

一种改进的 SVM 决策树及在遥感分类中的应用 *

丁胜锋^{1,2}, 孙劲光¹, 陈东莉³, 李 扬¹, 姜晓林²

(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; 3. 中国石化抚顺石化公司 石油二厂科技信息部, 辽宁 抚顺 113004)

摘 要: 针对遥感图像分类问题提出了一种基于遗传算法和 K 近邻的 SVM 决策树方法。算法以基于类分布的类间分离性测度为准则, 利用遗传算法对传统的 SVM 决策树进行优化, 生成最优(较优)决策树。在分类阶段, 对容易分的节点利用 SVM 进行分类, 而对可分离性差的节点采用 SVM 和 K 近邻相结合的分类方法, 最终实现多类别分类。实验结果表明, 与传统的分类方法相比, 该算法的实验效果较好, 可有效地提高遥感图像的分类精度。

关键词: 遗传算法; K 近邻; 支持向量机决策树; 遥感图像分类

中图分类号: TP391.41

文献标志码: A

文章编号: 1001-3695(2012)03-1146-03

doi:10.3969/j.issn.1001-3695.2012.03.095

Improved SVM decision-tree and its application in remote sensing classification

DING Sheng-feng^{1,2}, SUN Jin-guang¹, CHEN Dong-li³, LI Yang¹, JIANG Xiao-lin²

(1. Liaoning Technical University, Huludao Liaoning 125105, China; 2. School of Economics & Management, Liaoning Shihua University, Fushun Liaoning 113001, China; 3. Refinery NO.2 Science & Technology Department, PetroChina Fushun PetroChemical Company, Fushun Liaoning 113004, China)

Abstract: This paper presented a SVM decision-tree algorithm based on GA and KNN. First, GA was used to create optimal or near-optimal decision-tree, which defined a novel separability measure. Then in the class phase, standard SVM was used to make binary classification for the divisible nodes, and SVM combined with KNN were used to classify the fallible nodes. Finally, achieved the multi-classification by the SVM decision-tree. Experimental results show that the proposed method can effectively improve the classification precision of remote sensing image in comparison to traditional classification methods.

Key words: genetic algorithm; K-nearest neighbors; support vector machine(SVM) decision-tree; classification of remote sensing image

遥感图像的计算机分类是遥感技术应用的一个重要组成部分。在遥感图像分类中比较常用的监督分类方法是统计分类法, 包括平行管道方法、最小距离法、最大似然分类方法和光谱角制图等。但是传统的统计模式方法常常建立在统计分析和大数定律基础上, 只有当训练样本数目趋于无穷时, 分类性能才能达到最优, 而且样本数目随着波段数的增加而增加。遥感图像分类总是在样本有限的情况下进行的, 所以难以取得理想的分类效果^[1]。20 世纪 80 年代末, 神经网络开始应用于遥感图像分类, 并能够提供较理想的分类结果。然而神经网络存在着结构选取困难、局部最小化以及过学习等问题^[2]。

支持向量机(SVM)是由 Vapnik 及其合作者提出的一种新的学习方法^[3], 是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上, 根据有限样本信息在模型复杂性和学习能力之间寻求最佳折中, 以期获得最好的推广能力。在遥感图像的分类研究中, SVM 具有小样本学习、非线性、高维数、抗噪声性能、学习效率高与推广性好的优点。

1 经典的 SVM 多类分类方法

1) 一对一(one-against-one)^[4] 该方法在每两类样本间训

练一个 SVM 分类器, 共得到 $k(k-1)/2$ 个分类器, 与每一类相关的分类器个数为 $k-1$ 。预测未知样本时, 每个分类器都对其类别进行判断, 并为相应的类别投上一票, 最后得票最多的类别即作为该未知样本的类别。

2) 一对多(one-against-rest)^[5] 该方法依次用一个 SVM 分类器将每一类与其他类别区分开, 共得到 K 个分类器。预测未知样本时, 将样本归类于具有最大决策函数值的那类。

一对一和一对多这两种方法, 当类别数目较多时, 训练速度和分类速度比较慢, 而且均存在不可识别区域。后来出现了几种改进算法, 其中有 ECOC^[6]、DAG^[7]。ECOC 方法选择合适的码本比较困难; DAG 方法的缺点是对类别数目多的分类问题, 训练速度较低; 根节点的选择可能直接影响着分类的结果。

3) 决策树(DT-SVM)^[8] 该方法首先将所有类别划分为两个子类, 再对两个子类作进一步划分, 如此循环, 直到子类中只包含一个类别为止, 这样, 就得到了一个倒立的二叉树。然后对每个决策节点的二类分类问题用 SVM 解决。SVM 决策树方法对于 k 类分类问题, 只需构造 $k-1$ 个 SVM 分类器, 具有较高的训练速度和分类速度, 也不存在拒分区域。但 SVM 决策树存在错分累积问题, 分类错误在越靠近树根的地方发

收稿日期: 2011-07-13; 修回日期: 2011-09-03 基金项目: 辽宁省科技计划资助项目(2010401010)

作者简介: 丁胜锋(1981-), 男, 湖北蕲春人, 讲师, 博士研究生, 主要研究方向为图形图像、数据挖掘(jgdsf@163.com); 孙劲光(1962-), 女, 辽宁阜新, 教授, 博导, 主要研究方向为图形图像、数据挖掘; 陈东莉(1979-), 女, 辽宁抚顺人, 工程师, 本科, 主要研究方向为数据挖掘; 李扬(1985-), 男, 辽宁阜新, 硕士研究生, 主要研究方向为模式识别; 姜晓林(1975-), 男, 山东招远人, 讲师, 博士, 主要研究方向为系统工程、信息系统开发。

生,其分类性能越差。

2 基于遗传算法和 K 近邻的 SVM 决策树分类方法

传统的 SVM 决策树采用固定树结构,而且包括根节点在内各个节点的选择具有随意性,从而使得分类性能具有不确定性。为构造分类性能良好的决策树,可以考虑将容易分的类先分离出来,然后再分不容易分的类,这样就能够使可能出现的错分尽可能远离树根。本文在 SVM 决策树的训练过程中,以基于类分布的类间分离性测度为准则,利用遗传算法在每个节点得到最优的两类分类方案,最终自适应地生成最优(或较优)决策树。在分类过程中,对容易分的节点利用 SVM 进行分类。越靠近分类树最下层的节点,类间可分离性就越差,分类识别率越低。通过对 SVM 分类时错分样本的分布进行分析发现,其出错样本点都在分界面附近,要提高分类性能必须尽量利用分界面附近的样本提供的信息。SVM 分类器等价于每类只有一个代表点的最近邻分类器。所以在类间可分离性差的分类树节点采用 SVM 和 K 近邻相结合的方法进行分类。

2.1 决策树训练过程

2.1.1 遗传算法设计

利用遗传算法,在决策树每个节点将多类训练样本划分为两类,应使可分离性好的类先划分出来,最终生成最优(或较优)的分类树。

1)染色体的构造

染色体采用二进制的方法进行编码,染色体结构如图 1 所示。染色体长度 K 为当前训练样本的类别总数,其中每个基因位对应一个类别。解码时对应基因为“1”的类别分为一类,为“0”的分为另一类。在这些编码中,全“1”和全“0”的编码表示将所有的训练样本分为一类,是无用编码。另外,某种分类的等效编码也是应该避免的,如 110000 和 001111 的编码对应一种分类方法,都表示将第 1 类和第 2 类的训练样本分为一类,将其余 4 类的训练样本分为另一类。因此要对初始种群产生的染色体和交叉、变异后产生的新一代染色体进行合理性检验。

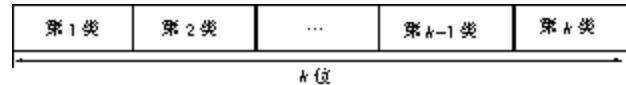


图 1 染色体编码方式

2)初始种群设定

随机产生初始种群的染色体,直到符合要求为止。

3)适应函数设计

这里以基于类分布的类间分离性测度作为适应度函数,以使每个节点类间可分离性尽可能大。根据训练数据估计各类间易分性,通常的做法是用类间的 Euclidean 距离作为分离性测度^[9],但这种方法的缺点在于类中心间的距离远近并不能够代表类间的分离度。本文采用基于类分布的类间分离性测度,既考虑了类中心间的距离,又考虑了类的分布情况。

已知两类模式的训练样本分别为 $\{x_1, x_2, \dots, x_{n_1}\}$ 和 $\{x'_1, x'_2, \dots, x'_{n_2}\}$,经非线性映射 Φ 作用,映射到高维特征空间 H 中,类中心分别为 C_Φ 和 C'_Φ ,则在特征空间中训练样本 x 到类中心 C_Φ 的距离为

$$d^H(x, C_\Phi) = \sqrt{k(x, x) - \frac{2}{n_1} \sum_{i=1}^{n_1} k(x, x_i) + \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(x_i, x_j)}$$

(1)

进而求得特征空间中的类方差为

$$\sigma^H = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} d^H(x, C_\Phi)$$

(2)

在特征空间中类中心 C_Φ 和 C'_Φ 间的 Euclidean 距离为

$$d^H(C_\Phi, C'_\Phi) = \sqrt{\frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K(x_i, x_j) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(x_i, x'_j) + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} K(x'_i, x'_j)}$$

(3)

在特征空间中类 i 和类 j 间的分离性测度定义如下:

$$sm_{ij}^H = \frac{d^H(C_\Phi^i, C'_\Phi^j)}{(\sigma_i^H + \sigma_j^H)}$$

(4)

则适应度函数为

$$fit(a_t) = sm_{ij}^H = \frac{d^H(C_\Phi^i, C'_\Phi^j)}{(\sigma_i^H + \sigma_j^H)}; t = 1, 2, \dots, n$$

(5)

4)遗传操作的设计

(1)选择算子 采用轮盘赌的选择方式,若个体 a_t 的适应度函数为 $fit(a_t)$,则选中 a_t 为下一代个体的概率为

$$p(a_t) = \frac{fit(a_t)}{\sum_{i=1}^n fit(a_i)}$$

(6)

(2)交叉算子 采取一点交叉的方法,以交叉概率 P_c 在交叉点相互交换两个配对个体的部分染色体。

(3)变异算子 采用基本位变异法,对个体编码串中以变异概率 P_m 随机指定某一位或某几位基因座上的基因值作变异运算,对每个指定的变异点,对其基因值进行取反运算或用其他等位基因值来代替,从而产生出新的个体。

5)停止条件

以最容易分割的类是否 f 代未变以及是否达到最大进化代数 $S_{\max}$ 作为终止条件。若满足,则算法停止;若不满足,转步骤 4)继续。

2.1.2 决策树生成算法

a)对全部类别按二进制的方法进行编码,在根节点调用遗传算法得到最优(或较优)染色体,按解码后的分类方法训练 SVM,将训练样本分为两类。

b)判断各子节点是否只包含一类样本,若是转 d);反之转 c)。

c)如果子节点包含两类以上样本,按编码方式生成新的染色体,并调用遗传算法得到最优(或较优)染色体,按解码后的分类方法训练 SVM,将训练样本分为两类,转 b);如果子节点包括两类样本,直接训练 SVM 将其分开,转向 b)。

d)算法结束,生成最优(或较优)分类树。

2.2 决策树分类过程

对于待分样本 x ,如果 x 分类时遍历的决策树节点类间可分性好,则利用 SVM 得出 x 的类别;否则先给定一个分类阈值 ε 。如果类间可分性差对应的 SVM 决策函数绝对值大于 ε ,则利用 SVM 得出 x 的类别;否则利用 KNN 得出 x 的类别。算法过程如下:

设 T 为测试集, S 为训练得到的可分性差类对应的分类器集合, ε 为分类阈值, t 为 KNN 的个数。

a)如果 $T \neq \emptyset$,取 $x \in T$,如果 $T = \emptyset$,停止。

b)如果 x 分类时遍历的分类器 $S_i \notin S$,则计算 $f(x) = \text{sgn} \{ \sum_t \alpha_t y_t K(x_t, x) + b \}$ 判断 x 的类别,转 e);否则转 c)。

c)令 $g(x) = \sum_t \alpha_t y_t K(x_t, x) + b$,如果 $|g(x)| > \varepsilon$,则计算 $f(x) = \text{sgn} \{ \sum_t \alpha_t y_t K(x_t, x) + b \}$ 判断 x 的类别,转 e);否则

转 d)。

d)以分类器 S_i 的支持向量集合 S_i^w 作为 KNN 代表点集合,计算 x 和每个支持向量在特征空间的距离 $d_i^H(x, x_i) = \|\Phi(x) - \Phi(x_i)\|^2 = K(x, x) - 2K(x_i, x) + K(x_i, x_i)$, $x_i \in S_i^w$,并对 $d_i^H(x, x_i)$ 从小到大进行排序,取前 t 个支持向量作为 x 的 t 个近邻, t 个近邻多数属于哪类, x 就属于哪类,转 e)。

e)如果 x 已分完类,则 $T = T - \{x\}$,转 a),否则转 b)。

3 实验结果及分析

3.1 实验数据及环境

本文所用的遥感影像数据为某地 Landsat-TM 影像,选取的影像为 TM1 ~ TM7 波段。图像大小为 640 × 400 像元。图 2 为 TM5 波段(中心波长为 1.65 μm)、TM4 波段(中心波长为 0.83 μm)、TM3 波段(中心波长为 0.66 μm)的合成图像。实验硬件平台为 Celeron® CPU,主频为 2.26 GHz,内存为 1 GB,软件为 ENV4.7/IDL 和 MATLAB 7.0。

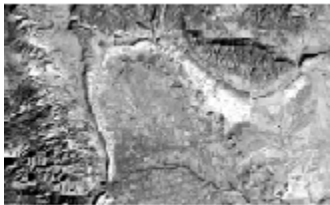


图2 原始图像

3.2 样本选择

样本选取要充分考虑各种地物的光谱结构和纹理特征,共选择有 6 种地类的 1 279 个训练样本和 3 099 个测试样本,如表 1 所示。

表 1 Landsat-TM 影像测试数据

类别号	土地覆盖类别	训练样本数	测试样本数	类别号	土地覆盖类别	训练样本数	测试样本数
C1	裸地	875	2073	C4	森林	146	105
C2	耕地	107	345	C5	草地	29	33
C3	沙地	105	506	C6	水域	17	37

3.3 实验结果

为了验证本文算法的有效性,实验还使用了传统的遥感分类方法和经典的 SVM 多类分类方法,通过分类结果图、混淆矩阵、分类精度和运行时间四个方面对比各种分类方法的效果。

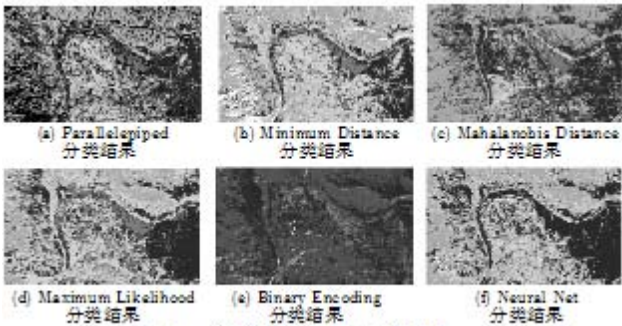


图3 传统分类方法的分类结果

1)传统的遥感分类方法的分类效果

传统的遥感分类方法主要包括平行管道法、最小距离法、马氏距离法、最大似然法、二进制编码法和神经网络法,分类结果如图 3 所示,分类精度和运行时间如表 2 所示。由于篇幅的限制,传统分类方法分类的混淆矩阵未列出。从实验结果看,

在传统分类方法中,最大似然分类方法的精度最高,Kappa 系数为 0.834 2,分类总体精度为 90.674 4%,运行时间为 3.09 s;神经网络方法的精度最低,Kappa 系数为 0.627 2,分类总体精度为 79.769 9%,运行时间为 45.29 s。实验结果说明,传统的遥感图像分类方法存在精度比较低和难以收敛等问题。

表 2 传统分类算法的精度和耗时统计

统计量	平行管道	最小距离	马氏距离	最大似然	二进制编码	神经网络
总体精度/%	72.120 0	73.927 1	87.738 0	90.674 4	79.769 9	75.701 8
Kappa 系数	0.568 1	0.610 6	0.786 9	0.834 2	0.627 2	0.487 0
运行时间/s	2.52	2.36	3.01	3.09	2.48	45.29

2)SVM 多类分类方法的分类效果

经典的 SVM 多类分类方法主要包括一对一 SVM 多类分类方法、一对多 SVM 多类分类方法和 SVM 决策树方法,其分类结果如图 4(a)~图 4(c)所示,混淆矩阵、分类精度和运行时间如表 3~7 所示。与传统的遥感分类方法比较,SVM 多类分类方法尽管时间有所损耗,但都是可行的,并且精度都得到了提高。在所有分类方法中,本文算法的分类精度最高,Kappa 系数为 0.916 8,分类总体精度为 95.579 2%。本文算法的分类精度优于一般的 SVM 决策树,说明本文算法能有效解决错分累积问题,提高 SVM 决策树的分类精度。本文算法的运行时间为 8.73 s,优于一对一 SVM 多类分类方法和一对多 SVM 多类分类方法。可见在遥感图像分类中,本文方法具有一定的优势。

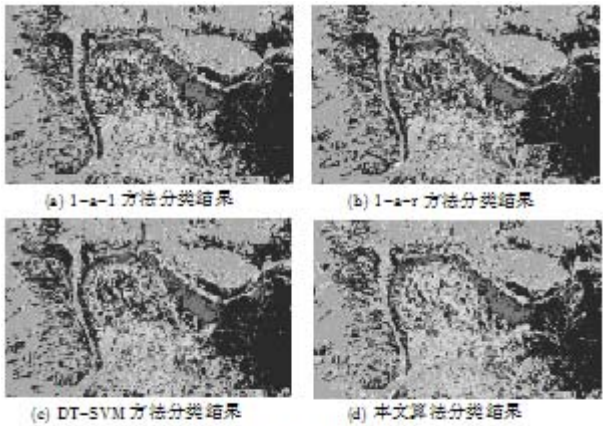


图4 SVM 多类分类结果

表 3 1-a-1 方法分类的混淆矩阵

类别	C1	C2	C3	C4	C5	C6	类正确率/%
C1	1 941	0	0	3	129	0	93.63
C2	0	345	0	0	0	0	100.00
C3	6	0	500	0	0	0	98.81
C4	0	0	0	105	0	0	100.00
C5	0	0	0	22	11	0	33.33
C6	0	0	0	12	0	25	67.57
总检验样本	3 099	正确分类样本		2 927	总正确率		94.449 8

表 4 1-a-r 方法分类的混淆矩阵

类别	C1	C2	C3	C4	C5	C6	类正确率/%
C1	1 930	0	0	4	139	0	93.10
C2	0	345	0	0	0	0	100.00
C3	5	0	501	0	0	0	99.01
C4	0	0	0	105	0	0	100.00
C5	0	0	0	22	11	0	33.33
C6	0	0	0	12	0	25	67.57
总检验样本	3 099	正确分类样本		2 917	总正确率		94.127 1

训练数据	AER	FRR	FAR
10	35.30%	32.48%	12.50%
15	26.78%	21.15%	17.78%
18	22.74%	18.73%	19.34%

表 3 采用 127 个中心点, $T=4000$

训练数据	AER	FRR	FAR
10	20.56%	21.78%	10.38%
15	18.38%	18.92%	12.42%
18	16.56%	17.21%	13.04%

从表 13 中可以发现,随着采用的中心点的增加,识别的准确率和误报率将下降。

作为对比,通过采集到的中心点,利用聚类(K-means)的方法将各个训练样本投影到特征空间中,然后求取测试样本与各个类别的空间距离,将测试样本赋予离其最近的签名类别。表 4、5 中显示了采用聚类方法获得的实验结果。

表 4 采用 127 个中心点, $T=4000$

聚类数目	AER	FRR	FAR
2	24.56%	25.87%	15.88%
4	17.34%	21.52%	19.32%
6	26.16%	29.11%	17.76%

在表 5 中,对比了本文方法和采用聚类方法进行静态签名识别的结果,表格数据为 AER。

表 5 本文方法与聚类方法实验对比 %

方法	15 个中心点	31 个中心点	127 个中心点
本文方法	20.56	21.78	10.38
聚类方法	27.37	25.11	17.34

从表 4、5 中可以看到,采用本文方法提取特征后,依据本文所设计的识别方法,其准确率要高于采用聚类方法的识别率。

(上接第 1148 页)

表 5 DT-SVM 方法分类的混淆矩阵

类别	C1	C2	C3	C4	C5	C6	类正确率/%
C1	1 928	0	0	0	145	0	93.01
C2	5	333	0	0	7	0	96.52
C3	77	0	429	0	0	0	84.78
C4	0	0	0	86	19	0	81.90
C5	0	0	0	17	16	0	48.48
C6	0	0	0	0	14	23	62.16
总检验样本	3 099	正确分类样本		2 816	总正确率		90.866 7

表 6 本文算法分类的混淆矩阵

类别	C1	C2	C3	C4	C5	C6	类正确率/%
C1	1 975	0	0	3	95	0	95.27
C2	0	345	0	0	0	0	100.00
C3	8	0	498	0	0	0	98.42
C4	0	0	0	105	0	0	100.00
C5	0	0	0	19	14	0	42.42
C6	0	0	0	12	0	25	67.57
总检验样本	3 099	正确分类样本		2 962	总正确率		95.579 2

表 7 SVM 多类分类算法的精度和耗时统计

统计量	1-a-l	1-a-r	DT-SVM	本文算法
总体精度/%	94.449 8	94.127 1	90.866 7	95.579 2
Kappa 系数	0.897 0	0.891 5	0.873 9	0.916 8
运行时间/s	18.90	16.48	2.74	8.73

4 结束语

本文针对传统遥感图像分类的缺点,提出了一种基于遗传

4 结束语

本文提出了一种新颖的、依据签名图像几何中心的特征提取方法,并在此基础上设计了一种签名图像的识别算法。实验结果显示,随着提取中心点个数的增加,识别准确率显著上升,并且本文的识别方法相对于其他识别算法具有更高的准确率。由于本文方法无须机器学习的过程而只需简单地查表比对,所以运算效率更高,更适合于移动设备或者手持设备的使用。

参考文献:

[1] DIMAURO G, IMPEDOVO S, LUCCHESI M G, et al. Recent advancement in automatic signature verification [C]//Proc of the 9th International Workshop on Frontiers in Handwriting Recognition. Washington DC:IEEE Computer Society, 2004:179-184.

[2] BAJAJ R, CHAUDHARY S. Signature verification using multiple neural classifiers [J]. Pattern Recognition, 1997, 30(1):1-7.

[3] 肖春景, 李春利, 张敏. 脱机手写体签名识别的小波包隐马尔可夫模型[J]. 计算机应用, 2010, 30(2):445-448.

[4] 陈刚, 李弼程, 庄军, 等. 基于矩特征和证据理论的离线签名识别[J]. 计算机工程与应用, 2005, 41(18):57-60.

[5] 戴斯获, 夏利民. 基于保局投影的离线签名识别[J]. 计算机工程与应用, 2010, 53(1):190-193.

[6] GURU D S, PRAKASH H N. On-line signature verification and recognition: an approach based on symbolic representation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(6):1059-1073.

[7] 李京兵, 黄席祺. 一种基于 DWT 抗几何攻击数字水印鲁棒算法[J]. 计算机仿真, 2007, 24(3):303-307.

算法和 K 近邻的 SVM 决策树方法。实验结果表明这种分类方法具有很高的分类精度,并且在运行时间上是可行的,与传统的统计模式方法、神经网络方法和经典的 SVM 多类分类方法相比具有一定的优势,是一种有效的遥感图像分类方法。

参考文献:

[1] 惠文华. 基于支持向量机的遥感图像分类方法[J]. 地球科学与环境学报, 2006, 8(2):93-95.

[2] 元昌安. 数据挖掘原理与 SPSS Clementine 应用[M]. 北京:电子工业出版社, 2009:247.

[3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1):32-42.

[4] LNOUE T, ABE S. Fuzzy support vector machines for pattern classification [C]//Proc of International Joint Conference on Neural Networks. 2001:1449-1454.

[5] KRESSEL U. Pairwise classification and support vector machines [M]//Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1999: 255-268.

[6] YAN Zhi-gang. Research on ECOC SVMs [C]//Proc of World Congress on Intelligent Control and Automation (WCICA). 2010:2838-2842.

[7] PLATT J C, CRISTIANINI N, SHAW T J. Large margin DAGs for multiclass classification [C]//Proc of Neural Information Processing Systems. Cambridge: MIT Press, 2000: 547-553.

[8] ARUN K M, GOPAL M. Fast multiclass SVM classification using decision tree based one-against-all method [J]. Neural Processing Letters, 2010, 32(3):311-323.

[9] TAKAHASHI F, ABE S. Decision-tree-based multi-class support vector machines [C]//Proc of ICONIP'02. [S. l.]: IEEE Press, 2002: 1419-1422.