文章编号:1000 - 582X(2006)01 - 0061 - 05

利用支持向量机 SVM★识别车辆类型*

肖汉光[1,2],蔡从中[1,2],王万录[1]
(1.重庆大学 应用物理系,重庆 400030; 2. 新加坡国立大学 计算科学系,新加坡 117543)

摘 要:支持向量机(Support Vector Machine, SVM)分类方法在实际二类分类问题的应用中显示出良好的学习和泛化能力,已被广泛地应用于多类分类问题的研究.以车辆轮廓特征为对象,将二类分类支持向量机 SVM★应用于多类车辆类型的识别,并与其它分类器的分类结果进行了对比.通过 9次交叉验证实验,结果表明 SVM★对车辆数据样本的测试准确率达到了 85.59%,其分类性能优于其它分类器.

关键词:支持向量机;车辆识别;轮廓特征
中图分类号:TP18; U49 **文献标识码**:A

目前,人们生活水平普遍提高,车辆拥有量不断上升,但道路建设的相对滞后使得交通更加拥挤.在有限的道路资源下,利用智能交通系统进行车辆管理能有效地缓解该类问题.智能交通的重要组成部分是车辆类型识别.对车辆类型识别而言,关键问题在于选择准确且有效的车体特征和模式识别方法.特征提取的有效性和准确性直接关系到识别系统的最终效果.现有的车辆特征提取方法包括图像特征提取,感应线圈特征提取和声表面波特征提取等.感应线圈的应用原理是利用不同的车辆通过环形线圈时,其电感量输出特征曲线不同来判断车的类型.但此种方法易受较多的不定因素的影响,安装和维护感应线圈较为麻烦,使用周期短.随着计算机的发展,基于图像处理技术的车辆特征(如汽车颜色、轮廓等)提取正迅速发展起来.利用汽车颜色特征进行车辆分类时,车辆的车身颜色易受外围灯光颜色的影响和背景的干扰,且随着光线强度的变化其识别结果也会有所不同[1],因而该方法的可信度不高.而汽车外部轮廓不易受外部环境的影响,并且特征信息较多,识别准确率受单个特征变化的影响不大.所以,车辆轮廓是一种较为可靠的识别特征.

支持向量机(Support Vector Machine, SVM)是由 Vapnik及其合作者[2]基于结构风险最小化原理提出的一种有监督的统计学习方法,被公认为小样本情况下统计及其学习的经典.由于其不需要确定各类的条件概率密度和先验概率就能找到全局最优解,并且具有较好的泛化能力,所以被广泛的应用于诸多领域,如文本分类,手写体数字识别,语音识别,图像识别与目标探测,人脸识别,商业时序预报,水文预报,空气质量预报,地球空间物理和实验高能物理数据分析与处理,肿瘤及癌症诊断,基因微阵列表达数据分析,药物设计,蛋白质 -蛋白质相互作用预测以及蛋白质结构与功能预测等[3-8].目前,支持向量机已经从 2类分类问题发展到多类分类问题,最为常见的 2种方式是:一对一(ONE - VS - ONE)[9]和一对多(ONE - VS - ALL)[10].也有通过向量增维将多类分类转换成两类分类的方法.笔者利用 ONE - VS - ALL的方法,以汽车轮廓特征为对象,对车辆类型进行分类,并将获得的结果和其他类别分类器(贝叶斯分类方法,神经网络方法,决策树方法 C4.5,最近邻方法(K - NN)等)进行了比较.

1 支持向量机(SVM)分类原理

1.1 线性可分情形

假定对于给定的 n个线性可分的训练样本(Training samples)(x_1, y_1), (x_2, y_2), …, (x_n, y_n), $y_i \in$

* 收稿日期:2005 - 08 - 26
基金项目:重庆大学与新加坡国立大学国际合作研究资助项目
作者简介:肖汉光(1980 -),男,湖北石首人,重庆大学硕士研究生,新加坡国立大学访问学者,主要从事机器学习、模式识别和特征提取等研究工作.

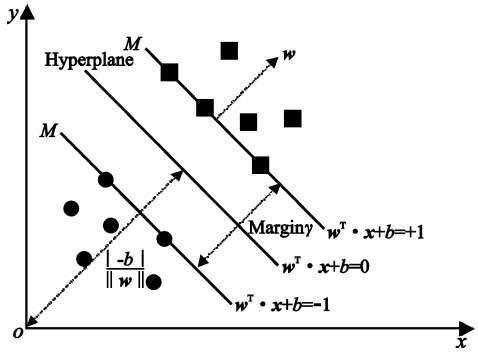
$\{-1, +1\}$,存在权向量 w 及偏置 b(参见图 1),满足:

$$w^T \cdot x_i + b \geq 1, \quad y_i = +1, \tag{1}$$

$$w^T \cdot x_i + b \leq -1, \quad y_i = -1, \tag{2}$$

合并式 (1)、式 (2)即得

$$y_i (w^T \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, n. \tag{3}$$



(图中圆点和方点分别代表 -1类和 +1类样本)

图 1 超平面及边界的定义

对于同一组训练样本,可以用不同的超平面 (Hyperplane)将它们区分开. SVM 分类的目的就是寻找出最佳权重 w_0 和最佳偏置 b_0,它们既能将两类数据最好地分离开来,同时又对测试样本 (Testing Samples)具有最佳泛化能力. 在众多超平面中,能使两类边界 (Margin)上的点到超平面的距离最大的超平面,称为最优超平面 (Optimal Separating Hyperplane, OSH),其对应两类的边界则称为最优边界 (Optimal Margin, OM). 最优超平面具有最优泛化能力.

设超平面 $H(w, b)$的方程为

$$w^T \cdot x + b = 0, \tag{4}$$

则两边界之间的距离 γ 为

$$\gamma(w, b) = \min_{\{x|y=+1\}} \frac{x^T \cdot w}{\|w\|} - \max_{\{x|y=-1\}} \frac{x^T \cdot w}{\|w\|}. \tag{5}$$

最优超平面 OSH (w_0, b_0)应使式 (5)值达到最大,即

$$\gamma(w_0, b_0) = \frac{2}{\|w\|}. \tag{6}$$

因而,求最优超平面问题就归结为求解如下二次规划 (Quadratic Programming, QP)问题:

$$\Phi(w, b) = \min \frac{1}{2} \|w\|^2,$$

$$\text{s.t.} \quad y_i (w^T \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, n \tag{7}$$

问题 (7)式的 Lagrange目标函数为

$$L(w, b, a) = \frac{1}{2} w^T \cdot w - \sum_{i=1}^{n} \alpha_i [y_i (w^T \cdot x_i + b) - 1]. \tag{8}$$

规划问题 (7)式的解由 Lagrange函数 (8)式的鞍点决定:

$$\left.\frac{\partial}{\partial w}\right|_{w=w_0} = w_0 - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, \tag{9}$$

$$\left.\frac{\partial}{\partial b}\right|_{b=b_0} = \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{10}$$

将式 (9)、式 (10)代入式 (8)可得

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T \cdot x_j),$$

$$\text{s.t.} \quad \alpha_i \geq 0,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{11}$$

通过求解式 (11),即在约束 $\alpha_i \geq 0$和 $\sum_{i=1}^{n} \alpha_i y_i = 0$ 下,求式 (11)极大值,便可求出 α_i 来. 由式 (9)可得,$\alpha_i \neq 0$所对应的样本点才对权向量 w_0 有贡献,这种样本点正好落在最优边界 OM 上,其所对应的向量 x_i 称为"支持向量 (Support Vector, SV)". 根据式 (1)、式 (2)及最优超平面 OSH的定义可得:

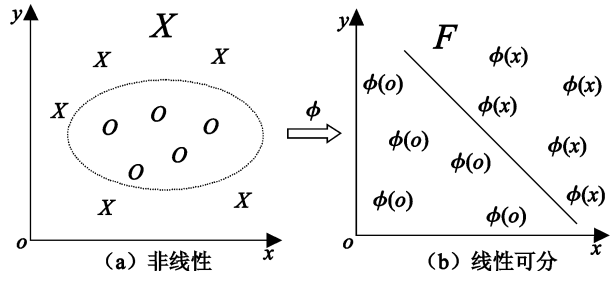
$$b_0 = -\frac{1}{2} (\min_{\{x|y=+1\}} w_0^T \cdot x + \min_{\{x|y=-1\}} w_0^T \cdot x). \tag{12}$$

因而分类决策函数 (decision function)即为:

$$f(x) = \text{sign}[w_0^T \cdot x + b_0] = \text{sign}[\sum_{i=1}^{n} \alpha_i y_i x_i^T \cdot x + b_0] = \text{sign}[\sum_{SV} \alpha_i y_i x_i^T \cdot x + b_0]. \tag{13}$$

1.2 非线性可分情形

由于自然界的许多实际问题往往是非线性的 (如图 2(a)),因此可将输入空间 X 中的输入向量 x 经一非线性变换 $(x \to z = \phi(x))$映射到另一高维特征空间 F 中,使之成为线性可分的样本 (图 2(b)).



(a) 非线性 (b) 线性可分

说明:第 1步:先通过转换函数 ϕ将训练数据非线性地投影到一高维特征空间;第 2步:构建将正负样本分离开的超平面.

图 2 SVM非线性分类原理示意图

在 F 空间中,$z = \phi(x)$为线性可分情形,1.1节中的推导完全适用,只须将 x 变换为 z 即可. 此时 (11)式变为:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (z_i^T \cdot z_j), \tag{14}$$

$$z_i^T \cdot z_j = \phi(x_j)^T \cdot \phi(x_j) = K(x_i, x_j). \tag{15}$$

K 是一个非负定对称函数,称为核函数 (Kernel func-

tion). K一般可取多种形式,如径向基函数,多项式函数等,经非线性变换后的二次规划问题变为求解约束 $\alpha_i \geq 0$和 $\sum_{i=1}^{n} \alpha_i y_i = 0$下极值问题:

$$\alpha = \arg\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j)\right]. \tag{16}$$

分类决策函数变为:

$$f(x) = \mathrm{sign}\left[\sum_{i\in\{SV\}}\alpha_i y_i K(x_i, x) + b_0\right], \tag{17}$$

其中:

$$b_0 = -\frac{1}{2}\left\{\min_{\{x|y_i=+1\}}\left(\sum_{j\in\{SV\}}\alpha_j y_j K(x_i, x_j)\right) + \max_{\{x|y_i=-1\}}\left(\sum_{j\in\{SV\}}\alpha_j y_j K(x_i, x_j)\right)\right\}. \tag{18}$$

1.3 支持向量机的算法及伪代码

为了求解特定条件下的最大值问题式(14),Lagrange目标函数可重构为:

$$L(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \lambda\sum_{i=1}^{n}\alpha_i y_i. \tag{19}$$

利用随机梯度上升算法:

$$\delta\alpha_k = \eta\frac{\partial L}{\partial\alpha_k} = \eta\left[1 - y_k\sum_{j=1}^{n}\alpha_j y_j K(x_i, x_k) - \lambda y_k\right], \tag{20}$$

可得:

$$\Delta L = L(\alpha_k + \delta\alpha_k) - L(\alpha_k) = \delta\alpha_k\left[1 - y_k\sum_{j=1}^{n}K(x_j, x_k) - \lambda y_k\right] - \frac{1}{2}(\delta\alpha_k)^2 L(x_k, x_k) = \frac{(\delta\alpha_k)^2}{\eta} - \frac{1}{2}(\delta\alpha_k)^2 K(x_k, x_k) = \left(\frac{1}{\eta} - \frac{K(x_k, x_k)}{2}\right)(\delta\alpha)^2. \tag{21}$$

给定 $\delta L > 0$,由方程(21)可以导出:

$$0 < \eta K(x_k, x_k) < 2, \tag{22}$$

即:

$$0 < \eta < \frac{2}{K(x_k, x_k)}. \tag{23}$$

对于高斯核函数,$K(x_k, x_k) \equiv 1$,所以 $0 < \eta < 2$ 对于多项式核函数,

$$0 < \eta < \frac{2}{\max\limits_{\{k=1,\cdots,n\}} K(x_k, x_k)}.$$

当 L 达到其最大值时,二次规划问题达到稳定解:$\delta\alpha_k = 0$. 所以方程(20)可为:

$$y_k\left[y_k - \sum_{j=1}^{n}\alpha_j y_j K(x_i, x_j) - \lambda\right] = 0. \tag{24}$$

基于上述分析,支持向量机的伪代码如下:

1) 初始化 α_i, λ;

2) For $k = 1, \cdots, ite_{\max}$执行第 3步至第 7步;

3) For $i = 1, \cdots, n$,执行第 4步至第 5步;

4) 定义并计算 q_i :$q_i = \sum_{j=1}^{n}\alpha_j y_j K(x_i, x_j)$;

5) 计算 α:

如果 $\alpha_i^t + \delta\alpha_i^t \leq 0$,则 $\alpha_i = 0$,

否则 $\alpha_i^t = \alpha_i^t + \delta\alpha_i^t$;

6) 计算 γ, λ:

$\gamma = (1/2)[\min_{\{i|y_i=+1\}}(q_i) - \max_{\{i|y_i=-1\}}(q_i)]$,

$\lambda = (1/2)[\min_{\{i|y_i=+1\}}(q_i) + \max_{\{i|y_i=-1\}}(q_i)]$;

7) 如果 $k = ite_{\max}$或者 γ达到 $\gamma_{\max}$,则终止循环,否则返回到第 2步,执行下一个 k值;

8) 执行后续计算.

根据 SVM的上述原理,构建了通用二类分类器 SVM*,并首次将它成功地应用于蛋白质功能家簇的系统分类与预测研究[3-5,11,13]. 文中应用 SVM*对多种车辆类型进行了识别研究.

2 结果分析

实验的车辆数据 Vehicle Dataset来源于公用数据库 UCI Repository[12]. 该数据库包含了 4种车型(分别为 Opel, Saab, Bus和 Van)的相关特征数据,共有 846个样本. 其中第 1类车型样本数为 212,第 2类车型样本数为 217,第 3类车型样本数为 218,第 4类车型样本数为 199. 该数据是利用照相机以 34.2 的俯角,分别在固定的水平高度移动照相机,并保持俯角不变,从不同的角度(0～360°)拍摄 4种车辆的侧面图像,然后分别以 37.5 和 30.8 的俯角重复上述实验,最后得到分辨率为 128×128像素、灰度级为 64的车辆侧面图像,通过采用分级图像处理系统(Hierarchical Image Processing System, HIPS)从车辆的侧面图像中提取出 18个特征值.

实验先后分别将 Bus, Van, Saab和 Opel作为正样本(Positive),其他的作为负样本(Negative),即:一对多(ONE-VS-ALL)多类分类方法,并采用 9次交叉验证法(9-Fold Cross-Validation)[14]进行训练和测试. 每次实验的训练准确率均为 100%,测试准确率结果见表 1. 测试准确率公式为:

$$Q_1 = \frac{(TP + TN)}{(TP + FN + TN + FP)}. \tag{25}$$

其中:TP(True Positive)代表在测试集中被准确判断为正样本的样本个数;TN(True Negative)代表在测试集中被准确判断为负样本的样本个数;FN(False Neg-

ative)代表在测试集中被错判为负样本的样本个数 *FP* (*False Positive*) 则代表在测试集中被错判为正样本的样本个数.

表 1 4种车型数据的 9次交叉验证结果 %

9次交叉验证	车型			
	Bus - P	Van - P	Saab - P	Opel - P
1	91.49	91.49	79.78	78.72
2	91.49	92.55	78.72	76.60
3	97.87	98.94	77.66	78.72
4	93.62	94.68	74.47	77.66
5	97.87	95.75	77.66	70.21
6	96.81	97.87	76.60	75.53
7	92.55	95.75	72.34	75.53
8	89.36	97.87	73.40	78.72
9	98.94	95.75	71.78	76.60
平均准确率	94.44	95.63	75.82	76.48

说明: Bus - P, Van - P, Saab - P, Opel - P分别代表该类车为正样本,其它车型为负样本, 1 - 9代表第 *j*次交叉验证, Average代表 9次测试结果的平均准确率.

总的测试准确率公式为:

$$Q = \sum_{i=1}^{4}\sum_{j=1}^{9}\frac{(TP + TN)_{i,j}}{(TP + FN + TN + FP)_{i,j}} \quad (26)$$

i, *j*分别表示第 *i*个正样本类和第 *j*次交叉验证.

从表 1可以看出,当 Bus和 Van这两类车作为正样本时,分类准确率比较高;当 Saab和 Opel这两类车作为正样本时,分类准确率比较低. 这说明利用 18个轮廓特征很难将 Saab和 Opel这两类车从其他类别中区别出来,同时说明了 Saab和 Opel两类车型的轮廓特征不太明显. King R. D. 等人[14]曾用不同的分类方法对本样本数据集进行了详细的分类研究,结果如表 2所示 (其中 SVM★为本实验的结果). 从表 2可以看出,除最近邻分类方法 (KNN)和 SVM外,其他方法训练准确率都未达到 100%; SVM★对测试集的测试准确率 *Q*最高,达到了 85.59%.

表 2 不同方法的训练准确率和测试准确率 %

名称	训练准确率	测试准确率
SVM★	100.00	85.59
Quadra	91.50	85.00
Alloc80	—	82.70
LogReg	83.30	80.90
BackProp	83.20	79.30
Discrim	79.80	78.40
Smart	93.80	78.30
C4.5	93.50	73.40
KNN	100.00	72.50
Cart	—	71.60
Bayes	48.10	44.20

3 结 语

应用基于随机梯度上升算法构建的支持向量机程序 SVM★对公用数据库 UCI Repository中的车辆数据 (Vehicle Dataset)样本进行了分类研究. 从本实验可以看出:支持向量机的 9次交叉验证的训练准确率和测试准确率均高于其它分类方法. 合理的提取特征向量对于提高分类准确率至关重要. 因此,要将多类车型较有效地进行分类并达到实用效果,必须改进特征向量的提取方法.

参考文献:

[1] 王运琼, 游志胜, 刘直芳. 利用支持向量机识别汽车颜色[J]. 计算机辅助设计与图形学学报, 2004, 16(5): 701 - 706.

[2] VAPNIK V. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995.

[3] CAI C Z, HAN L Y, JI Z L, et al. SVM-prot: Web-based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence[J]. Nucleic Acids Res, 2003, 31(13): 3 692 - 3 697.

[4] CAI C Z, HAN L Y, JI Z L, CHEN Y Z. Enzyme Family Classification by Support Vector Machines[J]. Proteins, 2004, 55(1): 66 - 76.

[5] CAI C Z, WANG W L, SUN L Z, CHEN Y Z. Protein Function Prediction Via Support Vector Machine Approach[J]. Mathematical Biosciences, 2003, 185: 111 - 122.

[6] HAN L Y, CAI C Z, LO S L, CHUNG M C M, CHEN Y Z. Prediction of RNA-binding Proteins from Primary Sequence by a Support Vector Machine Approach[J]. RNA, 2004, 10: 355 - 368.

[7] HAN L Y, CAI C Z, JI Z L, CAO Z W, CUI J, CHEN Y Z. Predicting Functional Family of Novel Enzymes Irrespective of Sequence Similarity: a Statistical Learning Approach[J]. Nucleic Acids Res, 2004, 32: 6 437 - 6 444.

[8] HAN L Y, CAI C Z, JI Z L, CHEN Y Z. Prediction of Functional Class of Novel Viral Proteins by a Statistical Learning Method Irrespective of Sequence Similarity[J]. Virology, 2005, 331: 136 - 143.

[9] VAPNIK V N. Statistical Learning Theory[M]. New York: Wiley, 1998.

[10] SCHOLKOPF B, BURGES C J C, SMOLA A J. Advances in Kernel Methods -Support Vector Learning[M]. Cambridge, Mass: MIT Press, 1999.

[11] CAI C Z, WANG W L, CHEN Y Z. Support Vector Machine Classification of Physical and Biological Datasets[J]. Inter J Modern Phys C, 2003, 14(5): 575 - 585.

[12] BLAKE C L, MERZ C J. UCI Repository of Machine Learning Databases[EB/OL]. Univ of California, Dept Information and Computer Science, Irvine, CA, USA, 1980. http://www.ics.uci.edu/mlearn/MLRepository.html, 2005 - 01 - 01.

[13] CAI C Z, HAN L Y, CHEN X, et al. Prediction of Functional Class of the SARS Coronavirus Proteins by a Statistical Learning Method[J]. J Proteome Res, 2005, 4(5): 1 855 - 1 862.

[14] KING R D, FENG C, SUTHERLAND A. StatLog: Comparison of Classification Algorithms on Large Real-world Problems[J]. Applied Artificial Intelligence, 1995, 9(3): 289 - 334.

Vehicle Type Recognition by Using Support Vector Machine SVM★

XIAO Han-guang[1,2], CAI Cong-zhong[1,2], WANG Wan-lu[1]

(1. Department of Applied Physics, Chongqing University, Chongqing 400030, China;
2. Department of Computational Science, National University of Singapore, Singapore 117543)

Abstract: The Support Vector Machine (SVM) has shown excellent learning and generalization ability in the practice problems of binary classification, and has been widely employed in multi-class classification. Based on the framework features of the vehicles, the SVM★ is used to classify 4 types of vehicles. The results of the SVM★ are compared with that of different classifiers. The testing accuracy to this vehicle dataset reaches 85.59% by means of 9-fold cross-validation which demonstrates that the classification performance of SVM★ is superior to those of other classifiers.

Key words: support vector machine (SVM); vehicle recognition; framework feature

(编辑　陈移峰)

(上接第 60页)

Image Recognising Based on Support Vector Machine

HE Jiang-ping[1], WEN Jun-hao[2], DENG Tian-jie[3], WANG Dao-qian[2]

(1. School of Mathematical Sciences, Chongqing Institute of Technology, Chongqing 400050, China;
2. School of Software Engineer, Chongqing University, Chongqing 400030, China;
3. Software Institute, Nanjing University, Nanjing Jiangsu 210093, China)

Abstract: Support vector machine is the uniform method of statistical learning method, and has become more and more popular in research field. Support vector machine has achieved excellence in pattern recognition and text classification for its high performance in veracity. Support vector machine method was used to process several Binary image and Gray scale image and got a good statistical result. A strategy to select feature vectors has been found out, and the criterion to judge the results are discussed. Compared with other methods, the important conclusion was draw: it performs perfectly in statistic when using support vector machine to detect the image edge.

Key words: support vector machine; image recognising; edge detection; statistical learning; digital image processing

(编辑　张　苹)