

基于总间隔的模糊 v -相对间隔机的研究

龙章勇¹, 刘建华², 卢涵宇^{3,4}

(1 南京铁道职业技术学院 通信信号学院, 江苏 南京 210015; 2 浙江工商职业技术学院 工学院, 浙江 宁波 310053;
3 贵州大学 计算机与信息工程系, 贵州 贵阳 550003; 4 成都理工大学 信息科学与技术学院, 四川 成都 610059)

摘 要: 类间间隔和类内聚类性是影响分类器分类性能的两种重要因素. 基于模糊支持向量机和总间隔思想, 提出一种基于总间隔的模糊 v -相对间隔机(TMf-vRMM), 本方法本质上是传统相对间隔机(RMM)的扩展, 但可取得比 RMM 更好的分类性能. TMf-vRMM 通过使用差异成本和引入总间隔和模糊隶属度, 同时解决了不平衡训练样本问题和传统软间隔分类机 RMM 的过拟合问题, 显著提升学习机的泛化能力. 分别采用人造和实际数据集进行分类实验, 结果显示 TMf-vRMM 具有优于相关方法的稳定分类性能.

关键词: 相对间隔; 总间隔; 模糊; 支持向量机; 核方法

中图分类号: TP181

文献标识码: A

文章编号: 1000-7180(2012)06-0167-05

Research of Fuzzy v -Relative Margin Machine Based on Total Margin

LONG Zhang-yong¹, LIU Jian-hua², LU Han-yu^{3,4}

(1 College of Communication Signals, Nanjing Institute of Railway Technology, Nanjing 210015, China;

2 College of Technology, Zhejiang Business Technology Institute, Ningbo 310053, China;

3 Department of Computer and Information Engineering, Guizhou University, Guiyang 550003, China;

4 College of Information Science and Technology, Chengdu University of Technology, Chengdu 610059, China)

Abstract: The between-class margin and within-class cohesion are two important factors impacting on the performance of classifiers. This paper presents a novel classifier called total margin based fuzzy v -relative margin machine (TMf-vRMM) based on the idea of fuzzy support vector machine (FSVM) and total margin. Although it can be seen as a modified class of the classic RMM, TMf-vRMM has better theoretical classification performance than RMM. The proposed method solves not only the over-fitting problem resulted from outliers with the approaches of fuzzification of the penalty and total margin algorithm, but also the imbalanced datasets by using different cost algorithm, thus obtaining a lower generalization error bound. Experimental results obtained with synthetic and real datasets respectively show that the algorithm proposed in the paper is stable and superior to other related diagrams.

Key words: relative margin; total margin; fuzzy; support vector; kernel method

1 引言

基于统计学习理论和大间隔思想的支持向量机(SVM)^[1]及其相关变体是目前实现模式分类的主流方法之一, 其通过最大化类间间隔来达到强泛化能力. 目前, SVM 在机器学习和模式识别领域得到了广泛而成功地应用^[2]. v -支持向量分类机(v -

SVM)^[3]是 SVM 的一个扩展变体.

尽管 SVM 方法具有较明显优势, 但也存在一些问题^[4]. 针对 SVM 易出现过拟合问题, Wang 等人^[4]提出模糊 SVM(FSVM), 其基本思想是根据样本在它所在类别中的相关性, 给每个样本分配不同的隶属度值, 使那些相关度较小的样本得到的惩罚也相应较轻. 当非均衡样本集中的负样本的个数远

收稿日期: 2011-09-19; 修回日期: 2011-11-21

基金项目: 国家“八六三”计划项目(2006AA06Z110); 国家科技支撑计划项目(2009BAH50B02)

远小于正样本的个数时, SVM 会发生类边界倾斜现象, 从而使 SVM 在非均衡数据集中的应用受到了限制, 为了自适应非均衡的训练集, 引文^[5]采用不同代价算法, 从而降低了虚警率, 提高了识别的准确率。

本文受文献^[2,6]思想启发, 提出一种新颖的总间隔模糊 v -相对间隔机 (TMF- v RMM), TMF- v RMM 在传统的 RMM 中融合 FSVM 和 TM-SVM 方法思想, 通过对训练集模糊化和引入总间隔算法, 训练样本不再采用相同的方式进行处理, 而是根据它们相关度的不同而采用不同的处理方式, 同时解决了 RMM 中过拟合和样本不均衡等问题。从而进一步提升 RMM 的泛化性能。同时, TMF- v RMM 通过引入参数 v , 使得 TMF- v RMM 具有与 v -SVM 相同的特性, 即参数 v 能控制间隔误差的上界和支持向量的下界。

2 总间隔 v -RMM (TMF- v RMM) 算法

2.1 TMF- v RMM 优化目标与求解

基于 FSVM 方法思想, 通过引入模糊隶属度概念, 对每个训练数据赋予一个相对重要程度的隶属度值, 指示数据点对所属类的重要度, 数据点的重要度与隶属度值成正比。

设训练集 S 被分成两个集合: 模糊正训练集 S_f^+ 和模糊负训练集 S_f^- , 分别表示如下:

$$S_f^+ = \{x_i, y_i, \mu_i^+\}, i = L_1, \dots, L_p$$

$$S_f^- = \{x_i, y_i, \mu_i^-\}, i = L_1, \dots, L_n$$

$\mu_i^+ \in [\epsilon, 1], \mu_i^- \in [\epsilon, 1]$ 表示点 $x_i \in S_f^+$ 和 $x_i \in S_f^-$ 分别在正负类中的重要度。变量 ϵ 是一个非常小的正实数。 L_p 和 L_n 分别是模糊正训练集和模糊负训练集的个数, 且 $L_p + L_n = N$ 。

对于一个包含 N 个模式的二类分类问题, 设其中一类(或正常类)含有 m_1 个模式 $\{x_i, y_i\}_{i=1}^{m_1}$, 另一类(或异常类)含有 m_2 个模式 $\{x_i, y_i\}_{i=m_1+1}^{m_1+m_2}$, 其中, $x_i \in \chi \subset R^d$ ($1 \leq i \leq m_1 + m_2 = N$) 为输入数据, y_i 为类标签, 且当 $1 \leq i \leq m_1$ 时, $y_i = 1$; 当 $m_1 + 1 \leq i \leq m_1 + m_2$ 时, $y_i = -1$ 。线性可分情况下, TMF- v RMM 描述为

$$\begin{aligned} \min f(R, \rho, \xi, a) = & \frac{1}{2} \|w\|^2 - \nu\rho + \\ & \frac{1}{v_1 m_1} \left(\lambda_1 \sum_{i=1}^{m_1} \mu_i^+ \xi_i - (1 - \lambda_1) \sum_{i=1}^{m_1} \mu_i^+ \delta_i \right) + \\ & \frac{1}{v_2 m_2} \left(\lambda_2 \sum_{j=m_1+1}^N \mu_j^- \xi_j - (1 - \lambda_2) \sum_{j=m_1+1}^N \mu_j^- \delta_j \right) \end{aligned} \quad (1)$$

s. t.

$$y_i(w x_i + b) \geq \rho - \xi_i + \delta_i, \xi_i \geq 0, \delta_i \geq 0, 1 \leq i \leq m_1 \quad (2)$$

$$y_j(w x_j + b) \geq \rho - \xi_j + \delta_j, \xi_j \geq 0, \delta_j \geq 0, m_1 + 1 \leq j \leq N \quad (3)$$

$$w \cdot x_i + b \leq B, 1 \leq i \leq N \quad (4)$$

$$w \cdot x_i + b \geq -B, 1 \leq i \leq N \quad (5)$$

式中, B 为一大于等于 1 的实数, 实数 $\rho \geq 0$ 为两类模式分别与超平面间的间隔距离, $\xi = \{\xi_i \in R^N\}_{i=1}^N$ 为松弛向量, v_1, v_2 为二个正常数, 其中, $0 < \lambda_i < 1$, ($i=1, 2$), 这确保样本 x_i 的 ξ_i 和 δ_i 至少有一个为 0, μ_i^+ 和 μ_j^- 分别指正负类的模糊隶属度值。设 $\bar{x}$ 为样本平均值, 对于正类的模糊隶属度值 μ_i^+ 计算公式:

$$\mu_i^+ = \begin{cases} 1 - \frac{\|x_i - \bar{x}\|^2}{\max_j \|x_j - \bar{x}\|^2}, & \text{if } \mu_i^+ \geq \epsilon, \\ \epsilon, & \text{otherwise} \end{cases} \quad (6)$$

式中, $\|\cdot\|$ 指欧式距离, 下界 ϵ 为一用户指定的小实数, 所有正类数据的模糊隶属度值取值范围为 $[\epsilon, 1]$ 。

对于负类数据的模糊隶属度值 μ_j^- 的计算依此类推可得。

根据 Lagrange 定理, 原始问题(1)~(5)的对偶形式为

$$\begin{aligned} L(w, b, \rho, \xi, \delta, \alpha, \eta, \tau, \beta, \gamma, s) = & f - \sum_{i=1}^{m_1} \alpha_i (y_i(w x_i + b) - \rho + \xi_i - \delta_i) - \\ & \sum_{j=m_1+1}^N \alpha_j (y_j(w x_j + b) - \rho + \xi_j - \delta_j) + \\ & \sum_{i=1}^N \eta_i (w x_i + b - B) - \sum_{i=1}^N \tau_i (w x_i + b + B) - \\ & \sum_{k=1}^N \beta_k \xi_k - \sum_{m=1}^N \gamma_m \delta_m - s\rho \end{aligned} \quad (7)$$

式中, $\alpha_i \geq 0, \beta_i \geq 0, \eta_i \geq 0, \tau_i \geq 0, \gamma_i \geq 0$ 分别为 Lagrange 乘子。在 $L(w, b, \rho, \xi, \delta, \alpha, \eta, \tau, \beta, \gamma, s)$ 方程中分别对 w, ρ, ξ, b 等原始变量求偏导数可得

$$\begin{aligned} \frac{\partial L}{\partial w} = & w - \sum_{i=1}^N \alpha_i y_i x_i + \sum_{i=1}^N (\eta_i - \tau_i) x_i = 0 \\ \rightarrow w = & \sum_{i=1}^N (\alpha_i y_i - \eta_i + \tau_i) x_i \end{aligned} \quad (8)$$

$$\frac{\partial L}{\partial \rho} = -v + \sum_{i=1}^N \alpha_i - s = 0 \rightarrow \sum_{i=1}^N \alpha_i \geq v \quad (9)$$

$$\frac{\partial L}{\partial \xi_i} = \lambda_1 \frac{\mu_i^+}{v_1 m_1} - \alpha_i - \beta_i = 0 \quad (10)$$

$$\frac{\partial L}{\partial \xi_j} = \lambda_2 \frac{\mu_j^-}{v_2 m_2} - \alpha_j - \beta_j = 0 \quad (11)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^N (\alpha_i y_i - \eta_i + \tau_i) = 0 \quad (12)$$

$$\frac{\partial L}{\partial \delta_i} = -\frac{(1-\lambda_1)\mu_i^+}{v_1 m_1} + \alpha_i - \gamma_i = 0 \quad (13)$$

$$\frac{\partial L}{\partial \delta_j} = -\frac{(1-\lambda_2)\mu_j^-}{v_2 m_2} + \alpha_j - \gamma_j = 0 \quad (14)$$

将式(8)~(14)代入方程(7)中,得到 TMF-vRMM 原始优化问题的对偶形式:

$$\min_{\alpha, \eta, \tau} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\alpha_i y_i - \eta_i + \tau_i)(\alpha_j y_j - \eta_j + \tau_j) x_i x_j - B\left(\sum_{i=1}^N \eta_i + \sum_{i=1}^N \tau_i\right) \quad (15)$$

s. t

$$\frac{(1-\lambda_1)\mu_i}{v_1 m_1} \leq \alpha_i \leq \lambda_1 \frac{\mu_i}{v_1 m_1}, 1 \leq i \leq m_1 \quad (16)$$

$$\frac{(1-\lambda_2)\mu_j}{v_2 m_2} \leq \alpha_j \leq \lambda_2 \frac{\mu_j}{v_2 m_2}, m_1 + 1 \leq j \leq N \quad (17)$$

$$\sum_{i=1}^N (\alpha_i y_i - \eta_i + \tau_i) = 0 \quad (18)$$

$$\sum_{i=1}^N \alpha_i \geq v \quad (19)$$

对于约束 $\sum_i \alpha_i \geq v$,事实上该约束条件可加强为 $\sum_i \alpha_i = v$,这是因为在最优解中必有 $\rho > 0$ 成立,若不然,所有的 α_i 与 b 均为零,这将得到一个平凡解。从而,由 KKT 条件有 $s=0$,即得 $\sum_i \alpha_i = v$ 。

从 TMF-vRMM 的对偶形式及其约束条件可看出如下两个特性:

(1) 根据式(16)和式(17),TMF-vRMM 所有 Lagrange 乘子变量大于零,因此, TMF-vRMM 的所有训练样本均为支持向量。与此不同,基于软间隔的 RMM 的优化分割超平面仅由少数几个对应于 Lagrange 乘子变量大于零的训练样本构成;

(2) RMM 的 Lagrange 乘子变量 α_i 满足 $0 \leq \alpha_i \leq C$,当自由变量 C 一旦确定, RMM 所有训练样本所对应的优化解的可行域便固定且相同。而根据特性 1), TMF-vRMM 方法的优化解的上、下界针对每个数据点均不同,即 TMF-vRMM 的优化解的可行域是动态变化的,这也意味着非重要的数据点可具有更窄的解可行域。

从 TMF-vRMM 的优化形式可看出, TMF-vRMM 在寻求二类间总间隔最大化的同时,使得分别包含二类模式的平行区域尽量最小,即使得类内

数据在分割平面权重向量 w 方向上的投影间距离最小,从而克服传统方法 SVMs 对仿射和伸缩变换的敏感缺陷。

当正负两类样本在输入空间不为线性 TMF-vRMM 可分时,可引入非线性映射 $\phi^{[6]}$,将输入空间的模式样本映射到高维甚至无限维特征空间 H 中,实现 TMF-vRMM 线性可分,在特征空间 H 中两个向量 $\phi(x_i), \phi(x_j)$ 的内积可利用 Mercer 核函数 $K(x_i, x_j)$ 映射来表示计算,即有 $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ 。

2.2 w 和 b 的求解

考虑两个支持向量集合:

$$SV_1 =$$

$$\left\{ x_i \mid \frac{(1-\lambda_1)\mu_i}{v_1 m_1} \leq \alpha_i \leq \lambda_1 \frac{\mu_i}{v_1 m_1}, 1 \leq i \leq m_1 \right\} \quad (20)$$

$$SV_2 = \left\{ x_j \mid \frac{(1-\lambda_2)\mu_j}{v_2 m_2} \leq \alpha_j \leq \lambda_2 \frac{\mu_j}{v_2 m_2}, 1 + m_1 \leq j \leq N \right\} \quad (21)$$

根据 K. K. T. 条件^[6],对于输入 $x_i \in SV_1 (1 \leq i \leq m_1)$ 式(2)在线性不可分情况下变成一个所有松弛变量等于 0 的等式,即

$$y_i(w x_i + b) = \rho, 1 \leq i \leq m_1 \quad (22)$$

同理,对于输入 $x_j \in SV_2 (m_1 + 1 \leq j \leq N)$ 式(3)在线性不可分情况下也变成一个等式:

$$y_j(w x_j + b) = \rho, m_1 + 1 \leq j \leq N \quad (23)$$

设 $n_1 = |SV_1|, n_2 = |SV_2|, |\cdot|$ 表示集合基数,由式(22)~(23)可得

$$b^* =$$

$$\frac{\left(m_1 \sum_{x_i \in SV_1} \sum_{j=1}^N \alpha_j y_j k(x_i, x_j) \right) + \left(m_2 \sum_{x_j \in SV_2} \sum_{j=1}^N \alpha_j y_j k(x_i, x_j) \right)}{-m_1 m_2} \quad (24)$$

类间间隔为

$$\rho^* = \frac{1}{n_1} \sum_{x_i \in SV_1} y_i(w x_i + b^*) \quad (25)$$

2.3 决策函数

TMF-vRMM 的决策函数为

$$f(x) = \text{sign}\left(\sum_i \alpha_i y_i \langle x, x_i \rangle + b^*\right) \quad (26)$$

2.4 算法复杂度分析

按照 RMM 的算法求解方法,上述对偶问题通过迭代方式求解,每次迭代过程中,仅对偶变量的一个子集得以优化。令 q, r, s 分别指在某次迭代过程

中,对偶变量 α, η, τ 中自由(固定)变量的索引, p 指每次优化迭代过程中所选元素的个数,即 $p=|q|+|r|+|s|$,TMF-vRMM 在每次迭代过程中时间复杂度为 $O(p^3+n\log(n)+nq)$.

2.5 参数 v 属性

本文称满足式(20)、(21)的训练样本 $x_i(1 \leq i \leq N)$ 为支持向量(SV),对应于松弛变量 $\xi_i > 0$ 的训练样本 $x_j(1 \leq j \leq N)$ 称为间隔误差(ME).为了更好地选择优化的参数值,现给出如下定理:

定理 1^[5] TMF-vRMM 中参数 v 具有和 v-SVM 中参数 v 同样的性质,即能同时控制支持向量的下界和间隔误差的上界.

3 实验结果

为了验证 TMF-vRMM 的有效性,在人造数据和 UCI 机器学习数据集上进行了一系列的仿真实验.分别将 TMF-vRMM 与 RMM、TAF-SVM^[7]和 FSVM 等方法在二元分类上进行测试比较,实验结果显示,所提方法在核诱导的高维特征空间,通过引入样本的总间隔和模糊隶属度,具有较强的模式分类泛化性能.

3.1 实验参数设置

实验中所有样本首先归一化为 $[-1, 1]$,所有参数的协调通过网格搜索策略来选取.实验算法均采用高斯核函数(RBF):

$$K(x, y) = \exp\left(-\frac{1}{g} ||x - y||^2\right),$$

式中, g 值在网格 $\{\sigma^2/8, \sigma^2/4, \sigma^2/2, \sigma^2, \sigma^2 \times 2, \sigma^2 \times 4\}$ 中搜索选取,其中 σ 为训练样本平均范数的平方根.

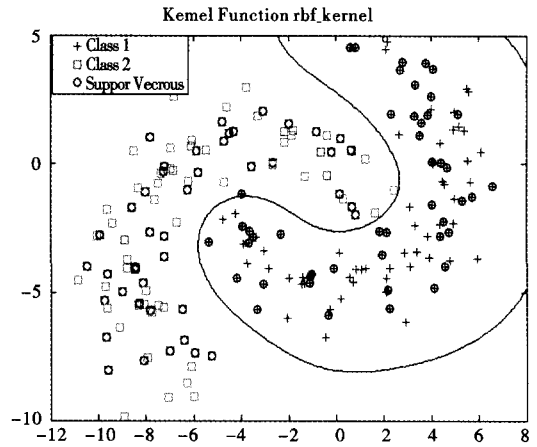
TMF-vRMM 方法中存在五个自由参数($v, v_1, v_2, \lambda_1, \lambda_2$)需要调节,实验中,令 $\lambda_1 = \lambda_2 \in [0, 1]$.根据文献^[5],参数 v 在网格 $\{0.01, 0.1, 0.5, 1\}$ 中选取中搜索选取,参数 v_1, v_2 从网格 $\{0.001, 0.01\}$ 中选取.

其他实验方法的参数均采取 10 重交叉验证法来选取.实验环境为 Matlab2009B.

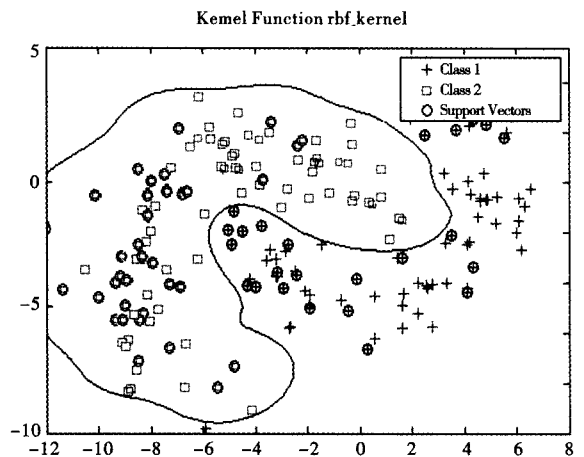
3.2 人造数据实验

本部分将通过与 RMM 方法进行比较,测试本文方法在 two-moons 数据集^[8]上的学习性能. two-moons 数据集大小为 200,其中正类的数据数为 101,负类数据数为 99,随机抽取训练集和测试集,重复 10 次,分别记录实验结果,且记录实验结果的平均值.二种方法的分类决策边界如图 1 所示.从图

1 结果可以看出:RMM 方法的分类边界过于平滑,导致学习过拟合;而所提方法 TMF-vRMM 由于引入了模糊隶属度和控制间隔误差的参数,从而取得了较好的泛化性能.



(a) RMM 分类精度:97.2%



(b) TMF-vRMM 分类精度:98.8%

图 1 two-moons 数据集上的决策超平面比较

3.3 UCI 实际数据集实验

为了进一步评价所提方法的分类性能,选取 8 个 UCI 实际数据集作为测试数据^[5],实验中所采用数据集如表 1 所示,其中多类数据集分类采取一对一(OAO)策略将其划分成多个二类分类问题.重点比较 TMF-vRMM 与 RMM、TAF-SVM、FSVM 等方法的分类性能.实验中各类数据集随机分割成训练集和测试集,通过 10 重交叉验证进行参数选择,每个数据集分别实验 10 次,取其平均值作为各自实验结论.表 2 中记录了 4 种方法各自的分类精度.从表 2 中数据可看出,所提方法对于所有实际数据集

均具有优于或可比较的模式分类性能.从而可得出结论:在考虑数据分布的同时,充分考虑数据的模糊隶属度和分类总间隔能明显提升传统支持向量学习机的泛化性能.

表 1 模式分类数据集

Datasets	# pos	# neg	m_1	m_2	d
Iris-setosa	50	100	50	100	4
Sonar-mines	111	97	111	97	60
Imports	71	88	71	88	25
Ecoli-periplasm	284	52	284	52	7
Cancer-wpbc	151	47	151	47	33
Breast	458	241	458	241	9
Heart	120	150	120	150	13
Abalone class1-8	1 407	2 770	1 407	2 000	10
Glass building nonfloat	76	138	76	138	9
Glass vehicle float	17	197	17	197	9

表 2 二元模式分类平均分类精度 %

Dataset	RMM	TAF-SVM	FSVM	TMF-vRMM
Iris-setosa	99.4	98.7	98.3	100
Sonar-mines	53.8	53.5	52.57	54.3
Imports	71.7	72.1	63.79	76.4
Ecoli-periplasm	95.2	97.6	83.2	100
Cancer-wpbc	84.5	66.7	86.4	87.7
Breast	97.9	98.6	96.5	99.6
Heart	84.1	85.4	83.8	86.2
Abalone class1-8	78.6	78.4	76.6	82.6
Glass building nonfloat	62.9	65.6	64.5	68.2
Glass vehicle float	93.7	92.2	93.2	94.2

4 结束语

基于模糊支持向量和总间隔方法思想的启发,提出了一种基于总间隔的模糊 v -相对间隔机 TMF-vRMM,其能显著增强支持向量机的模式分类泛化

性能.人造数据和 UCI 实际数据实验结果均显示 TMF-vRMM 具有相较于其他相关方法明显或相比较的模式分类性能优势.有待进一步研究的问题:TMF-vRMM 模糊隶属度的测量方法对学习机的泛化性能影响研究.

参考文献:

[1] Vapnik V. The nature of statistical learning theory [M]. Berlin: Springer Verlag, 1995.

[2] 彭新俊,王翼飞.总间隔 v -支持向量机及其几何问题[J].模式识别与人工智能,2009,22(1):8-16.

[3] 郭瑞华,程国建.基于核向量机的模式分类及其实验测试[J].微电子学与计算机,2010,27(9):190-196.

[4] 刘斌,王立梅.一种在线向量机增强学习算法[J].微电子学与计算机,2008,25(11):94-100.

[5] Schölkopf B, Smola A J, Williamson R, et al. New support vector algorithms [J]. Neural Computation, 2000, 12(5):1207-1245.

[6] Shivaswamy P, Jebara T. Maximum relative margin and data-dependent regularization [J]. Journal of Machine Learning Research, 2010, 11(3): 747-788.

[7] Yoon M, Yun Y, Nakayama H. A role of total margin in support vector machines [J] Proc Int Joint Conf Neural Network, 2003(3): 2049-2053.

[8] Belkin M, Niyogi P, Sindhwan V. Manifold regularization: a geometric framework for learning from examples [J]. Journal of Machine Learning Research, 2006(7):2399-2434.

作者简介:



龙章勇 男,(1979—),硕士,讲师.研究方向为通信及模式识别、计算机网络技术.

刘建华 男,(1979—),硕士,讲师.研究方向为计算机网络通信、软件工程.

卢涵宇 男,(1978—),博士研究生,讲师.研究方向为现代信息处理技术、计算机应用技术.