cdot n^+\cdot\bar{\delta}_+^2=C_-^2\cdot n^-\cdot\bar{\delta}_-^2 \tag{13}$$

设惩罚参数为：

$$C_+=C\cdot\bar{\delta}_-\cdot\sqrt{\frac{n^-}{n}},\ C_-=C\cdot\bar{\delta}_+\cdot\sqrt{\frac{n^+}{n}} \tag{14}$$

至此，本文中所提出的面向不均衡分类的隶属度加权模糊支持向量机(IFM-WFSVM)总结为算法1。

算法1 面向不均衡分类的隶属度加权模糊支持向量机(IFM-WFSVM)

输入：$\{x_i,y_i\}_{i=1}^{n},\varepsilon,k\text{-}\max,\varepsilon_0,\lambda$

输出：测试样本点的标签 $\hat{y}_i=\mathrm{sgn}((w^*\cdot\varphi(x_i))+b^*)$

步骤1 对原始训练集标准化。以标准支持向量机训练分类超平面并计算得到隶属度 $\delta_i(i=1,2,\cdots,n)$。核函数参数 σ 及惩罚参数 C 由 k-折交叉验证进行选择取最优。

步骤2 根据隶属度求得三角模糊数 $\tilde{y}$ 及模糊训练集，以此可削弱甚至剔除噪点和野值点。

步骤3 根据式(14)求得平衡调节因子。

步骤4 在置信水平 λ 下用本文的IFM-WFSVM算法对新数据集进行训练，并最终构造决策函数。

4　实验与结果分析

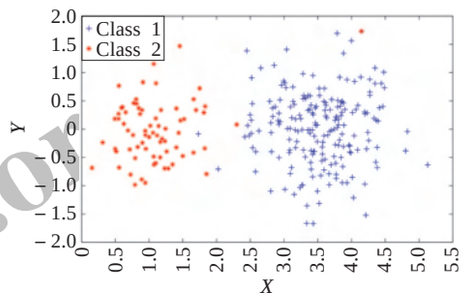
在本章，用实验结果验证了本文所提出的算法的准确率和有效性。所有的实验均是在Intel® Core™ i3 CPU 2.53 GHz 2 GB RAM PC机MATLAB R2013a软件上实现的。

4.1　人工数据集实验

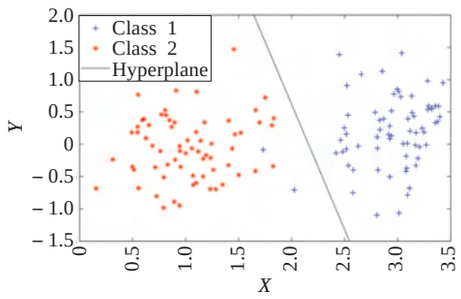
为了验证本文提出的方法对不平衡数据集的鲁棒性，下面首先用人工数据集给出两个直观的例子。

4.1.1　线性可分的数据集

在二维空间中随机生成不平衡比为200∶70的不平衡数据集的正态分布的样本点，如图1(a)所示，这些样本点共两类，类1([2.5;0],[0.3 0;0 0.4])含有200个样本点；类2([1;0],[0.2 0;0 0.3])含有70个样本点。采用本文中提出的基于样本点到超平面的距离的不平衡数据预处理方法对数据集进行预处理，图1(b)显示了经过IFM-WFSVM算法预处理后数据集的分布情况。将图1(a)与(b)进行比较，可以清晰地看出，经过预处理之后类1数据规模明显减小，噪点和野值点也被剔除，样本集的整体不平衡比例也有所减小，说明本文中提出的方法对不平衡数据集具有一定的鲁棒性。



(a)随机生成的线性样本集

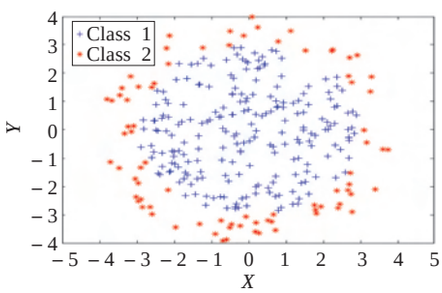


(b)IFM-WFSVM处理后

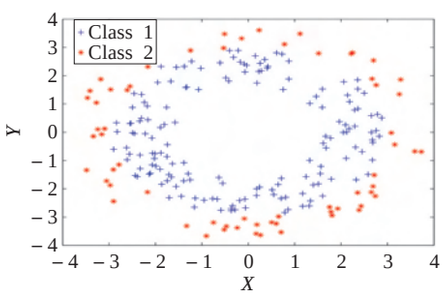
图1　IFM-WFSVM处理线性数据集

4.1.2　非线性可分的数据集

继续在非线性可分的数据集上验证IFM-WFSVM算法，随机生成如图2(a)所示的两类非线性可分的数据集，一类是以(0,0)为中心的方形数据集，共220个样本点；另一类是以(0,0)为圆心、以3为半径的圆环数据集，共80个样本点。经预处理后的数据集分布情况如



(a)随机生成的非线性样本集



(b)IFM-WFSVM处理后

图2　IFM-WFSVM处理非线性数据集

图2(b)所示。分析结果,依然可以得出本文中提出的IFM-WFSVM算法确实可提高分类性能。

4.2 UCI数据集实验

为了充分验证本文提出的IFM-WFSVM方法的有效性,本节采用7个取自UCI机器学习知识库(http://www.ics.uci.edu)的实际不平衡数据集进行实验,分别对算法中参数对准确率的影响、与不同的5种算法分类性能的对比以及训练的时间效率的对比等三个方面对本文中提出的IFM-WFSVM 算法的有效性进行了说明。表1给出了7个数据集的详细信息。

表1　实验所用的7个数据集的详细情况

Dataset	Total	Dim	Pos	Neg	Ratio
Balances	625	4	49	576	11.755
Ecoli	336	7	35	301	8.600
Spect	267	44	55	212	3.855
Hepatitis	155	19	32	123	3.844
WPBC	198	33	47	151	3.213
Haberman	306	3	81	225	2.778
Sonar	208	60	97	111	1.144

本文中,采用评价不平衡数据集时通常采用的*SE*、*SP*和*Gm*来评价分类性能,其定义如下所示:

$$SE=\frac{TP}{TP+FN},SP=\frac{TN}{TN+FP},Gm=\sqrt{SE\cdot SP} \quad (15)$$

其中,*TP*、*TN*、*FP*、*FN*分别为真正、真负、假正、假负的样本点数目。本文中,分别列出了核函数为线性和非线性的实验结果,其中,非线性的核函数采用应用广泛的Gauss径向基核函数:

$$K(x_i,x_j)=\exp\left(\frac{-\|x_i-x_j\|^2}{p^2}\right) \quad (16)$$

除了本文中提出的算法,还选取了如下5种已有的方法来做实验结果的比较。所有算法中的惩罚参数 C 和非线性核参数 p 的候选集为 $\{2^{-8},\cdots,2^8\}$,参数 λ 的候选集为 $\{0.1,\cdots,0.9\}$,均取最优时的数值参加算法。

(1)SVC:不另外添加任何约束的标准支持向量分类机算法。

(2)FSVM:引入模糊事件理论,将训练样本点的输出转化为三角模糊数,并以可能性测度为基础,把模糊分类问题转化为模糊机会约束规划问题,数学意义上较为完全的模糊支持向量机模型。

(3)DEC:为正负类松弛变量给出不同惩罚因子的方法。

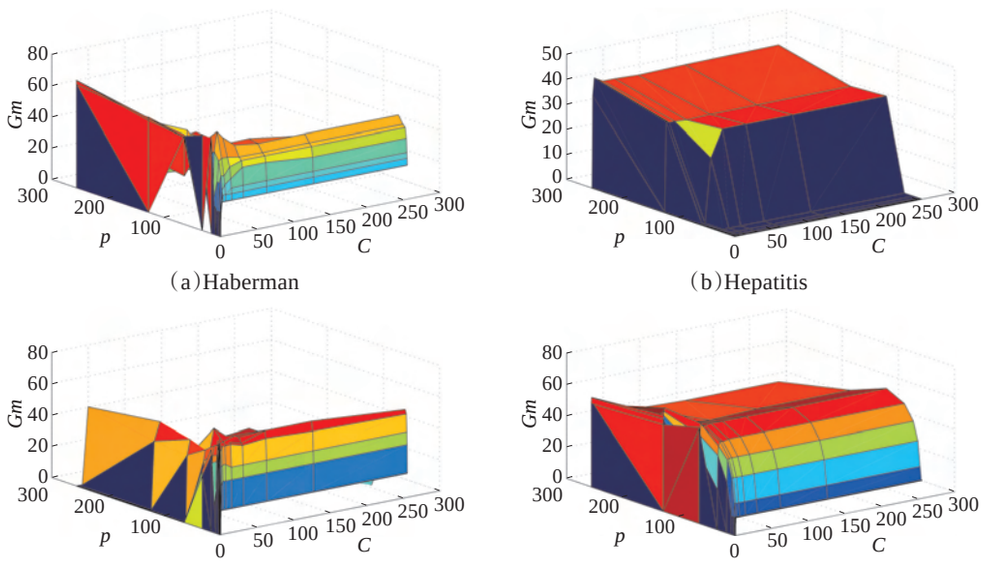
(4)LS-SVM[27]:支持向量机的一种改进,算法中将经验风险由偏差的一次改为二次方,用等式约束代替标准SVC中的不等式约束,将求解二次规划问题转化为直接求解线性方程。

(5)FSVM-CIL[28]:将FSVM与CIL(Class Imbalanced Learning)结合的一种算法,在本文的实验中,CIL采用DEC。

4.2.1 参数对算法准确率 Gm 的影响

在本小节中,讨论并给出了实验参数对IFM-WFSVM算法准确率 Gm 的影响。首先,给出的是惩罚参数 C 和RBF核参数 p 对准确率 Gm 的影响。其中,C、p 的值均从 2^{-8} 取到 2^8。相应的两参数对于 Gm 的影响效果如图3所示(本文中只给出了Haberman、Hepatitis、Balances、WPBC等四个数据集的效果图)。观察图3可以发现,准确率 Gm 较易受核参数 p 的影响,而惩罚参数 C 的影响不是很明显。

继续给出置信水平 λ 的取值变化对于 Gm 的影响,相应结果如图4。在这里,置信水平 λ 的值从0.1取到0.9。可以看出,当 λ 的值发生变化时,Gm 变化范围不是很大,但一般在取最优值下的参数值时准确率会有较

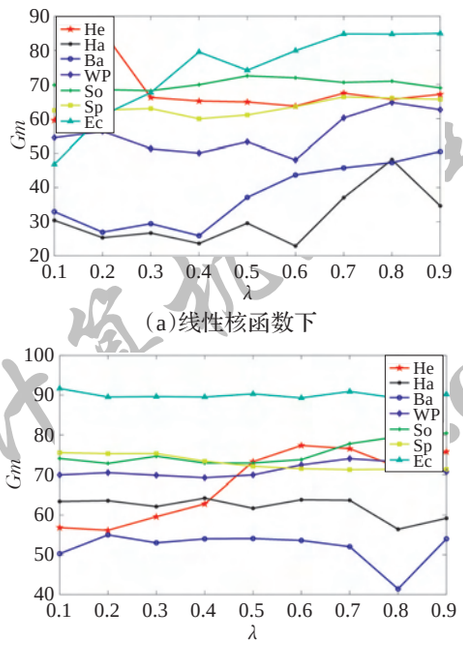


(a)Haberman　(b)Hepatitis

(c)Balances　(d)WPBC

图3　线性核函数下参数 C、p 对准确率 Gm 的影响

为明显的突出，并且对于不同的数据集，最优的参数 λ 值一般互不相同。



(a)线性核函数下

(b)非线性核函数下

图4　参数 λ 对准确率 Gm 的影响

4.2.2　算法性能的比较

在本小节中，具体给出了本文中提出的IFM-WFSVM算法与SVC、FSVM、DEC、LS-SVM、FSVM-CIL等6种算法的分类结果。其中，表2是线性核函数的结果，表3是非线性核函数的结果，包括各算法对于各数据集的 SE、SP、Gm 的均值和标准差以及 P -值。 P -值是在显著性水平5%下由T检验所得，零假设为本文中提出的IFM-WFSVM算法的准确率与其他5种算法的相比没有显著不同。从表2中可以看出，在使用线性核函数的情况下，除了对于WPBC数据集其准确率略输于LS-SVM和DEC的准确率、对于Sonar数据集略输于LS-SVM的 Gm 值之外，总体来说，IFM-WFSVM算法处理各数据集都较为有优势，在高度不平衡数据集的处理上性能也有所提升，同时，通过分析标准差可看出，IFM-WFSVM算法在稳定性方面也有一定的优势，且在Spect、Hepatitis等数据集优势明显，综合对于准确率 Gm 值的分析对比结果，可以认为该算法能有效改进不平衡数据的分类性能。同样，从表3的非线性核函数的分类结果中亦可以得出相似的结论。

对实验结果进行综合分析，可以看出，本文中所提出的IFM-WFSVM算法在分类性能上的确有较好的提

表2　采用线性核函数的各算法实验结果对比

Dataset	Index	SVC	FSVM	DEC	LS-SVM	FSVM-CIL	IFM-WFSVM
Balances	*SE*/%	0.000 0±0.00	35.156 3±5.12	0.000 0±0.00	92.156 8±0.02	37.153 1±2.38	45.223 0±2.22
	SP/%	100.000 0±0.00	61.204 9±0.46	100.000 0±0.00	0.000 0±0.00	65.187 5±0.67	54.540 4±1.10
	Gm/%	0.000 0±0.00	46.289 6±3.43	0.000 0±0.00	0.000 0±0.00	49.187 7±1.43	49.651 5±1.41
	P-value	0.000 0	0.135 1	0.000 0	0.000 0	0.998 2	—
Ecoli	*SE*/%	—	15.041 7±3.28	—	90.874 5±0.34	8.944 4±6.30	37.838 0±4.99
	SP/%	—	38.317 2±0.77	—	39.166 6±9.11	48.679 2±1.21	54.258 4±1.40
	Gm/%	—	23.900 4±2.47	—	40.486 2±8.80	19.549 1±9.20	45.214 5±2.73
	P-value	—	0.007 6	—	0.040 1	0.000 0	—
Spect	*SE*/%	86.627 6±3.86	43.463 4±1.39	67.814 0±6.51	25.333 3±10.32	45.809 1±5.08	50.093 2±1.73
	SP/%	38.462 6±3.74	99.464 2±0.73	56.411 9±6.93	79.932 9±0.54	93.982 5±1.78	98.850 0±1.59
	Gm/%	57.632 8±2.61	65.744 5±1.21	61.579 1±2.59	25.678 7±8.71	65.524 4±3.59	70.352 7±0.93
	P-value	0.000 0	0.000 1	0.000 1	0.000 0	0.019 7	—
Haberman	*SE*/%	0.000 0±0.00	58.011 3±2.69	13.000 0±4.47	47.666 7±11.39	58.794 0±3.11	26.740 7±3.74
	SP/%	100.000 0±0.00	38.703 1±1.09	87.714 2±4.35	75.003 0±0.57	48.560 9±5.85	87.289 6±2.92
	Gm/%	0.000 0±0.00	47.363 1±0.93	33.289 9±4.67	46.668 8±11.10	53.293 9±2.27	48.169 4±2.79
	P-value	0.000 0	0.568 6	0.000 0	0.437 2	0.004 6	—
WPBC	*SE*/%	89.540 0±3.12	52.705 7±5.20	87.020 7±1.78	57.066 6±11.93	53.979 1±2.38	86.172 4±4.26
	SP/%	41.453 1±6.14	55.666 7±2.08	44.828 5±5.62	84.764 6±1.62	53.238 0±9.41	47.695 2±11.14
	Gm/%	60.784 2±4.51	54.153 3±3.68	62.351 8±3.78	62.738 1±13.19	53.348 1±3.69	63.837 8±8.43
	P-value	0.978 9	0.283 7	0.821 5	0.892 3	0.028 0	—
Hepatitis	*SE*/%	56.583 3±9.31	72.583 3±6.36	65.380 9±3.94	88.511 4±1.00	53.250 0±11.73	91.488 1±0.92
	SP/%	89.641 5±2.40	82.612 2±3.33	84.531 9±2.21	73.190 4±4.66	82.129 8±2.98	77.420 8±4.31
	Gm/%	71.055 8±4.91	77.421 4±4.94	74.311 0±2.51	74.322 3±.40	65.901 4±8.42	84.138 2±1.92
	P-value	0.072 6	0.348 0	0.051 9	0.059 3	0.022 5	—
Sonar	*SE*/%	72.658 7±2.83	69.988 4±4.58	73.201 4±4.59	80.710 4±1.50	78.734 0±3.56	72.292 4±4.67
	SP/%	66.203 9±6.47	55.961 3±3.05	68.142 0±3.07	75.754 4±2.60	56.421 8±6.11	69.732 6±4.28
	Gm/%	69.316 7±4.39	62.578 1±3.65	70.560 2±1.73	76.976 4±2.47	66.601 5±4.76	70.968 3±3.76
	P-value	0.541 0	0.007 2	0.831 2	0.601 1	0.146 4	—

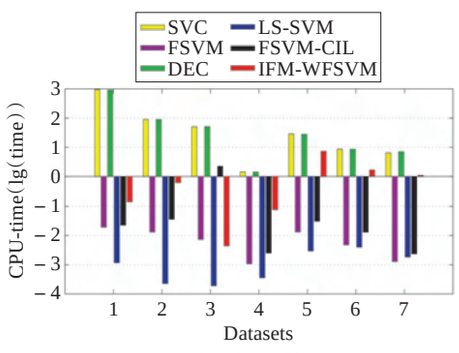
表3 采用非线性核函数的各算法实验结果对比

Dataset	Index	SVC	FSVM	DEC	LS-SVM	FSVM-CIL	IFM-WFSVM
Balances	*SE*/%	92.148 3±0.01	19.485 2±6.21	100.000 0±0.00	92.155 2±0.02	48.903 9±3.32	47.376 2±9.96
	SP/%	0.000 0±0.00	89.276 0±2.78	0.000 0±0.00	0.000 0±0.00	51.657 3±0.58	60.003 5±1.08
	Gm/%	0.000 0±0.00	41.304 0±6.18	0.000 0±0.00	0.000 0±0.00	50.234 1±1.57	53.081 9±5.54
	P-value	0.000 0	0.013 2	0.000 0	0.000 0	0.301 3	—
Ecoli	*SE*/%	95.367 2±0.31	85.349 2±2.00	5.288 9±0.48	95.130 5±0.21	76.666 7±4.18	79.750 0±3.45
	SP/%	59.761 9±5.89	63.999 7±8.23	98.631 8±0.03	71.103 5±6.13	67.817 0±6.11	85.581 5±0.30
	Gm/%	68.942 4±6.29	73.821 3±5.11	22.821 7±1.02	77.710 7±5.50	72.070 9±4.76	82.599 1±1.73
	P-value	0.010 3	0.040 5	0.000 0	0.051 2	0.022 9	—
Spect	*SE*/%	87.507 6±1.59	62.626 8±1.12	88.204 5±1.43	55.321 8±5.56	70.408 3±0.60	67.188 7±1.11
	SP/%	46.692 8±3.91	87.126 1±3.31	42.373 0±5.59	84.449 9±0.68	79.974 6±4.55	89.122 0±1.52
	Gm/%	63.894 3±3.25	73.849 6±1.21	61.030 9±4.05	62.193 8±5.33	75.012 3±2.04	77.380 3±1.16
	P-value	0.000 0	0.001 5	0.000 0	0.000 0	0.054 3	—
Haberman	*SE*/%	97.460 9±4.65	40.833 6±4.10	7.866 7±8.33	75.051 9±0.43	77.808 4±3.81	63.054 0±4.13
	SP/%	20.803 0±0.85	43.526 7±0.50	96.904 2±4.80	44.283 3±8.32	51.694 7±5.69	64.974 9±2.24
	Gm/%	44.340 5±3.49	41.961 0±4.19	20.727 8±19.48	44.360 1±8.20	63.285 9±2.61	63.956 7±1.70
	P-value	0.000 3	0.000 0	0.001 1	0.001 5	0.643 1	—
WPBC	*SE*/%	86.996 3±1.29	93.263 3±0.98	86.286 6±1.64	66.676 1±0.20	50.083 9±2.91	63.983 2±1.71
	SP/%	41.283 3±6.70	14.457 2±3.14	44.641 2±7.70	82.156 2±1.06	58.781 7±4.96	71.728 0±6.91
	Gm/%	59.750 0±4.53	36.559 7±4.18	61.876 8±5.67	65.337 8±8.08	54.176 8±2.07	67.597 7±4.11
	P-value	0.021 0	0.000 0	0.105 6	0.387 8	0.000 0	—
Hepatitis	*SE*/%	58.958 3±7.51	86.077 4±7.24	71.293 7±3.23	88.209 5±0.70	68.700 0±5.54	90.750 0±1.61
	SP/%	88.930 4±1.45	75.318 0±0.71	82.845 4±2.19	67.704 8±6.09	89.967 3±0.33	74.167 4±1.73
	Gm/%	72.320 1±4.98	80.465 9±3.61	76.829 0±1.59	69.106 2±6.97	78.565 1±3.22	82.036 6±2.02
	P-value	0.043 1	0.436 5	0.368 1	0.0318	0.223 1	—
Sonar	*SE*/%	73.868 6±1.48	79.371 7±1.69	69.100 6±3.56	86.463 4±0.35	72.728 3±4.10	73.950 0±2.45
	SP/%	65.371 9±1.81	65.160 5±3.19	66.949 2±3.79	90.432 9±0.99	55.196 9±4.54	90.116 7±1.82
	Gm/%	69.483 1±1.24	71.888 1±1.53	67.964 0±2.11	88.065 2±1.08	63.342 1±4.13	81.631 0±2.10
	P-value	0.003 1	0.010 7	0.002 8	0.088 3	0.000 0	—

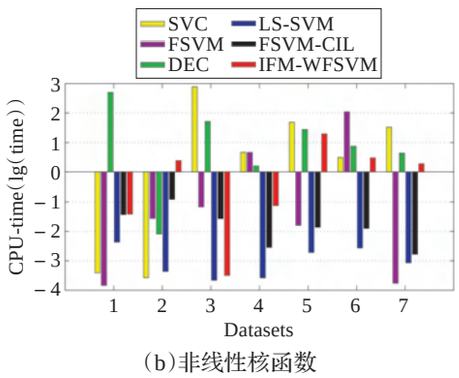
升,特别是对于数据集的非平衡比例较大的情况。这是因为IFM-WFSVM算法在处理高度不平衡的数据集时,如数据集Balances,综合考虑了影响不平衡分类性能的因素,所以才表现出了一定的优越性。*SE*值虽然有所降低,但*Gm*值提升明显。随着数据集的不平衡度逐渐降低,该算法的性能仍比其他算法要好。综合上述分析,本文中所提出的IFM-WFSVM算法在综合考虑了正负类样本数目及分布情况对分类器的影响的条件下,表现出了更强的适应性及稳定性,即使在处理不平衡比较小的数据集时,依然能保持一定的性能优势。

4.2.3 算法效率的比较

在本小节中,列出了本文中提出的IFM-WFSVM算法与SVC、FSVM、DEC、LS-SVM、FSVM-CIL等算法在7个数据集中的训练时间(图中时间数据取lg(time))对比。其中,图5(a)是线性核函数的结果,图5(b)是非线性核函数的结果。从图5(a)和(b)中可以看出,提出的IFM-WFSVM算法在计算时间上的效率属于较为中等的水平。这是因为,虽然IFM-WFSVM以隶属度作为权重进行加权时,剔除了一部分对分类性能影响不大的样本点,可以相对提升训练速度,但是在用标准SVC计算初始超平面的时候花费时间较多。



(a)线性核函数



(b)非线性核函数

图5 6种算法在7个数据集下的训练时间

5 结论

本文针对不均衡分类问题,综合考虑正负类样本数目、样本分布情况及其与分类超平面的距离等因素,提出了一种基于面向不均衡分类的隶属度加权模糊支持向量机模型IFM-FSVM。该算法利用各样本点对分类超平面的贡献率(即各样本点到分类超平面的距离)构造隶属度δ,为样本点赋予不同的权值,提高了算法的抗噪性能,并在一定程度上起到了简约样本的作用。当正类与负类的样本数相差较大时,引入平衡调节因子C_+和C_-,可以根据正负类样本点数和样本分布等两个因素进行调节,缓解了数据不平衡时造成的分类超平面的偏移现象。实验的结果,证明了该算法的有效性和实用性,在分类性能上相较于SVC、FSVM、DEC、LS-SVM、FSVM-CIL等算法有较为显著的提升。在未来的研究中,将探索如何进一步提升加权模糊支持向量机的分类性能及训练速度等问题。

参考文献:

[1] Cortes C, Vapnik V.Support vector networks[J].Machine Learning, 1995, 20(3): 273-297.

[2] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2004.

[3] Witten I H, Frank E, Hall M A.Data mining: practical machine learning tools and techniques[M].3rd ed.[S.l.]: Morgan Kaufmann Publishers Inc, 2011: 206-207.

[4] Vapnik V.Statistical learning theory[M].New York: Wiley, 1998.

[5] Morales N, Toledo J, Acosta L.Path planning using a multiclass support vector machine[J].Applied Soft Computing, 2016, 43: 498-509.

[6] Santamaria-Bonfil G, Reyes-Ballesteros A, Gershenson C. Wind speed forecasting for wind farms: a method based on support vector regression[J].Renewable Energy, 2015, 85: 790-809.

[7] Ping Y, Chang Y F, Zhou Y.Fast and scalable support vector clustering for large-scale data analysis[J].Knowledge & Information Systems, 2014, 43(2): 281-310.

[8] Zhang Xuegong.Using class-center vectors to build support vector machines[C]//IEEE Conference on Neural Networks for Signal Processing, 1999: 3-11.

[9] Inoue T, Abe S.Fuzzy support vector machine for pattern classification[C]//International Joint Conference on Neural Networks, 2001: 1449-1454.

[10] Lin Chunfu, Wang Shengde.Fuzzy support vector machines[J].IEEE Trans on Neural Networks, 2002, 13(2): 464-471.

[11] 杨志民,邓乃扬.基于可能性理论的模糊支持向量分类机[J].模式识别与人工智能,2007,20(1):7-14.

[12] Japkowicz N, Stephen S.The class imbalabced problem: a systematic study[J].Intelligent Data Analysis, 2002, 6(5): 429-449.

[13] Peng X, Wang Y.A Bi-Fuzzy Progressive Transductive Support Vector Machine (BFPTSVM) algorithm[J].Expert Systems with Applications, 2010, 37(1): 527-533.

[14] Chawla N V, Bowyer K W, Hall L O, et al.SMOTE: synthetic minority over-sampling technique[J].Journal of Artificial Intelligence Research, 2011, 16(1): 321-357.

[15] Zheng Z, Cai Y, Li Y.Oversampling method for imbalanced classification[J].Computing & Informatics, 2015, 34(5): 1017-1037.

[16] Galar M, Fernández A, Barrenechea E, et al.EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling[J].Pattern Recognition, 2013, 46(12): 3460-3471.

[17] Yang W Y, Liu S X, Jin T S, et al.An optimization criterion for generalized marginal fisher analysis on undersampled problems[J].International Journal of Automation and Computing, 2011, 8(2): 193-200.

[18] Veropolos K, Campbell C, Cristianini N.Controlling the sensitivity of support vector machines[C]//Proceedings of Artificial Intelligence, 1999: 55-60.

[19] Benjamin X W, Nathalie J.Boosting SVM for imbalanced datasets[J].Journal Foundations of Intelligent Systems, 2008, 4994: 38-47.

[20] 袁兴梅,杨明.一种面向不平衡数据的结构化SVM集成分类器[J].模式识别与人工智能,2013,26(3):315-320.

[21] Zadeh L A.Fuzzy sets as a basis for a theory of possiblility[J].Fuzzy Sets and Systems, 1978, 1(1).

[22] 刘宝碇,赵瑞清.随机规划与模糊规划[M].北京:清华大学出版社,1998.

[23] Gu X, Ni T, Wang H.New fuzzy support vector machine for the class imbalance problem in medical datasets classification[J].Scientific World Journal, 2013.

[24] Jaya T, Dheeba J.Detection of hard exudates in colour fundus images using fuzzy support vector machine-based expert system[J].Journal of Digital Imaging, 2015, 28: 761-768.

[25] 汪廷华,田盛丰.特征加权支持向量机[J].电子与信息学报,2009,31(3):514-518.

[26] 杨志民,刘广利.不确定性支持向量机——算法及应用[M].北京:科学出版社,2012.

[27] Suykens J A K.Least squares support vector machines[M].[S.l.]: World Scientific Pub Co Inc, 2003.

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