文章编号：2095-2163(2023)02-0161-04 中图分类号：TP301 文献标志码：A

HOG 特征值的笔迹鉴定算法

杨 东，王以松
（贵州大学 计算机科学与技术学院，贵阳 550025）

摘 要：机器学习作为人工智能的一个分支，在工程实践中已经产生了较大的经济价值和科技价值。机器学习创建基于样本的数学模型，通过训练预测或者作出决策解决人工智能中的问题。支持向量机是按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。本文通过提取笔迹中的 HOG 特征，再利用支持向量机对该特征值进行训练，得到笔迹鉴定的模型，并通过该模型鉴定笔迹。

关键词：机器学习；支持向量机；笔迹鉴定；HOG 特征

Handwriting identification based on HOG feature values

YANG Dong, WANG Yisong
(College of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

【Abstract】 Machine learning, a branch of artificial intelligence, has generated a great economic and scientific value in engineering practice. Machine learning solves problems in artificial intelligence by creating sample-based mathematical models that are trained to predict or make decisions. Support Vector Machine (SVM) is a class of generalized linear classifier that performs binary classification of data in a supervised learning manner, where the decision boundary is a maximum margin hyperplane solved for the learned samples. In this paper, we extract the HOG features in handwriting and then use the feature values for training using SVM, thus obtaining a model for handwriting identification. The experimental results verify its good effect.

【Key words】 machine learning; SVM; handwriting identification; HOG feature

0 引 言

统计学习理论(Statistical Learning Theory, SLT)是在 20 世纪 60 年代由 Vapnik 等人提出并于 20 世纪 90 年代中期建立的一种针对小样本情况研究统计学习规律的理论，由于解决了机器学习中小样本、过学习、欠学习和局部最小点等实际问题，从而成为 20 世纪 90 年代末发展最快的研究方向之一，其核心思想是学习机器要与有限样本相适应，从而实现最佳推广能力[1-2]。统计学习理论为小规模样本的机器学习问题提供了良好的理论框架，机器学习中的支持向量机(SVM)是最重要的算法之一，在诸多领域得到了运用[3]。郭辉[4]等人提出一种改进鲸鱼优化算法，同步优化 SVM 的特征选择模型，该算法的原理是利用 Levy 飞行策略对鲸鱼优化算法的螺旋更新位置进行变异扰动，改进了单纯形策略中的反射操作对种群中的精英个体，标准函数的测试结果证明，其改进能有效提高算法的收敛速度和计算精度。利用 SVM 核参数和特征选择目标作为共同优化对象，对 UCI 标准数据集和真实乳腺癌数据集进行特征选择仿真实验，真实乳腺癌数据集上的分类精度与传统支持向量机相比提高了 11.053%。

方向梯度直方图(Histogram of Oriented Gradient, HOG)特征是在计算机视觉中用来对图片进行处理的一种特征描述子。HOG 特征通过计算和统计图像局部区域的梯度方向直方图来构成特征[5]。韩松来[6]等人提出主成分分析(Principal Component Analysis, PCA)和 HOG 特征的遥感异源图像匹配算法，利用 HOG 提取图像间的几何结构共性特征，能有效克服异源图像非线性灰度畸变的问题，实现多种遥感异源图像匹配性能的明显提升；宋建辉[7]等采用融入 HOG 特征对 ResNet 残差模型进行改进，利用自裁残差(Cropping Inside Residual, CIR)模型塑造的孪生目标可以增强跟踪网络的骨干网络对图形几何变化的鲁棒性；高达义[8]通过注

作者简介：杨 东(1987-)，男，硕士研究生，主要研究方向：人工智能；王以松(1975-)，男，博士，教授，主要研究方向：人工智能。
通讯作者：王以松 Email：ys_wang168@ sina.com

收稿日期：2022-04-12

意力模型(Convolutional Block Attention Module, CBAM)调节网络上下文信息的 HOG 特征比例,使网络中各特征图发挥出最好的效果。实验结果表明,该算法在 OTB100 的精确率和成功率分别达到 81.9%和 60.6%。

笔迹鉴定是司法领域的一个重要鉴定手段,张伟[9]等通过一起刑事申诉中的笔迹鉴定案例,阐述了笔迹鉴定意见在刑事审判中的关键性作用;张乐[10]等对中国裁判文书网平台的笔记鉴定意见资料进行归纳和分类,进而讨论笔迹鉴定在司法案件的应用,增强了笔迹鉴定意见的公信力,从而最大限度发挥笔记鉴定意见在诉讼案件中的证据作用。

目前,深度学习、强化学习等模型训练呈现训练规模大、时间长,对计算资源要求高的问题,在实际使用过程应用较为困难。本文探索提取笔迹 HOG 特征对笔迹的 SVM 模型进行训练,对笔迹的真伪进行鉴定。实验结果表明,该模型在较小的样本和较短的训练时间、资源条件下,能成功鉴定伪造笔迹。

1 算法思想

SVM(支持向量机)是定义在特征空间上通过求解间隔最大的线性分类器的二分类模型。SVM 的求解问题可以形式化求解的凸二次规划问题中的间隔最大化,也等价于正则化的合页损失函数的最小化问题。SVM 的学习算法就是求解凸二次规划的最优化算法[8]。本文提取笔迹的 HOG 特征向量,利用 SVM 算法训练出笔迹特征的 HOG 特征向量超平面。

1.1 SVM 算法原理

SVM 的基本算法思想是能够对训练数据集正确划分并且求解几何间隔最大的分离超平面。如图 1 所示,$\omega \cdot x + b = 0$ 即为分离超平面,对于线性可分的数据集来说,虽然分离超平面可能有无穷多个,但是几何间隔最大的分离超平面却是唯一的。

几何间隔:对于给定的数据集 T 和超平面 $\omega * x + b = 0$,定义超平面关于样本点 (x_i, y_i) 的几何间隔,公式(1):

$$\gamma_i = y_i\left(\frac{\omega}{\|\omega\|} \cdot X_i + \frac{b}{\|\omega\|}\right) \tag{1}$$

SVM 模型的求解最大分割超平面通过求解二次规划问题得到最小化泛函,式(2):

$$\Phi(\omega) = \frac{1}{2}(\omega \cdot \omega) \tag{2}$$

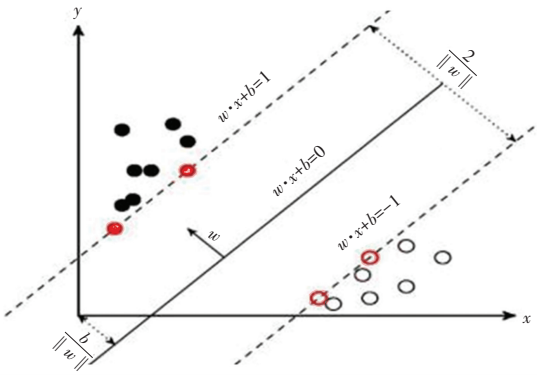

图 1 SVM 超平面图

Fig. 1 SVM hyperplane

约束条件为不等式类型:$y_i[(x_i \cdot \omega) - b] \geqslant 1$, $i = 1,2,3,\cdots,l$

这个优化问题的解是由拉格朗日泛函的鞍点给出的,式(3):

$$L(\omega,b,\alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{l}\alpha_i\{[(x_i \cdot \omega) - b]y_i - 1\} \tag{3}$$

其中,α_i 为拉格朗日乘子。

对拉格朗日函数关于 ω, b,求解最小值和关于 $\alpha_i > 0$,求其最大值。在鞍点上,解 ω_0, b_0 和 α_0 必须满足以下条件:$\frac{\partial(\omega_0,b_0,\alpha_0)}{\partial b} = 0$ 和 $\frac{\partial(\omega_0,b_0,\alpha_0)}{\partial \omega} = 0$。

最优超平面具有两个特性:

(1)系数 α_i 必须满足约束,式(4):

$$\sum_{i=0}^{l}\alpha_i y_i = 0, \alpha_i \geqslant 0, i = 1,2,\cdots,l \tag{4}$$

(2)最优超平面,向量 $\boldsymbol{\omega}_0$ 是训练集中的向量的线性组合,式(5):

$$\boldsymbol{\omega}_0 = \sum_{i=0}^{l} y_i\alpha_i x_i, \alpha_i \geqslant 0, i = 1,2,\cdots,l \tag{5}$$

只有所谓的支持向量可以在 $\boldsymbol{\omega}_0$ 的展开式中 j 具有非零的系数 α_i。不等式 $y_i[(x_i \cdot \omega) - b] \geqslant 1$, $i = 1,2,3,\cdots,l$ 在等号成立时得到支持向量,其最优超平面,式(6):

$$\boldsymbol{\omega}_0 = \sum_{\text{支持向量}} y_i\alpha_i x_i, \alpha_i \geqslant 0, i = 1,2,\cdots,l \tag{6}$$

1.2 HOG 特征向量的计算

纹理特征是图片的一种独特的特征。纹理作为反映图像同质现象的一种视觉特征,当物体表面具有缓慢变化或者周期性变化时,纹理就会体现出相应的变化。纹理的 3 大标志:非随机排列、某种局部序列性不断重复以及纹理区域内大致为均匀的统一

体。纹理特征包含统计型纹理特征和模型纹理特征。统计型纹理特征从像元及其邻域内的灰度属性出发，研究纹理区域中的统计特征；模型纹理特征是假设图片纹理是某种参数控制的分布模型的形式，再以纹理图像的实现来估计计算模型参数[5]。HOG 特征是描述局部纹理的有效特征向量 ，计算笔迹图片的 HOG 特征的方法为：将一张笔迹图片按照 8×8 像素的大小分割为若干个细胞单元，每个细胞单元中梯度角度的取值范围介于 0~180°之间，将角度范围分成 9 份，每 20°为一个组；在细胞单元中，对内部所有的像素的梯度进行统计；将每一组中所有像素对应的梯度值进行累加，可以得到 9 个数值。直方图就是由这 9 个数值组成的数组，每个细胞单元就会得到一个 9 维的特征向量，特征向量每一维对应的值是累加的梯度幅值。在获得每个细胞单元的梯度方向直方图后，再对细胞单元组合形成区块。在笔迹图片中，选取 2×2 个细胞单元作为一个区块，每次滑动 8 个像素得到一个新的区块，按照此步骤循环结束后得到整个笔迹图片的 HOG 特征向量。结合 SVM 的算法，可求解出笔迹图片中的 HOG 特征向量超平面。

2　实验步骤

本文的实验包括 3 个步骤：分别是图像预处理、计算梯度直方图并进行池化处理并训练模型，通过测试样本对模型进行检测。

（1）图像预处理包括伽马校正和灰度化。使用伽马校正减少光度对实验的影响。灰度化是将彩色图片变成灰度图，降低图片处理的计算量。考虑到本实验的笔迹均为黑白色，因此采用灰度图来处理。笔迹图片预处理后获取外部纹理特征如图 2 所示。

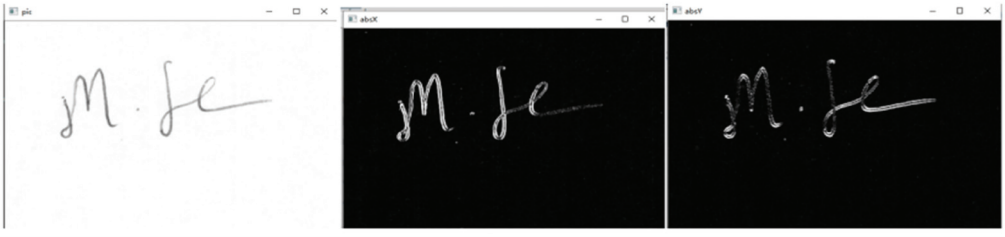

图 2　笔迹图片预处理

Fig. 2　Handwriting image preprocessing

（2）提取笔迹图片的 HOG 特征值，再对 HOG 特征值矩阵进行平均值池化处理，最终得到笔迹图片的 HOG 特征向量，如图 3 所示。

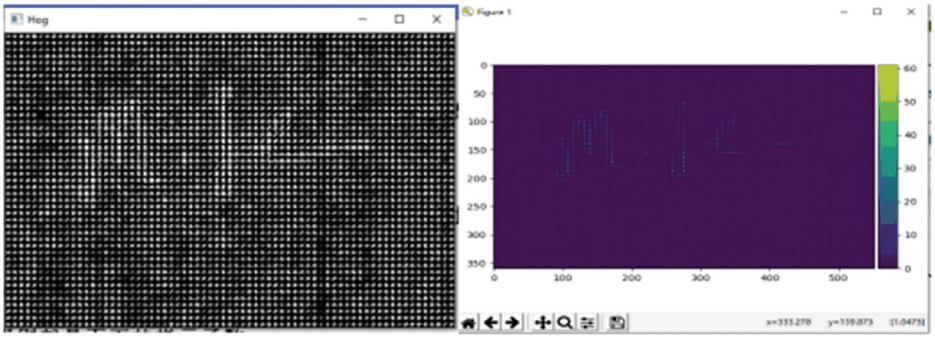

图 3　笔迹图片的 HOG 特征向量

Fig. 3　HOG feature vector of handwriting image

（3）通过测试样本对 SVM 模型的训练结果进行验证，如用仿冒的笔迹进行验证，则系统判定出属于仿冒的笔迹，且给出其属于仿冒笔迹的相应的概率值。两组数据分为训练数据组和测试数据组，而测试数据组则是用来检验模型是否可以对笔迹进行鉴别。通过人工标注方式对数据进行标注：本人笔迹标记 label 为 1，仿冒笔迹标记为 0。

3　实验结果

为了评估实验训练模型的有效性，本文采用数据集 CEDAR 的部分英文笔迹作为实验对象进行鉴定。采用人工标注数据，真迹标签值为 1，仿冒标签值为 0。利用 HOG 特征值训练的 SVM 模型鉴定笔迹的鉴别结果见表 1。

表 1 HOG 笔迹特征模型验证统计表

Tab. 1 HOG handwriting feature model validation statistics table

标识	笔迹类型	判断概率	判断是否正确
[1]	original writing	0.604 115 25	错误
[1]	original writing	0.397 693 24	错误
[1]	original writing	0.400 313 22	错误
[0]	forge writing	0.693 873 66	正确
[0]	forge writing	0.808 294 28	正确
[0]	forge writing	0.927 800 52	正确
[0]	forge writing	0.848 881 62	正确
[0]	forge writing	0.873 913 93	正确
[0]	forge writing	0.885 834 99	正确
[0]	forge writing	0.798 513 69	正确
[0]	forge writing	0.891 944 56	正确
[0]	forge writing	0.832 572 69	正确
[0]	forge writing	0.900 081 41	正确
[0]	forge writing	0.725 670 61	正确
[0]	forge writing	0.846 654 11	正确
[0]	forge writing	0.916 083 94	正确
[0]	forge writing	0.736 325 48	正确
[0]	forge writing	0.730 263 74	正确
[1]	original writing	0.494 431 95	错误
[1]	original writing	0.469 890 96	错误
[1]	original writing	0.461 088 81	错误
[1]	original writing	0.510 670 97	错误
[1]	original writing	0.470 710 05	错误

本实验实现了仿冒笔迹的基本鉴别。总体的 23 个样本中,有 15 个样本被成功鉴别,识别成功率达 65.21%。本实验在较短时间内和较低的计算资源下基本鉴定了笔迹的真伪,在要求快速响应的机器学习应用场景下有一定的参考价值。

4 结束语

针对深度学习在训练过程中资源要求过高的问题,本文提出了针对笔迹 HOG 特征值训练 SVM 模型的方法,该模型可以对笔迹进行基本的鉴别,可以通过少数的样本特征和较少的计算资源即可训练出模型。但是,由于 SVM 在多分类场景时将耗费大量的计算资源和时间,本文提出的 SVM 算法不适用于多分类的场景。

参考文献

[1] 张学工. 统计学习理论的本质[M]. 北京:清华大学出版社,2000:1-11.

[2] 张植明. 双重随机样本的结构风险最小化原则[J]. 计算机工程与应用,2009,45(1):51-55.

[3] 周志华. 机器学习[M]. 北京:清华大学出版社,2016-01.

[4] 郭辉,付接递,李振东,等. 基于改进鲸鱼算法优化 SVM 参数和特征选择[J/OL]. 吉林大学学报(工学版):1-22[2022-04-08].

[5] RAFAEL C. Gonzalez / Richard E.Woods.数字图像处理(第三版)[M]. 北京:电子工业出版社,2011.

[6] 韩松来,王钰婕,王星,等. 多尺度 PCA-HOG 遥感异源图像匹配算法[J]. 国防科技大学学报,2022,44(1):146-155.

[7] 宋建辉,孙晓南,刘晓阳,等. 融合 HOG 特征和注意力模型的孪生目标跟踪算法[J/OL]. 控制与决策:1-9[2022-04-08].

[8] 高达义. 基于局部特征的人脸表情识别算法[D]. 南京邮电大学,2021.

[9] 张伟. 浅析笔迹鉴定在刑事申诉案件中的作用[J]. 法制博览,2020(15):150-151.

[10] 张乐,羡闻鹏. 笔迹鉴定意见证据应用研究[J]. 法制与社会,2020(16):86,91.

(上接第 160 页)

[2] NELSON B,OLOVSSON T. Security and privacy for big data: A systematic literature review [C]//2016 IEEE international conference on big data (big data). IEEE, 2016: 3693-3702.

[3] ZHOU S G. Privacy Preservation in Database Applications: A Survey[J]. Chinese Journal of Computers, 2009, 32(5):847-861.

[4] MIN Z, YANG G,SANGAIAH A K, et al. A privacy protection-oriented parallel fully homomorphic encryption algorithm in cyber physical systems [J]. EURASIP Journal on Wireless Communications and Networking, 2019, 2019(1): 1-14.

[5] TEMUUJIN O, AHN J, IM D H. Efficient L-diversity Algorithm for Preserving Privacy of Dynamically Published Datasets [J]. IEEE Access, 2019, PP(99):1.

[6] DWORK C. Differential privacy: A survey of results [C]// International conference on theory and applications of models of computation. Springer, Berlin, Heidelberg, 2008: 1-19.

[7] FANG Y, ZHU J, ZHOU W. A Surveyon Data Mining Privacy Protection Algorithms[J]. Netinfo Security, 2017, 2:6-11.

[8] XIONG J, REN J, CHEN L, et al. Enhancing privacy and availability for data clustering in intelligent electrical service of IoT [J]. IEEE Internet of Things Journal, 2018, 6(2): 1530-1540.

[9] 傅彦铭,李振铎. 基于拉普拉斯机制的差分隐私保护 k-means++聚类算法研究[J]. 信息网络安全,2019(2):43-52.

[10] 李洪成,吴晓平,陈燕. MapReduce 框架下支持差分隐私保护的 k-means 聚类方法 [J]. 通信学报,2016,37(2):124-130.

[11] 马银方,张琳. 基于差分隐私保护的 KDCK-medoids 动态聚类算法[J]. 计算机科学,2016,43(S2):368-372.

[12] DWORK C, ROTHBLUM G N, VADHAN S P. Boosting and Differential Privacy [C]// 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA. IEEE, 2010.