基于混合模糊隶属度的模糊双支持向量机研究*

丁胜锋[1,2]，孙劲光[1]

(1. 辽宁工程技术大学，辽宁 葫芦岛 125105；2. 辽宁石油化工大学 经济管理学院，辽宁 抚顺 113001)

摘 要：双支持向量机是近年提出的一种新的支持向量机。在处理模式分类问题时，双支持向量机速度远远超过传统支持向量机，而且显示出较好的推广能力。但双支持向量机没有考虑不同输入样本点可能会对分类超平面的形成产生不同影响，在某些实际问题中具有局限性。为了克服这个缺点，提出了一种基于混合模糊隶属度的模糊双支持向量机。该算法设计了一种结合距离和紧密度的模糊隶属度函数，给不同的训练样本赋予不同的模糊隶属度，构建两个最优非平行超平面，最终实现二值分类。实验证明，该模糊双支持向量机的分类性能优于传统的双支持向量机。

关键词：模糊隶属度；支持向量机；双支持向量机；模式分类

中图分类号：TP18 **文献标志码**：A **文章编号**：1001-3695(2013)02-0432-04

doi:10.3969/j.issn.1001-3695.2013.02.031

Research on fuzzy twin support vector machine based on hybrid fuzzy membership

DING Sheng-feng[1,2], SUN Jin-guang[1]

(1. *Liaoning Technical University*, *Huludao Liaoning* 125105, *China*; 2. *School of Economics & Management*, *Liaoning Shihua University*, *Fushun Liaoning* 113001, *China*)

Abstract: As a new version of support vector machine(SVM), twin support vector machine(TWSVM) is proposed recently. TWSVM is not only more faster than a conventional SVM, but shows good generalization for pattern classification. But the different effects of the different training samples on the classification hyperplanes are ignored in TWSVM, and the limitation is existed for some actual applications. Therefore, this paper presented a fuzzy twin support vector machine based on hybrid fuzzy membership. It designed a fuzzy membership function combined distance with affinity, and modified TWSVM by applying the fuzzy membership to every training sample. Finally it built two optimal nonparallel hyperplanes to achieve classification. The experiments indicate that the classification performance of the algorithm is more superiorer than a traditional TWSVM.

Key words: fuzzy membership; support vector machine; twin support vector machine; pattern classification

支持向量机(SVM)作为一种新的有效的统计学习方法，具有小样本学习、非线性、高维数与推广性好的优点，近年来成为模式识别与机器学习领域一个新的研究热点[1,2]。在运用SVM对输入样本分类时，要解决二次规划问题就非常费时。为了避免SVM训练时间过长的问题，最近出现了几种改进算法，其中有邻近支持向量机(proximal support vector machine, PSVM)[3~6]、基于广义特征值的多平面邻近支持向量机(proximal SVM based on generalized eigenvalues, GEPSVM)[7,8]和双支持向量机(twin SVM, TWSVM)。PSVM以求解方程组方式来代替求解凸规划问题，在处理模式分类问题中速度远远超过SVM；GEPSVM摈弃了PSVM平行约束的条件，通过求解两个广义特征值问题来求解两个非平行超平面[7,8]，其计算效率优于SVM，同时可以得到较好的分类性能；TWSVM是通过解决两个形如SVM的凸规划问题来构建两个非平行超平面，能将训练时间约减到原SVM的1/4[9,10]，但它没有考虑不同输入样本点对分类超平面的形成产生不同影响，所以在某些应用中具有一定的局限性。

为了解决此问题，本文对每个样本都赋予一个属于其类别的模糊隶属度值，使不同的样本对判别函数的学习有不同的贡献，以此来改进TWSVM。实验证明，这种改进的模糊TWSVM比TWSVM分类性能更好。

1 双支持向量机

给定 m 个线性可分样本点$\{(x_i, y_i), i = 1, \cdots, m\}$，其中 $x_i \in \mathrm{R}^n$，$y_i \in \{-1, 1\}$表示 x_i 的类别。SVM的思想是：通过一对带有最大间隔的平行超平面，将两类样本尽可能地分开(图1)。其数学模型可描述为

$$\min_{w,b,\xi} \quad \frac{1}{2}w^{\mathrm{T}}w + C(\sum_{i=1}^{m}\xi_i) \tag{1}$$

$$\text{s.t.} \quad y_i[(w.x_i) + b] + \xi_i \geqslant 1 \tag{2}$$

$$\xi_i \geqslant 0 \quad i = 1, \cdots, m \tag{3}$$

TWSVM摈弃了平行约束的条件，构建两个非平行超平面，使得每一类样本离一个超平面尽可能近，而离另一个超平面尽可能远(图2)。假设属于1类和-1类的样本点分别由矩阵 A 和 B 来表示，那么TWSVM分类器可以通过以下一对二次规划问题得到

收稿日期：2012-06-05；**修回日期**：2012-07-20 **基金项目**：辽宁省重点实验室资助项目(2008s115)

作者简介：丁胜锋(1981-)，男，湖北蕲春人，博士研究生，主要研究方向为数据挖掘(jgdsf@163.com)；孙劲光(1962-)，女，辽宁阜新人，教授，博导，主要研究方向为图形图像、数据挖掘.

(TWSVM1)

$$\min_{w_1,b_1,\xi} \frac{1}{2}(Aw_1+e_1b_1)^{\mathrm{T}}(Aw_1+e_1b_1)+c_1e_2^{\mathrm{T}}\xi \tag{4}$$

$$\text{s.t.}\ -(Bw_1+e_2b_1)^{\mathrm{T}}+\xi\geqslant e_2\quad \xi\geqslant 0 \tag{5}$$

(TWSVM2)

$$\min_{w_2,b_2,\xi} \frac{1}{2}(Bw_2+e_2b_2)^{\mathrm{T}}(Bw_2+e_2b_2)+c_2e_1^{\mathrm{T}}\xi \tag{6}$$

$$\text{s.t.}\ (Aw_2+e_1b_2)^{\mathrm{T}}+\xi\geqslant e_1\quad \xi\geqslant 0 \tag{7}$$

其中:c_1 和 c_2 是惩罚参数;e_1 和 e_2 是全为 1 组成的列向量。TWSVM 为每一个类都找到一个超平面,式(4)和(6)的目标函数用平方距离来度量本类样本离本类超平面的距离,因此最小化它可以保证本类样本离本类超平面尽可能近。目标函数的第 1 项是属于 1 类的所有样本点到与之对应的分类面距离的平方和,第 2 项是误差变量的和。不等式约束可以理解为它类样本离超平面至少为 1。通过式(4)和(6)的第 2 项可以看出,每个样本的误差系数都相同,每个训练样本对分类面产生的影响是相同的。通过解优化问题式(4)和(6)可以得到 TWSVM 的两个分类面为

$$x^{\mathrm{T}}w_1+b_1=0 \tag{8}$$

$$x^{\mathrm{T}}w_2+b_2=0 \tag{9}$$

TWSVM 可以看成是 SVM 的分解算法。除了约束条件不包括所有的数据以外,TWSVM 的每一个二次规划问题都类似于 SVM。TWSVM 是解决一对二次规划问题,而 SVM 是解决一个二次规划问题。如果两类样本数目相等,则 TWSVM 的计算时间为 $2\times(m/2)^3=m^3/4$,而 SVM 的计算时间为 m^3,则 TWSVM 可以得到较 SVM 快 4 倍的速度。与 SVM 相比,TWSVM 大大降低了算法的时间复杂度。

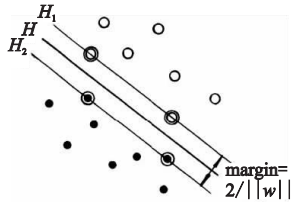

图1　SVM分类示意图

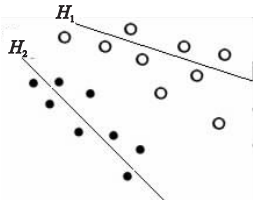

图2　TWSVM分类示意图

2　模糊双支持向量机

TWSVM 规定每个训练样本点属于两个类别中的一个。在每一个类中,每个训练样本都会对分类面产生相同影响,而在某些实际应用中,不同的训练样本对分类面的作用可能会不同。例如,那些正常的样本点应该比野值和噪声数据更加重要;分类算法应该更加注重那些重要训练样本的分类情况,而 TWSVM 没有考虑这个问题。因此,本文对不同样本采用不同的权重系数,使得在构造目标函数时,对野值或噪声赋予较小的权值,以削弱其影响,并由此产生新的优化问题。

2.1　线性可分模糊双支持向量机

给定训练样本集合$\{(x_i,y_i),i=1,\cdots,m\}$,引入模糊因子 $s_i(0<s_i\leqslant 1,i=1,\cdots,m)$,表示不同样本点属于相应类别的程度,这样就得到模糊化后的样本集$\{(x_i,y_i,s_i),i=1,\cdots,m\}$。则求解分类超平面的优化问题为

(FTWSVM1)

$$\min_{w_1,b_1,\xi} \frac{1}{2}(Aw_1+e_1b_1)^{\mathrm{T}}(Aw_1+e_1b_1)+c_1S_Ae_2^{\mathrm{T}}\xi \tag{10}$$

$$\text{s.t.}\ -(Bw_1+e_2b_1)^{\mathrm{T}}+\xi\geqslant e_2\quad \xi\geqslant 0 \tag{11}$$

(FTWSVM2)

$$\min_{w_2,b_2,\xi} \frac{1}{2}(Bw_2+e_2b_2)^{\mathrm{T}}(Bw_2+e_2b_2)+c_2S_Be_1^{\mathrm{T}}\xi \tag{12}$$

$$\text{s.t.}\ (Aw_2+e_1b_2)^{\mathrm{T}}+\xi\geqslant e_1\quad \xi\geqslant 0 \tag{13}$$

其中:S_A、S_B 分别表示样本集 A 和 B 中每个样本的模糊隶属度。每个样本的误差 ξ_i 乘以相应的隶属度 s_i,表示这种误差对于分类问题的影响,s_i 越小,相应的输入样本 x_i 在模糊双支持向量机中的作用就越低,这样就减少了野值和噪声的影响,提高了分类性能。

FTWSVM1 的 Lagrangian 函数为

$$L(w_1,b_1,\xi,\alpha,\beta)=\frac{1}{2}(Aw_1+e_1b_1)^{\mathrm{T}}(Aw_1+e_1b_1)+$$
$$c_1S_Ae_2^{\mathrm{T}}\xi-\alpha^{\mathrm{T}}(-(Bw_1+e_2b_1)^{\mathrm{T}}+\xi-e_2)-\beta^{\mathrm{T}}\xi \tag{14}$$

根据 KKT 条件,得到以下线性方程组:

$$A^{\mathrm{T}}(Aw_1+e_1b_1)+B^{\mathrm{T}}\alpha=0 \tag{15}$$

$$e_1^{\mathrm{T}}(Aw_1+e_1b_1)+e_2^{\mathrm{T}}\alpha=0 \tag{16}$$

$$c_1S_Ae_2-\alpha-\beta=0 \tag{17}$$

$$\alpha^{\mathrm{T}}(-(Bw_1+e_2b_1)^{\mathrm{T}}+\xi-e_2)=0 \tag{18}$$

$$\beta^{\mathrm{T}}\xi=0 \tag{19}$$

$$\alpha\geqslant 0,\beta\geqslant 0 \tag{20}$$

因为 $\beta\geqslant 0$,可得出 $0\leqslant\alpha\leqslant c_1S_A$。由式(15)(16)可得

$$[A^{\mathrm{T}}\quad e_1^{\mathrm{T}}][A\quad e_1][w_1,b_1]^{\mathrm{T}}+[B^{\mathrm{T}}\quad e_2^{\mathrm{T}}]\alpha=0 \tag{21}$$

令 $H=[A\quad e_1]$,$G=[B\quad e_2]$,$u=[w_1,b_1]^{\mathrm{T}}$,可得

$$H^{\mathrm{T}}Hu+G^{\mathrm{T}}\alpha=0 \tag{22}$$

$$u=-(H^{\mathrm{T}}H)^{-1}G^{\mathrm{T}}\alpha \tag{23}$$

根据式(14)~(19)可得到 FTWSVM1 的对偶问题为

(DFTWSVM1)

$$\max_{\alpha} e_2^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}G(H^{\mathrm{T}}H)^{-1}G^{\mathrm{T}}\alpha \tag{24}$$

$$\text{s.t.}\quad 0\leqslant\alpha\leqslant c_1S_A \tag{25}$$

同理,令 $P=[A\quad e_1]$,$Q=[B\quad e_2]$,$v=[w_2,b_2]^{\mathrm{T}}$,可得

$$v=(Q^{\mathrm{T}}Q)^{-1}P^{\mathrm{T}}\gamma \tag{26}$$

则 FTWSVM2 的对偶问题为

(DFTWSVM2)

$$\max_{\gamma} e_1^{\mathrm{T}}\gamma-\frac{1}{2}\gamma^{\mathrm{T}}P(Q^{\mathrm{T}}Q)^{-1}P^{\mathrm{T}}\gamma \tag{27}$$

$$\text{s.t.}\quad 0\leqslant\gamma\leqslant c_2S_B \tag{28}$$

显然,模糊双支持向量机与标准双支持向量机的区别是在对偶问题中变量 α 和 γ 的上界约束是随模糊因子 s_i 变化的,这也相当于对每个样本使用一个惩罚因子 s_ic。当样本的数量远大于样本的维数时,即 $m>>n$,矩阵 $H^{\mathrm{T}}H$ 和 $Q^{\mathrm{T}}Q$ 的维数为 $(n+1)\times(n+1)$,大大减少了计算复杂度。一旦求出了 u 和 v,两个非平行超平面就可以确定。这使得分类面对重要数据的分类精度显著提高,并且有较强的抗噪声能力。

对于待分样本 x,根据式(29)确定它的类别。

$$x^{\mathrm{T}}w_r+b_r=\min_{l=1,2}|x^{\mathrm{T}}w_l+b_l| \tag{29}$$

其中:$|\cdot|$为 x 到平面 $x^{\mathrm{T}}w_l+b_l=0(l=1,2)$ 的垂直距离。

2.2　非线性可分模糊双支持向量机

对于非线性情况,通过 $w_1=C^{\mathrm{T}}u_1$ 和 $w_2=C^{\mathrm{T}}u_2$,分别代替 w_1 和 w_2,同时引入核矩阵 $K(x^{\mathrm{T}},C^{\mathrm{T}})=\Phi(x^{\mathrm{T}}).\Phi(C^{\mathrm{T}})$,两个分类平面分别为

$$K(x^{\mathrm{T}},C^{\mathrm{T}})u_1+b_1=0 \tag{30}$$

$$K(x^{\mathrm{T}},C^{\mathrm{T}})u_2+b_2=0 \tag{31}$$

其中:$C^{\mathrm{T}}=[A \quad B]^{\mathrm{T}}$。则求解分类平面的优化问题为

(KFTWSVM1)

$$\min_{u_1,b_1,\xi}\frac{1}{2}\|(K(A,C^{\mathrm{T}})u_1+e_1b_1\|^2+c_1S_Ae_2^{\mathrm{T}}\xi \tag{32}$$

$$-(K(B,C^{\mathrm{T}})u_1+e_2b_1)^{\mathrm{T}}+\xi\geqslant e_2 \quad \xi\geqslant 0 \tag{33}$$

(KFTWSVM2)

$$\min_{u_2,b_2,\xi}\frac{1}{2}\|(K(B,C^{\mathrm{T}})u_2+e_2b_2\|^2+c_2S_Be_1^{\mathrm{T}}\xi \tag{34}$$

$$(K(A,C^{\mathrm{T}})u_1+e_1b_2)^{\mathrm{T}}+\xi\geqslant e_1 \quad \xi\geqslant 0 \tag{35}$$

KFTWSVM1 的 Lagrangian 函数为

$$L(u_1,b_1,\xi,\alpha,\beta)=\frac{1}{2}\|K(A,C^{\mathrm{T}})u_1+e_1b_1\|^2+c_1S_Ae_2^{\mathrm{T}}\xi-$$

$$\alpha^{\mathrm{T}}(-(K(B,C^{\mathrm{T}})u_1+e_2b_1)+\xi-e_2)-\beta^{\mathrm{T}}\xi \tag{36}$$

根据 KKT 条件得出

$$K(A,C^{\mathrm{T}})^{\mathrm{T}}(K(A,C^{\mathrm{T}})u_1+e_1b_1)+K(B,C^{\mathrm{T}})^{\mathrm{T}}\alpha=0 \tag{37}$$

$$e_1^{\mathrm{T}}(K(A,C^{\mathrm{T}})u_1+e_1b_1)+e_2^{\mathrm{T}}\alpha=0 \tag{38}$$

$$c_1S_Ae_2-\alpha-\beta=0 \tag{39}$$

$$\alpha^{\mathrm{T}}(-(K(B,C^{\mathrm{T}})u_1+e_2b_1)+\xi-e_2)=0 \tag{40}$$

$$\beta^{\mathrm{T}}\xi=0 \tag{41}$$

$$\alpha\geqslant 0,\beta\geqslant 0 \tag{42}$$

因为 $\beta\geqslant 0$,可得出 $0\leqslant\alpha\leqslant c_1S_A$。由式(37)和(38)可得

$$[K(A,C^{\mathrm{T}})^{\mathrm{T}} \quad e_1^{\mathrm{T}}][K(A,C^{\mathrm{T}}) \quad e_1][u_1 \quad b_1]+$$

$$[K(B,C^{\mathrm{T}})T \quad e_2^{\mathrm{T}}]\alpha=0 \tag{43}$$

令 $S=[K(A,C^{\mathrm{T}}) \quad e_1]$,$R=K(B,C^{\mathrm{T}})T \quad e_2^{\mathrm{T}}]$,$z=[u_1,b_1]^{\mathrm{T}}$,可得出

$$S^{\mathrm{T}}Sz+R^{\mathrm{T}}\alpha=0 \tag{44}$$

$$z=-(S^{\mathrm{T}}S)^{-1}R^{\mathrm{T}}\alpha \tag{45}$$

因此可得到 KFTWSVM1 的对偶问题为

(KDFTWSVM1)

$$\max_{\alpha} \quad e_2^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}R(S^{\mathrm{T}}S)^{-1}R^{\mathrm{T}}\alpha \tag{46}$$

$$\text{s.t.} \quad 0\leqslant\alpha\leqslant c_1S_A \tag{47}$$

同理,令 $L=[K(A,C^{\mathrm{T}}) \quad e_1]$,$N=[K(B,C^{\mathrm{T}}) \quad e_2]$,$z_2=[u_2,b_2]^{\mathrm{T}}$,可以得到

$$z_2=(N^{\mathrm{T}}N)^{-1}L^{\mathrm{T}}\gamma \tag{48}$$

则 KFTWSVM2 的对偶问题为

(KDFTWSVM2)

$$\max_{\gamma} \quad e_1^{\mathrm{T}}\gamma-\frac{1}{2}\gamma^{\mathrm{T}}L(N^{\mathrm{T}}N)^{-1}L^{\mathrm{T}}\gamma \tag{49}$$

$$\text{s.t.} \quad 0\leqslant\gamma\leqslant c_2S_B \tag{50}$$

其分类过程同线性可分情形。

3 混合模糊隶属度函数设计

目前构造隶属度函数的方法有很多,但并没有一个可遵循的一般性准则,多数学者都用样本点到类中心的距离来定义其模糊隶属度,这种方法不能有效地区分支持向量与孤立点,会降低算法的分类精度。为了构造一个合适的模糊隶属度函数,不仅要考虑样本点到类中心的距离,还要考虑样本间的密切程度。本文基于此思想设计了一种混合模糊隶属度函数。设两类样本点在特征空间中的中心分别为 C_A、C_B,两类样本点在特征空间中到各自类中心的距离分别为 d_{iA}、d_{iB},半径分别为 r_A、r_B,点与点之间的距离为 d_{ij},则

$$C_A=\frac{1}{n_A}\sum_{x_i\in A}\Phi(x_i) \tag{51}$$

$$C_B=\frac{1}{n_B}\sum_{x_i\in B}\Phi(x_i) \tag{52}$$

$$d_{iA}^2=\|\Phi(x_i)-C_A\|^2 \tag{53}$$

$$d_{iB}^2=\|\Phi(x_i)-C_B\|^2 \tag{54}$$

$$r_A^2=\max_{x_i\in A}\|\Phi(x_i)-C_A\|^2 \tag{55}$$

$$r_B^2=\max_{x_i\in B}\|\Phi(x_i)-C_B\|^2 \tag{56}$$

$$d_{ij}=\|\Phi(x_i)-\Phi(x_j)\| \tag{57}$$

其中:n_A 是 1 类样本的数目;n_B 是 -1 类样本的数目;Φ 是从原始空间中映射到高维特征空间的非线性映射。基于距离的模糊隶属度为

$$p_i=\begin{cases}1-\dfrac{1}{1+(r_A^2-d_{iA}^2)+\delta} & x_i\in A\\ 1-\dfrac{1}{1+(r_B^2-d_{iB}^2)+\delta} & x_i\in B\end{cases} \tag{58}$$

其中:δ 为一个充分小的正数。把 d_{ij} 按大小进行排序,$d_{i1}\leqslant d_{i2}\leqslant\cdots\leqslant d_{i(l-1)}$,$i\in l$,取离 x_i 最近的 k 个点,则

$$t_i=\frac{1}{\sum_{j=1}^{k}d_{ij}} \tag{59}$$

$$T=\max(t_1,t_2,\cdots,t_l) \tag{60}$$

基于紧密度的模糊隶属度为

$$q_i=\frac{t_i}{T} \tag{61}$$

则结合距离和紧密度的模糊隶属度为

$$s_i=p_iq_i \tag{62}$$

4 仿真实验

为了评价本文算法的性能,笔者进行了如下的人工数据和实际问题的仿真实验。SVM 算法是用 MATLAB 6.5 语言实现的,实验所采用的计算机配置为 Celeron® CPU,主频 2.26 GHz,内存 1 GB。在实验中使用高斯核函数 $K(x,x_i)=\exp(-\|x-x_i\|^2/2\sigma^2)$。参数的选择采用 10 折交叉验证法。实验中比较了支持向量机(SVM)、双支持向量机(TWSVM)、基于距离模糊隶属度的模糊双支持向量机(FTWSVM-Distance)、基于紧密度模糊隶属度的模糊双支持向量机(FTWSVM-Affinity)和本文算法的分类性能。

4.1 含野点样本的实验结果

仿真数据来自于计算机随机产生的 400 个二维空间点,其中 100 个用于训练,300 个用于测试。在训练样本中随机加入 5% 的噪声。核函数参数 σ 分别取(1,2,10,50,100),C_1 和 C_2 分别取(1,10,30,50,70,100)进行训练。在最大正确率对应的 C 附近再取 C 值,找更高的正确率。不同的 C 值会得到不同的分类正确率。最佳参数为 $c_1=12$,$c_2=30$,$\sigma=50$。实验结果如表 1 所示。从表 1 可看出,本文算法在分类精度上高于普通的 TWSVM,能够降低噪声和孤立点的影响。本文算法的训练时间与普通的 TWSVM 基本相同,但优于传统的 SVM。

表 1 人工数据集上的分类结果比较

统计量	SVM	TWSVM	FTWSVM-Distance	FTWSVM-Affinity	本文算法
分类精度/%	78.32	79.16	82.59	82.35	84.26
训练时间/s	4.05	1.28	1.30	1.31	1.30

4.2 标准数据库的实验结果

对 UCI 机器学习数据库中的数据[11]进行了实验,结果

如表 2 所示。选择前 10% 的数据集作为训练样本,剩余的样本集作为测试样本。核函数参数 σ 分别取(1,2,10,50,100),C_1 和 C_2 分别取(1,10,30,50,70,100)进行训练,选取正确识别率最高时的 C 值。对每组数据集采用相同的参数 $c_1=75$,$c_2=100$,$\sigma=2$。实验结果如表 3 所示。从表 3 可以看出,本文算法在所有数据集上的分类精度均为最高,优于普通的 TWSVM;与传统的 SVM 相比,本文算法可以大大缩短训练时间。可见在二值分类中,本文方法具有一定的优势。

表 2　标准数据集

数据集	样本数	属性数
Australian	690	14
Heart-stalog	270	14
Hepatitis	155	19
Pima-Indian	768	8
Sonar	208	60

表 3　真实数据集上的分类结果比较

统计量	SVM	TWSVM	FTWSVM-distance	FTWSVM-affinity	本文算法
分类精度/%	85.51	85.80	86.08	86.02	88.07
训练时间/s	399.6	71.05	71.41	71.50	71.50
分类精度/%	84.07	84.44	85.65	85.12	87.92
训练时间/s	25.45	5.65	5.83	5.83	5.83
分类精度/%	80.00	80.79	83.36	83.32	85.60
训练时间/s	6.35	2.43	3.19	3.20	3.19
分类精度/%	75.68	75.70	76.20	76.19	76.36
训练时间/s	539.3	115.75	119.82	117.45	119.96
分类精度/%	79.79	80.26	80.61	80.57	82.61
训练时间/s	12.45	3.32	3.51	3.54	3.52

5　结束语

在模式分类问题中,双支持向量机的速度远远超过传统支持向量机。但在训练过程中,双支持向量机没有考虑到每个训练样本对分类超平面的不同作用。针对这个缺点,本文提出了一种基于混合模糊隶属度的模糊双支持向量机,给不同的训练样本赋予不同的模糊隶属度。实验证明,该模糊双支持向量机的分类性能优于传统的双支持向量机。

参考文献:

[1] BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery,1998,2(2):121-167.

[2] VAPNIK V N. The nature of statistical learning theory[M]. 2nd ed. New York: Springer,2000.

[3] FUNG G, MANGASARIAN O L. Proximal support vector machine classifiers[C]//Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2001:77-82.

[4] RESHMA K, JAYADEVA S C. Knowledge based proximal support vector machines [J]. European Journal of Operational Research,2009,195(3):914-923.

[5] YANG Xu-bing, CHEN Song-can, CHEN Bin, *et al*. Proximal support vector machine using local information[J]. Neurocomputing, 2009,73(1):357-365.

[6] SARAVANAN N, KUMAR V N S, RAMACHANDRAN K I. Fault diagnosis of spur bevel gear box using artificial neural network (ANN),and proximal support vector machine (PSVM)[J]. Applied Soft Computing,2010,10(1): 344-360.

[7] MANGASARIAN O L, WILD E W. Multisurface proximal support vector machine classification via generalized eigenvalues[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2006,28(1):69-74.

[8] XU Xiao-ming, JIANG Nan, DING Qiu-lin. Improved classification approach via GEPSVM[J]. Journal of Southwest Jiaotong University,2009,17(4):292-296.

[9] JAYADCVA R, KHEMCHANDANI S C. Twin support vector machines for pattern classification[J]. IEEE Trans on Pattern Analysis and Machinc Intelligence,2007,29(5):905-910.

[10] YE Qiao-lin, ZHAO Chun-xia, YE Ning. Least squares twin support vector machine classification via maximum one-class within class variance[J]. Optimization Methods & Software,2012,27(1): 53-69.

[11] http://mlearn. ics. uci. edu/databases/[EB/OL].

(上接第 431 页)规范缩写。因此,如果在预处理阶段采取一种自动的方法对文本进行拼写校对和名词术语规范化,得到较高的组全率是完全可能的,这也是本文下一阶段的研究重点。

4　结束语

首先探讨了维吾尔文传统分词方法导致的词义不完整性,以及这种分词结果在各种文本处理中产生的不良影响。以提取文本中语义具体及独立的语义词为主要目的,提出了维吾尔文组词的全新概念并实现了两种组词算法。分词过程中组词算法不考虑词间自然分隔符,而完全用双词互信息来度量相邻单词之间的结合能力,从而较准确地找到了应该切分的此间位置。对传统方法和本文组词算法的切分结果进行对比分析发现,整个文本集 40% 左右的单词独立性很差,当它们与某几个单词上下文共现时,它们在文本中出现才有真正的意义。本文算法较准确地发现并提取文本中真正有意义的语义词(单词或多词关联模式),因此在这种分词结果的基础上得到更好的文本处理效果是完全可能的。本文提出的算法也可以直接用到同语系的哈萨克文和柯尔克孜文的组词中,具有一定的推广意义。

参考文献:

[1] 曹勇刚,曹羽中,金茂忠,等. 面向信息检索的自适应中文分词系统[J]. 软件学报,2006,17(3):356-363.

[2] 孟春艳. 用于文本分类和文本聚类的特征抽取方法的研究[J]. 微计算机信息,2009,25(3):149-150.

[3] 阿力木江·艾沙,吐尔根·依布拉音,艾山·吾买尔,等. 基于机器学习的维吾尔文文本分类研究[J]. 计算机工程与应用,2011,36(7):110-112.

[4] 费洪晓,康松林,朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用,2005,30(7):67-69.

[5] 赵秦怡,王丽珍. 一种基于互信息的串扫描中文文本分词方法[J]. 情报杂志,2010,29(7):161-163.

[6] 黄巍,高兵,刘异,等. 基于词条组合的军事类文本分词方法[J]. 计算机科学,2010,37(2):171-173.

[7] 刘兵. Web 数据挖掘[M]. 北京:清华大学出版社,2009:12-44.

[8] 古丽拉·阿东别克,米吉提·阿布力米提. 维吾尔语义词切分方法初探[J]. 中文信息学报, 2004,18(6):61-65.

[9] 吐尔地·托合提,维尼拉·木沙江,艾斯卡尔·艾木都拉. 维、哈、柯全文搜索引擎中查询处理研究与实现[C]//第四届全国信息检索与内容安全学术会议. 2008:217-223.