针对图像来源鉴别中支持向量机的研究

黄 曜[1],许华虎[2],欧阳杰臣[1],高 珏[3]
(1. 上海大学 计算机工程与科学学院,上海 200444;
2. 上海上大海润信息系统有限公司,上海 200444;
3. 上海大学 计算中心,上海 200444)

摘 要:随着数码图像的普及,图像盲取证成为时下的研究热点之一,如何识别图像来源是其主要的研究内容。作为图像来源鉴别最关键的阶段,构造鉴别的支持向量机(SVM)分类模型直接影响最终的鉴别率。由于不同核函数以及核参数对分类器性能有着相异的影响,故分析对比了各种核函数,然后选取了细分效果更好的高斯径向基函数作为核函数。针对核参数选择问题,分析了各种核参数寻优算法,并通过实验验证了各个算法的效果,以及最终构造的分类模型的效果。实验结果表明,选用高斯径向基函数作为核函数,利用粒子群算法选出的核参数所构造的分类模型取得了最好的图像来源鉴别率。

关键词:图像盲取证;支持向量机分类模型;核函数;核参数;图像来源鉴别率
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2016)10-0001-05
doi:10.3969/j.issn.1673-629X.2016.10.001

Research on Support Vector Machines for Image Source Identification

HUANG Yao[1], XU Hua-hu[2], OUYANG Jie-chen[1], GAO Jue[3]
(1. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;
2. Shang Da Hai Run Information System Co., Ltd., Shanghai 200444, China;
3. Computer Center of Shanghai University, Shanghai 200444, China)

Abstract: With the popularity of digital images, blind image forensics has become one of the hotspots nowadays. The main research content of blind image forensics is how to identify the image source. As the most critical stage of image source identification, the SVM classification model for identification directly affects the final identification rate. Because the different kernel function and kernel parameters has distinct effect on the performance of the classification model, the various kernel functions are analyzed and compared, then the Gaussian radial basis function with better subdivision is selected as the kernel function. In view of the kernel parameter selection, the various kernel parameter optimization algorithms are analyzed, and the effectiveness of each algorithm and the effect of the final classification model by experiments are verified. The results show that choosing Gaussian radial basis function as the kernel function, using the kernel parameters selected by particle swarm algorithm to construct the classification model will achieve the best image source identification rate.

Key words: blind image forensics; SVM classification model; kernel function; kernel parameter; image source identification rate

0 引 言

随着现代数字技术的发展以及数码相机的普及,数字图像在日常生活和工作中得到了广泛应用。相应地,篡改图像内容并使得人眼难以觉察出伪造的痕迹变得越来越频繁,由此带来的影响轻则干扰人们的正常生活,重则影响国家、社会和政治稳定[1]。因此,鉴别图像的真实性显得日益迫切,图像盲取证技术作为研究要点被提及并成为时下热点之一。

图像盲取证技术主要涉及四个方面的问题[2-3],其中之一便是如何确认图片是由相机、手机等设备所拍摄的自然图像,还是经过计算机制作的图像,亦或是扫描仪直接扫描生成的图像。传统的图像来源鉴别算法主要包括特征提取、特征选择以及构造分类器等多项技术。构造分类器作为整个算法流程最后也是最重

收稿日期:2015-12-16 **修回日期**:2016-04-08 **网络出版时间**:2016-08-23
基金项目:上海张江国家自主创新示范区专项发展资金重点项目(一期)(201411-ZB-B204-012)
作者简介:黄 曜(1991-),男,硕士,研究方向为图像多媒体技术;许华虎,教授,博士生导师,CCF 高级会员,研究方向为人机交互、图像处理、多媒体网络技术等。

要的一环,直接关系到最终的鉴别效果。然而,现有的图像来源鉴别算法大多只是将现成的分类模型投入鉴别使用,例如 LIBSVM[4] 默认的分类模型提供了一些基本参数。但是,这些现有的分类模型是否适用于图像来源鉴别并没有得到实际的验证。但众所周知,SVM 的一大优点是通过引入核函数,将输入的特征空间中的线性不可分问题转化为高维空间中的线性可分问题[5-6],所以核函数自然是影响分类器效果的一大因素。核函数的种类颇多,选用何种核函数是构建适合的 SVM 模型的必经之路。另外,研究表明,在确定核函数后,选择适合的误差惩罚因子 C 和核参数 σ 对分类器的性能影响甚至比选择一个适合的核函数更大。所以,有可能现有的 SVM 的相应参数并不适用于图像来源鉴别问题,然而这些参数深深影响着图像来源鉴别效果,所以研究改进支持向量机对于图像来源鉴别问题是十分必要的。

分类器的性能深深影响着图像来源鉴别的正确率,文中针对图像来源鉴别中支持向量机的性能进行了研究,特别是针对核函数、误差惩罚因子与核参数选择给出了研究结果。

1 图像来源鉴别中 SVM 核函数的选择

1.1 常用核函数

如前文所述,SVM 的核心思想在于通过核函数将低维的线性不可分问题转换为高维的线性可分问题,高维空间的内积运算因此可转化为核函数的运算。不同的内积运算形成不同的核函数,这意味着特征在其他的核函数下无法保持,所以选择合适的核函数对于支持向量机的应用至关重要。

根据统计学理论,如果一个函数满足 Mercer 条件,则就可以将之作为 SVM 的核函数。所以,从 Mercer 定理出发,可以明确核函数需要满足的条件。

Mercer 定理:令 Ω 是有限维欧氏空间中的有界闭集,并设 K 是连续对称函数,则存在积分算子 T_k:$L^2(\Omega) \to L^2(\Omega)$,使得 $(T_k f)(\bullet)$ 是正的。

$$(T_k f)(\bullet) = \int_{\Omega} K(\bullet, x) f(x) \mathrm{d}x \tag{1}$$

则对于任意的 $f \in L^2(\Omega)$,可以得到

$$\int_{\Omega \times \Omega} K(z,x) f(z) f(x) \mathrm{d}x \mathrm{d}z \geqslant 0 \tag{2}$$

其中,函数 $K(x_i, y_i)$ 就是核函数。由此可见,Mercer 定理很好地将核函数的性质表现了出来。即核函数可以将非线性样本转换为线性样本,避免增加问题的复杂性。

目前常见的核函数有以下几种:

线性核函数:

$$K(x_i, x_j) = (x_i \cdot x_j) \tag{3}$$

线性核函数在核函数里面使用频率相对较低,主要是因为它针对的是在低维空间可分的样本,这样就可以直接在低维空间进行分类,而不需转换到高维空间。但其实大部分的样本在低维空间都是线性不可分的,这样线性核函数就失去了意义。

多项式核函数:

$$K(x, x_i) = [(x \cdot x_i) + 1]^q \tag{4}$$

由多项式核函数可以得到 q 阶多项式分类器,q 代表了核函数的维数,q 越大,映射函数的维数越高,意味着样本更容易被分类,但计算复杂度也会相应增大。

高斯径向基函数(RBF):

$$K(x, x_i) = \mathrm{Exp}\left(-\frac{|x - x_i|^2}{\sigma^2}\right) \tag{5}$$

其中,σ 可看作高斯径向基函数的作用范围,由高斯径向基函数可得到高斯径向基函数分类器。

Sigmoid 函数:

$$K(x, x_i) = \tanh[v(x \cdot x_i) + c] \tag{6}$$

其中

$$\tanh(x) = [1 - \exp(-2x)]/[1 + \exp(-2x)] \tag{7}$$

由该式可得到带隐层的多层感知器网络。

1.2 核函数的确定

对于多项式核函数来说,因为属于全局核函数,所以相对位置相差很远的样本点都能对分类器产生影响。越复杂的多项式分类器分类效果越好,但随之而来的是计算复杂度的增加以及对新样本分类效果较差的问题;对于高斯径向基函数来说,它的局部性非常好,对于相对位置比较近的样本点也可以有较好的细分效果。但是当参数 σ 越小,该函数的推广能力越低,全局性相对较差;对于 Sigmoid 函数来说,它在神经网络中使用较为广泛,在 SVM 中的性能还没有得到充分的证明,只是理论可行,并且因为一定要满足一定的条件,所以实际应用也偏少。

综合以上分析并根据行业研究经验,文中选用高斯径向基函数作为 SVM 的核函数。这主要是基于该函数首先收敛域较宽,对样本点有较好的细分效果的优点。其次,它的实际应用非常广泛,性能得到了充分证明,是目前使用最多且表现相对优异的核函数。

2 图像来源鉴别中 SVM 核参数选择

为 SVM 选择一个高效的核函数固然重要,可是 Vanpik 等通过研究发现[7],相比核函数,不同的核参数以及惩罚因子产生的效果区分度更明显。所以,选择适合的核参数以及惩罚因子对 SVM 性能的影响更

显著。

在第一节的分析中,文中选择有着良好效果的高斯径向基函数,所以本节主要针对参数 σ 以及惩罚因子 C 的选择进行分析。参数 σ 主要用来控制高斯分布的距离。如果 σ 的值过小,甚至小于样本点之间最小相对距离时,所有的样本点都将成为支持向量,这将直接导致分类器对新样本的分类效果不理想,即“过拟合”现象;如果 σ 的值过大,甚至大于样本点之间最大相对距离时,分类器将完全没有分类能力。惩罚因子 C 表示对错分样本偏离值的惩罚系数,通过调节数据子空间中学习机器的置信区间范围,对其推广性产生影响。如果 C 的值越大,类的相对距离越小,分类器泛化能力越低,性能提高;如果 C 的值越小,类的相对距离越大,分类器的泛化能力越高,性能降低。

综合以上分析,无论是参数 σ 还是惩罚因子 C,过大或过小都会影响 SVM 的性能,特别是对于惩罚因子,要综合考虑 SVM 的性能与泛化能力。所以,选择适合的 σ 值与惩罚因子 C 的值至关重要。

常用核参数选择算法如下所述。

2.1　交叉验证法

机器学习的大意即是通过已知样本对待测样本进行预测。交叉验证法[8]的主要思想是将已知的部分样本集作为训练集训练模型,剩下的部分样本集作为测试集验证模型。它是用来验证分类器性能的一种统计分析方法。它以分类器的分类准确率来评价分类性能。具体实施办法如下:

1)按照一定规则将原始数据样本进行分组,一部分作为训练集,另一部分作为验证集;

2)利用训练集对分类器进行训练,再利用验证集来测试训练得到的模型,计算分类准确率。

正因为交叉验证法不仅可以有效地避免过学习和欠学习状态的发生,还能在做到良好的参数估计的同时,避免较高的计算复杂度,所以交叉验证法是统计学中一种著名的方法,并得到了广泛应用。发展到后来,产生了 K 折交叉验证法,它的主要思想是将样本集分为 K 组子集,将其中 $K-1$ 组子集作为训练集训练模型,再用剩下一组子集作为测试集验证模型的精度。再用另外一组子集作为测试集,剩下 $K-1$ 组子集作为训练集,这样依次调换测试集 $K-1$ 次,直到每组子集均作为测试集验证过模型的精度。最后再选择一组最优参数作为模型参数。由于经过了 K 次平均化的计算,该交叉操作避免了分类器中过学习或欠学习状态的发生,有一定的实用性。

2.2　网格搜索法

网格搜索法[9]是一种典型的试凑方法。主要思想是直接将一组组作为核参数,并求相应的分类函数,再根据分类模型的性能和经验调整参数值。由此可见,在网格搜索法最初,需要给出参数的取值范围,也可以理解为参数值的调整区间,最优解通常一定在这区间内产生。

使用网格搜索法确定核参数的步骤大致如下:

选定惩罚因子 C 与核参数 σ 的取值范围,一般遵从 $C \in (2^{-5},2^{-3},\cdots,2^{15})$,$1/\sigma^2 \in (2^{-15},2^{-13},\cdots,2^{3})$ 的原则。

设置搜索步长为 1,在以 C、σ 为横纵坐标的坐标系上构建一个二维网络,每个坐标点代表一个潜在解,可以用上文提到的 K 折交叉验证法计算各个参数预测准确率的均值,最后确定最佳解。

为了使结果更加精确,可进一步做更细致的网格搜索。将搜索步长减小为 0.1 进行二次搜索。

2.3　群智能法

交叉验证法以及网格搜索法虽能取得一定效果,但还是存在精度偏低的缺点。针对这些缺点,精确度更高而又更高效的群智能法应运而生,并且在 SVM 核参数的选择中取得了良好的效果。常见的群智能法包括遗传算法[10]、粒子群算法[11]、蚁群算法[12]、蛙跳算法[13]等。

1)遗传算法。

遗传算法是一种迭代算法,兼具繁衍、监测和评价的特性。每个个体在种群演化过程中都被评价优劣并得到其适应度值,个体在选择、交叉以及变异算子的作用下向更高的适应度进化,以达到寻求问题最优解的目标[14-15]。

遗传算法的大致步骤如下:

(1)初始化设置进化代数计数器 t;设置最大进化代数 T;随机生成 N 个个体作为初始种群 $p(t)$;

(2)通过个体评价计算种群 $p(t)$ 中各个个体的适应度;

(3)选择运算将选择算子作用于种群;

(4)交叉运算将交叉算子作用于种群;

(5)变异运算将变异算子作用于种群,种群 $p(t)$ 经过选择、交叉、变异运算后可得到下一代种群 $p(t+1)$;

(6)终止条件判断,若 $t < T$,则 $t = t + 1$,转到(2);若 $t > T$,则以进化过程中所得到的具有最大适应度的个体作为最优解输出,终止运算。

遗传算法的优点体现在不易表现为局部最优,但同时该方法受初值的影响较大,且对于不同的情况,需要重新设计相应的选择算子、交叉算子以及变异算子。

2)粒子群算法。

相对遗传算法来说,粒子群算法参数更少,操作更简便,整个流程更容易理解,所以在许多问题中得到了

更广泛的应用。算法的主要思想是将粒子经历过的最好位置记录下来并作为粒子最优解,也称作局部极值 pbest ,将整个群体经历过的最好位置记录下来并作为群体的最优解,也称作全局极值 gbest 。粒子通过这两个值调整飞行,最终产生新粒子。

粒子群算法的大致步骤如下:

(1)初始化一个种群规模为 N 的粒子群,在允许的范围内随机设定每个粒子的初始位置和初始速度,并把每个粒子的局部极值 pbest 设定为其初始位置,把 pbest 中的最好值赋给全局极值 gbest 。

(2)根据适应度函数计算每个粒子的适应值。

(3)将每个粒子的适应值与相应的 pbest 进行比较,若优于 pbest ,则将其作为新的 pbest 。

(4)将每个粒子的适应值与 gbest 进行比较,若优于 gbest ,则将其作为新的 gbest 。

(5)更新粒子的速度和位置。

(6)检验是否满足终止条件(达到最大迭代次数或最小适应度阈值),若是,则输出最优解,否则返回第(2)步。

3)蚁群算法、蛙跳算法。

蚁群算法以及蛙跳算法都存在算法收敛速度慢的问题,所以相对来说应用并没有遗传算法以及粒子群算法广泛。

3　核参数选择实验

3.1　实验步骤

3.1.1　数据预处理

因为文中最终的目的是验证各个核参数分类模型对最终鉴别效果的影响,所以选取事先已知成像设备的图像作为实验的源数据。然而确定核参数是一个复杂的过程,所以实验以验证各个核参数选择算法的效果为主。

为了研究的延续性,拟采用之前的研究成果,提取图像的混合特征作为实验数据。有针对性地将实验数据分为以下几类:

(a)自然图像与计算机生成图像类;

(b)自然图像与扫描仪生成图像类;

(c)计算机生成图像与扫描仪生成图像类;

(d)自然图像、计算机生成图像与扫描仪生成图像类。

理所应当地,再将每类数据分为训练与测试两组。

3.1.2　具体算法

(1)核参数选择算法。

按照前文所述,实验采用4种算法:K 折交叉验证法、网格搜索法、遗传算法、粒子群算法。

(2)理论方法。

将训练组数据通过各个核参数选择算法得到最优的核参数,再根据最优核参数对样本重新训练得到训练模型,再利用测试组数据对该模型的鉴别效果进行验证。

3.2　实验结果与分析

(1) K 折交叉验证法实验效果。

使用交叉验证法进行实验时,将 K 设置为20,由该算法得到的实验结果见表1。

表1　交叉验证法实验结果

数据类别	训练样本数	测试样本数	最优核参数 σ	最优惩罚因子 C	鉴别率/%
(a)	578	1 126	1.75	9	81.36
(b)	476	912	1.325	26	73.24
(c)	502	843	0.35	6	78.75
(d)	521	932	0.655	3	71.87

(2)网格搜索法实验效果。

使用网格搜索法时, $C \in (2^{-5},2^{-3},\cdots,2^{15})$, $1/\sigma^2 \in (2^{-15},2^{-13},\cdots,2^{3})$ 。先令步长为1进行粗网格搜索,搜索完毕后再令步长为0.1进行细网格搜索,得到最优核参数对。使用网格搜索法得到的实验结果见表2。

表2　网格搜索法实验结果

数据类别	训练样本数	测试样本数	最优核参数 σ	最优惩罚因子 C	鉴别率/%
(a)	578	1 126	2.55	8	82.88
(b)	476	912	0.215	72	72.83
(c)	502	843	0.45	9	81.63
(d)	521	932	0.025	2	72.48

(3)遗传算法实验效果。

使用遗传算法时,将进化代数设置为200,个体个数设为30,交叉率设置为0.8,变异率设置为0.15。使用遗传算法得到的实验结果见表3。

表3　遗传算法实验结果

数据类别	训练样本数	测试样本数	最优核参数 σ	最优惩罚因子 C	鉴别率/%
(a)	578	1 126	0.78	529.53	83.64
(b)	476	912	0.89	678.58	74.38
(c)	502	843	0.23	325.26	81.78
(d)	521	932	0.14	852.76	73.59

(4)粒子群算法实验效果。

为了对比粒子群算法与遗传算法的效果,使用粒子群算法时,将最大进化代数也设置为200,个体个数设为30。使用粒子群算法得到的实验结果见表4。

为了更直观地展现并对比各个算法的效果,将各个分类模型的鉴别率综合在一张表内,结果见表5。

表 4　粒子群算法实验结果

数据类别	训练样本数	测试样本数	最优核参数 σ	最优惩罚因子 C	鉴别率/%
(a)	578	1 126	0.93	675.53	84.98
(b)	476	912	2.34	376.87	76.48
(c)	502	843	0.09	402.47	82.73
(d)	521	932	0.24	356.49	76.66

表 5　各算法实验结果综合　　%

数据类别	交叉验证法	网格搜索法	遗传算法	粒子群算法
(a)	81.36	82.88	83.64	84.98
(b)	73.24	72.83	74.38	76.48
(c)	78.75	81.63	81.78	82.73
(d)	71.87	72.48	73.59	76.66

由表 5 可以看出,通过各种方法寻找到的核参数构成的分类模型对于各类数据的鉴别率整体呈上升趋势。(a)类数据的鉴别效果在四组数据里面相对较好。因为数据涉及到三类图像,所以(d)类数据鉴别效果相对不理想。

综合所有表的结果可以看出,除了对于(b)类数据,网格搜索法相对交叉验证法在各类数据的鉴别效果略微提升,但是整体表现不尽人意。就网格搜索法来说,寻找最优参数的效率取决于初始范围与设定的步长,两者的选择稍不准确,则非常容易错过最优解。遗传算法与粒子群算法同属群智能算法,由表 5 可以看出,二者的鉴别效率整体比前两种算法都要好。群智能算法的特点表现在寻找过程充分智能化,能够有效避免陷入局部最优的情况。粒子群算法更是如此,最终取得了最好的鉴别效率。

4　结束语

针对在图像来源鉴别如何构造分类模型进行了讨论分析,提出了行之有效的方法。首先比较分析了关于 SVM 的几种常用核函数,决定选取细分效果更明显的高斯径向基函数,这有利于图像鉴别中的多类鉴别问题。其后,为确定最优核参数与惩罚因子,首先分析了现有的核参数寻优算法,再对各个算法的性能进行实验验证。实验结果表明,粒子群算法在核参数寻优问题上,不仅达到了速度较快的效果,而且其得到的核参数与惩罚因子所构造的分类模型鉴别率最高,达到了对图像来源鉴别率预期的效果。但是,维数问题始终是分类模型绕不开的问题,如何避免特征集维数过高影响最终鉴别效果更是分类模型亟待解决的问题,也是图像来源鉴别的重点研究方向。

参考文献:

[1] Yong I Y. Detection of digital forgeries using an image interpolation from digital images[C]//Proc of IEEE international symposium on consumer electronics. [s. l.]:IEEE,2008:1-4.

[2] Sencar H T,Memon N. Overview of state-of-the-art in digital image forensics[C]//Proc of WSPC. [s. l.]:World Scientific Press,2008.

[3] Khanna N,Mikkilineni A K,Martone A F. A survey of forensic characterization methods for physical devices[J]. Digital Investigation,2006,3:17-28.

[4] Chang C C,Lin C J. LIBSVM:a library for support vector machines[J]. ACM Transactions on Intelligent Systems & Technology,2011,2(3):389-396.

[5] Schlkopf B, Smola A J. Learning with kernels[M]. Cambridge:MIT Press,2001.

[6] Schlkopf B,Smola A J. Support vector machines and kernel algorithms[M]. [s. l.]:John Wiley and Sons,2003.

[7] Vladimir N V. The nature of statistical learning theory[M]. New York:Springer-Verlag,1995.

[8] 邓　蕊,马永军,刘尧猛. 基于改进交叉验证算法的支持向量机多类识别[J]. 天津科技大学学报,2007,22(2):58-61.

[9] 王兴玲,李占斌. 基于网格搜索的支持向量机核函数参数的确定[J]. 中国海洋大学学报:自然科学版,2005,35(5):859-862.

[10] 刘东平,单甘霖,张岐龙,等. 基于改进遗传算法的支持向量机参数优化[J]. 微计算机应用,2010(5):11-15.

[11] 朱家元,杨　云,张恒喜,等. 基于优化最小二乘支持向量机的小样本预测研究[J]. 航空学报,2004,25(6):565-568.

[12] 张培林,钱林方,曹建军,等. 基于蚁群算法的支持向量机参数优化[J]. 南京理工大学学报:自然科学版,2009,33(4):464-468.

[13] 张潇丹,胡　峰,赵　力. 基于改进的蛙跳算法与支持向量机的实用语音情感识别[J]. 信号处理,2011,27(5):678-689.

[14] 熊　军,高敦堂,都思丹,等. 变异率和种群数目自适应的遗传算法[J]. 东南大学学报:自然科学版,2004,34(4):553-556.

[15] 吴秋玲,杨启文. 改进型自适应遗传变异算子[J]. 河海大学常州分校学报,2005,19(4):12-15.