

# 融合混合优化组合的大规模场景图像分类算法

王 燕<sup>1</sup>, 曹建芳<sup>1,2</sup>, 李艳飞<sup>2</sup>

(1. 忻州师范学院 计算机系, 山西 忻州 034000;

2. 太原科技大学 计算机科学与技术学院, 山西 太原 030024)

**摘 要:** 图像获取设备的普及和网络技术的发展导致数字图像迅速增长, 面对海量图像, 传统的单节点架构的分类算法性能急剧下降。针对上述问题, 以场景图像为研究对象, 提出了一种集群环境下的融合混合优化和组合技术的大规模图像分类方法。将 ABC 算法和 PSO 算法优化后的 SVM 作为弱分类器, 由 Adaboost 算法组合弱分类器输出构建强分类器; 利用 Hadoop 平台下的 MapReduce 并行编程模型对算法进行并行化设计, 提出 P-Adaboost-(ABC-PSO-SVM) 算法, 构造了大规模场景图像的自动分类模型。多组对比实验表明, 相对于传统的单机平台下的分类算法, 当图像数量达到 50 000 时, 该算法在 SUN Database 场景图像库上的平均分类准确率达 87.6%, 训练耗时仅为 98 s。实验结果充分说明, 该算法适合大规模场景图像的自动分类预测。

**关键词:** 混合优化; Adaboost 算法; 集群环境; MapReduce 并行编程模型; 分类模型

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2019)09-0086-06

doi: 10.3969/j.issn.1673-629X.2019.09.017

## A Classification Algorithm for Large-scale Scene Images Fusing Hybrid Optimization and Combination

WANG Yan<sup>1</sup>, CAO Jian-fang<sup>1,2</sup>, LI Yan-fei<sup>2</sup>

(1. Department of Computer, Xinzhou Teachers University, Xinzhou 034000, China;

2. School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

**Abstract:** The popularity of image acquisition devices and the development of network technology have resulted in the rapid growth of digital images. In the face of large amounts of images, the performance of classification algorithms under traditional single machine platforms has dropped dramatically. To solve the above problems, we propose a classification algorithm for large-scale scene images fusing hybrid optimization and combination technology in a cluster environment. Taking SVM optimized by ABC algorithm and PSO algorithm as weak classifiers, we use Adaboost algorithm to build a strong classifier through combining the outputs of weak classifiers. Then the MapReduce parallel programming model in the Hadoop platform is applied to carry on parallel design to the proposed algorithm, namely P-Adaboost-(ABC-PSO-SVM) algorithm. Finally, the automatic classification model for large-scale scene images is constructed. Compared with the traditional classification algorithms using single platform, the experiment shows that the average classification accuracy of the proposed algorithm is 87.6% and the training time is only 98 seconds in the SUN Database when the image number reaches 50 000. The experimental results further verify that the proposed algorithm is suitable for automatic classification and prediction for large-scale scene images.

**Key words:** hybrid optimization; Adaboost algorithm; cluster environment; MapReduce parallel programming model; classification model

## 0 引 言

随着图像获取设备的普及和计算机网络技术、多媒体技术的快速发展, 各类图像数据正在迅速增长<sup>[1]</sup>。

作为最为常见的一类图像数据, 场景图像的数量更是呈现指数级增长趋势。人工智能和机器学习的创新发展使得利用计算机自动提取场景图像特征并对其分类

收稿日期: 2018-11-05

修回日期: 2019-03-05

网络出版时间: 2019-04-24

基金项目: 山西省大学生创新创业训练项目(2017382); 山西省自然科学基金(201701D121059); 山西省艺术科学规划课题(2017F06); 山西省教育科学规划课题(GH-17059)

作者简介: 王 燕(1990-), 女, 研究方向为数字图像理解; 曹建芳, 博士, 教授, CCF 高级会员(25770S), 研究方向为数字图像理解、大数据技术。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190424.1055.064.html>

成为人工智能和计算机视觉领域的重要研究课题之一<sup>[2]</sup>。在机器学习领域,常见的分类算法有 K 最近邻 (KNN)<sup>[3]</sup>、贝叶斯 (Bayes)<sup>[4]</sup>、神经网络<sup>[5]</sup>、决策树<sup>[6]</sup>和支持向量机<sup>[7]</sup>。然而,上述算法都在不同层面存在一定的缺陷。KNN 算法本身的时间复杂度和空间复杂度都很高;Bayes 分类算法需要计算事件的先验概率,而且对输入数据的表达形式非常敏感;神经网络算法收敛速度慢并且易陷入局部最优;决策树对信息缺失的处理较为困难,易有过拟合的现象发生;而 SVM 因其广义属性,能够提供较高的分类精度,是一种应用广泛的分类算法<sup>[8]</sup>。但随着研究的不断深入,研究者们发现,由于受存储能力和计算能力的约束,传统的单节点架构的 SVM 算法在处理海量数据时,在内存需求和计算时间方面会产生“瓶颈”,分析处理效率会急剧下降。

2004 年,Google 公司推出了一个处理海量数据的并行编程模型 MapReduce<sup>[9]</sup>,因其具有良好的接口和运行支持库,并隐藏了实现的复杂细节,支持并行执行大规模的计算任务,从而在众多领域得到了广泛应用。在医疗领域,高汉松等<sup>[10]</sup>设计了一个对海量医疗数据进行挖掘和分析的医疗大数据挖掘平台;在生物领域,涂金金等<sup>[11]</sup>在 Hadoop 平台下应用 MapReduce 模型分析基因的表达数据;在通信领域,中国移动为了解决大流量数据业务快速、廉价处理的问题,提出了“大云”数据挖掘系统的构想<sup>[12]</sup>。MapReduce 并行编程模型在各个领域的应用愈来愈多。

因此,针对上述单机平台下 SVM 算法在处理大规模数据时面临的“瓶颈”,为进一步提高分类准确率,基于混合优化组合的思想,文中提出一种新的集群环境下的场景图像分类方法:P-Adaboost-(ABC-PSO-SVM)模型。该模型首先应用人工蜂群(ABC)和粒子群(PSO)算法对 SVM 的参数进行混合优化,以得到最优的 SVM 参数对;然后使用 Adaboost 算法加强多个 SVM 分类器的结果,进一步提高分类准确率;最后利用 MapReduce 并行编程模型对算法进行并行化设计,以更好地改进该算法在处理海量数据时的时间性能。通过在场景图像库 SUN Database 上进行实验,并与传统单节点架构的 SVM 分类模型和集群环境下的并行 SVM 分类模型进行对比,验证了 P-Adaboost-(ABC-PSO-SVM)模型对海量场景图像的分类效果。

## 1 P-Adaboost-(ABC-PSO-SVM)分类算法

### 1.1 ABC-PSO 混合优化 SVM 参数

#### 1.1.1 SVM 参数优化分析

SVM 算法以结构风险最小化原则为基础,使用核

函数将低维的线性不可分问题映射到高维空间,转换成线性可分问题进行处理。核函数的构造对 SVM 算法的性能起着关键性作用<sup>[13]</sup>。局部核函数和全局核函数这两类核函数是 SVM 最常使用的核函数,局部核函数的优点是学习能力强,但存在的问题是泛化能力弱;全局核函数虽然泛化能力强,但学习能力却很弱。为了兼顾两者的学习能力和泛化能力,研究者们提出了将局部核函数和全局核函数相结合构造混合核函数的思想,最常见的是将径向基函数(RBF)和多项式核函数进行线性组合,构造出满足 Mercer 条件的混合核函数。

$$K_{\text{mix}} = \lambda K_{\text{poly}} + (1 + \lambda) K_{\text{rbf}} \tag{1}$$

其中,  $\lambda \in (0, 1)$ ;  $K_{\text{poly}} = [(x \cdot x_i) + 1]^q$  为多项式核函数;  $K_{\text{rbf}} = -\gamma \|x - x_i\|^2$  为 RBF 核函数。

SVM 分类性能的优劣与 SVM 参数有很大关系,因此有必要对 SVM 的参数做优化调整。文中需要优化的参数是:惩罚因子  $C$ 、核参数  $\gamma$  和调节因子  $\lambda$ 。惩罚因子  $C$  用于决定 SVM 重视离群点带来损失的程度,核参数  $\gamma$  决定支持向量之间存在的关联程度,调节因子  $\lambda$  的经验值寻优范围一般取值在 0.50 ~ 0.99 之间。

#### 1.1.2 ABC-PSO-SVM 算法

PSO 算法<sup>[14]</sup>是一种群体智能进化算法,其思想源于鸟群觅食行为,优点是实现容易、精度高、收敛快,但存在局部寻优能力差、易产生早熟收敛等问题。ABC 算法<sup>[15]</sup>是一种全局群智能优化算法,其思想源于蜂群采蜜行为,应用的主要优势体现在不需要了解实际问题的特殊信息,只需要对问题进行优劣的比较,通过各人工蜂个体的局部寻优行为,最终在群体中使全局最优值突现出来,收敛速度较快。如果能充分利用 PSO 算法和 ABC 算法的优势,将两者结合起来进行混合优化,就会很好地克服 PSO 算法的缺陷,增强优化算法的鲁棒性。基于此,文中采用 ABC-PSO 算法对 SVM 的参数实行混合优化,以形成最优的 SVM 分类模型。优化算法步骤为:

Step1:初始化 PSO 和 ABC 算法的参数。主要包括:种群规模、最大迭代次数、PSO 算法的速度、ABC 算法的食物源数量和控制参数、ABC 和 PSO 算法的初始  $(C, \gamma, \lambda)$ 。

Step2:将  $(C, \gamma, \lambda)$  作为 SVM 的参数,对 SVM 进行训练和测试。

Step3:计算和更新适应度值。如果满足最大迭代次数,即可得到 SVM 的最优参数  $(C, \gamma, \lambda)$ ,算法停止;否则,执行 Step4。

为提高 SVM 算法的分类精度,文中将 PSO 算法和 ABC 算法的适应度函数分别定义为:  $\text{fit}_{\text{ps}} = -v_{\text{acc}}$  和

$\text{fit}_{\text{abc}} = \frac{v_{\text{acc}}}{1 + v_{\text{acc}}}$ , 其中  $v_{\text{acc}}$  为 SVM 的分类准确率。

Step4: 返回 Step2 执行, 继续迭代寻优。

## 1.2 Adaboost 算法组合参数优化的 SVM

Adaboost 算法是对同一问题集成多个弱分类器的结果共同决策的一种机器学习技术, 通过执行基本的分类算法, 获得多个不同的弱分类器, 训练过程中自适应改变样本权重, 错分的样本被赋予较大的权重, 反复迭代, 最后使用加权投票的方法获得最终的判决结果<sup>[16]</sup>。文中使用 Adaboost 算法对 ABC-PSO 算法优化后的 SVM 分类器进行加强, 提出 Adaboost-(ABC-PSO-SVM) 算法, 将 ABC-PSO-SVM 分类器作为弱分类器提供给 Adaboost 集成学习, 在训练过程中不断改变输入样本的权重以重构少数类别样本, 加强对错分样本的训练, 最终构建强分类器。算法步骤为:

Step1: 数值初始化。输入训练样本集  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$  和训练迭代次数  $T$ , 优化的 SVM 参数  $(C, \gamma, \lambda)$ , 并将训练样本集权值分布  $D_t(i)$  设置为  $1/N$ 。

Step2: 训练弱分类器 ABC-PSO-SVM。

Step2.1: 根据权值分布  $D_t(i)$  得到第  $t$  次弱分类器  $H_t = L(D, D_t)$ 。

Step2.2: 计算  $H_t$  的错误率  $\varepsilon_t$ 。

$$\varepsilon_t = \sum_{i=1}^m D_t(i), H_t(x_i) \neq y_i \quad (2)$$

Step2.3: 根据  $\varepsilon_t$  的值判断是否更新样本权重。

如果  $0 < \varepsilon_t \leq 0.5$ , 先根据式 3 计算权重值  $\alpha_t$ , 然后根据式 4 更新样本权重, 继续执行 Step2.1; 否则执行 Step3。

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \quad (3)$$

$$D_{t+1}(i) = \frac{D_t(i)}{|D_t|} \cdot \begin{cases} \exp(-\alpha_t), & h_t(x_i) = y_i \\ \exp(\alpha_t), & h_t(x_i) \neq y_i \end{cases} \quad (4)$$

Step3: 线性组合  $T$  轮训练后得到的  $T$  组弱分类器为  $H_t(x)$ , 得到强分类器  $H(x)$ 。

$$H(x) = \text{sign} \left( \sum_{t=1}^T \alpha_t H_t(x) \right) \quad (5)$$

## 1.3 Adaboost-(ABC-PSO-SVM) 算法的并行化设计

### 1.3.1 MapReduce 并行编程模型

Hadoop 是一个用于分布式处理海量各类型数据的软件框架, 其中 HDFS 和 MapReduce 是 Hadoop 平台的两个核心设计。HDFS 是采用主/从模式体系结构的分布式文件系统, 将大量数据分布存储于多台相关的计算机上, 以实现大规模数据集的流式访问; MapReduce 是一种并行编程模型, 能够将计算任务和

数据分配到 Hadoop 集群的各个节点上, 让各节点并行执行任务, 得到中间结果后进行汇总并再次向各节点分配计算, 以获得最终结果。MapReduce 在执行任务的过程中, 借助函数式编程方法, 将计算分为 Map 和 Reduce 两个任务, 每个任务的处理均以键值对的形式进行输入和输出, 通过定义 Mapper() 和 Reducer() 函数实现一个键值对到另一个键值对的映射<sup>[17-18]</sup>。Mapper() 函数将大数据集分割成小数据集分配给各节点进行并行处理, Reducer() 函数汇总各节点的处理结果, 实现了分布式并行处理。

### 1.3.2 P-Adaboost-(ABC-PSO-SVM) 算法

文中利用 MapReduce 并行编程模型在 Hadoop 集群环境中对提出的 Adaboost-(ABC-PSO-SVM) 算法进行并行化设计, 即 P-Adaboost-(ABC-PSO-SVM) 算法, 以解决单机平台下的 Adaboost-(ABC-PSO-SVM) 算法在处理大规模场景图像时硬件开销大、运行耗时长, 尤其是训练时间急剧增加的问题。算法主要包括对 Adaboost-(ABC-PSO-SVM) 算法并行化的 Map 和 Reduce 两个阶段, 另外, 考虑到如果在 Reduce 阶段直接对 Map 阶段产生的中间结果进行汇总, 集群环境各节点之间会产生很大的通信开销。因此, 文中在 Reduce 阶段之前设计了一个 Combine() 函数, 对 Map 阶段产生的结果在一定程度上进行了本地处理, 以减小节点之间的通信开销。算法伪代码描述如下:

Adaboost-(ABC-PSO-SVM)-Mapper()

Input: <弱分类器 ID, 样本特征值>

Output: <弱分类器 ID, 预测误差  $\varepsilon_t$ >

{

// 对每个弱分类器

ABC-PSO 优化 SVM 参数  $(C, \gamma, \lambda)$ ;

训练弱分类器;

{

计算预测误差  $\varepsilon_t$ ;

if ( $0 < \varepsilon_t \leq 0.5$ )

更新样本权重  $\alpha_t$ ;

}

获取弱分类器预测函数  $H_t(x)$  和更新的  $\varepsilon_t$ ;

输出 <弱分类器 ID, 预测误差  $\varepsilon_t$ >;

}

Adaboost-(ABC-PSO-SVM)-Combine()

Input: <弱分类器 ID, 预测误差  $\varepsilon_t$ >

Output: <弱分类器 ID, 弱分类器  $H_t(x)$ >

{

count ← 0; // 统计训练弱分类器数

// 对每个弱分类器

解析处理  $\varepsilon_t$  坐标值;

count ← count + 1;

在本地归约处理 ID 相同的键值对, 更新  $\varepsilon_t$ , 获得弱分类器



$H_i(x)$  的输出;

```
输出<弱分类器 ID,弱分类器  $H_i(x)$ >;
};
Adaboost-(ABC-PSO-SVM)-Reducer( )
Input:<弱分类器 ID,弱分类器  $H_i(x)$ >
Output:<预测误差  $\varepsilon_i$ ,强分类器  $H(x)$ >
{
  对各节点 ID 相同的键值对再次进行合并处理;
  对弱分类器  $H_i(x)$  进行线性组合,更新  $\varepsilon_i$ ,得到强分类器  $H(x)$ ;
  输出<预测误差  $\varepsilon_i$ ,强分类器  $H(x)$ >;
}
```

2 大规模场景图像分类实现

2.1 特征提取

尺度不变特征变化 ( scale - invariant feature transform,SIFT)是 David G. 提出的基于尺度空间的特征描述算子,是一种通过检测多尺度图像金字塔极值点提取图像的位置、尺度、旋转等关键点的计算方法<sup>[19]</sup>。由于 SIFT 特征不受图像大小、旋转、光线等变换的影响,区分力强,具有很好的鲁棒性,现已成为应用广泛的图像特征提取算法。文中选取场景图像的 SIFT 特征作为分类器的输入特征参数,在 Hadoop 平台下对 SIFT 特征的提取进行并行化设计,算法步骤为:

Step1:将不同类别的场景图像设置为 Map 任务的输入,其中 key 为图像名称,value 为场景图像。

Step2:Map 阶段,利用 OpenCV 函数库中的 Dense SIFT 算法提取场景图像的 SIFT 特征,输出形如<图像名称,128 维特征向量>的键值对,建立多个文件夹,每个文件夹保存一幅场景图像的 SIFT 特征向量。

由于 SIFT 特征提取过程较为简单,考虑到节约时间和通信开销,在并行提取场景图像 SIFT 特征时未设置 Reduce 任务。

2.2 分类模型构建及实现

文中设计的海量场景图像分类模型框架如图 1 所示。

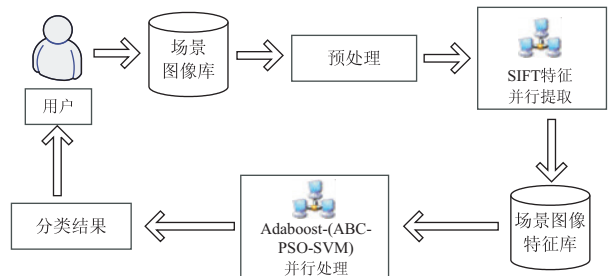


图 1 海量场景图像分类模型框架

分类模型基于 Hadoop 大数据平台架构,在提取场景图像的 SIFT 特征后,将其作为分类器的输入参数,

然后使用提出的 Adaboost-(ABC-PSO-SVM)算法建模,并利用 MapReduce 并行编程模型对 SIFT 特征提取和 Adaboost-(ABC-PSO-SVM)算法进行并行化设计,对大规模场景图像进行类别预测,将得到的分类结果反馈给用户。具体步骤为:

Step1:并行提取场景图像的 SIFT 特征,生成特征矩阵。

Step2:确定 Adaboost-(ABC-PSO-SVM)分类器的结构。

Step3:搭建 Hadoop 集群,并行训练 Adaboost-(ABC-PSO-SVM)分类器,不断更新样本权重,组合输出结果。

Step4:使用训练好的分类器,对大规模场景图像进行分类识别,并将结果反馈给用户。

3 实验结果及分析

3.1 实验环境和数据来源

实验环境:采用局域网内 5 台计算机搭建了 Hadoop 集群环境,其中 1 台作为主(master)节点,其余 4 台作为从(slave)节点。所有节点计算机硬件配置都采用酷睿 i7 四核八线程 4.2 G 处理器,8 G 内存,4 T 硬盘空间的基本配置;软件配置:操作系统是 64 位的 Ubuntu 14. 04, Java 环境为 jdk1. 7. 0\_79, Hadoop 为 Hadoop-2. 5. 1(64 位编译)。

数据来源:实验数据来源于 SUN Database 场景图像库,包含 131 067 幅、908 个类别的场景图像。该图像库是供研究者们免费使用的,其规模不断扩大。为便于处理,文中实验用的图像都被预处理成 200×200 像素的大小。

3.2 实验对比分析

为验证该算法的有效性,在 SUN Database 图像库中随机选取 50 000 张场景图像,构造了不同规模的数据集,从最优参数组合及进化迭代次数、分类准确率和分类器训练耗时等几个方面进行了实验对比分析。

3.2.1 SVM 参数优化对比

ABC 算法和 PSO 算法同时对 SVM 参数进行混合优化,不需要遍历所有的参数点就能很快找到全局最优解。在 SUN Database 图像库中随机选取了 10 个类别,共 5 000 张场景图像,对 ABC-PSO 算法参数优化与 PSO 算法参数寻优及进化代数做了对比,如表 1 所示。

表 1 的数据表明,数据类别不同,数据的复杂性也不同,两种算法在不同数据类别下的进化速度会有快有慢,得到的参数组合也不一样。但无论针对哪种数据类别,ABC-PSO-SVM 算法的进化速度都比 PSO-SVM 算法的进化速度快得多,而且 ABC-PSO-SVM 算

法能找到优于 PSO-SVM 算法的参数组合,充分说明 了 ABC 算法和 PSO 算法进行混合优化的优越性。

表 1 最优参数组合及进化代数对比

数据集	PSO-SVM		ABC-PSO-SVM	
	( $C, \gamma$ )	进化代数	( $C, \gamma, \lambda$ )	进化代数
Airport terminal	(12.43,4.392)	61	(7.21,1.793,0.548)	16
Beach	(45.28,0.713)	172	(11.69,0.964,0.845)	65
Bedroom	(20.43,0.449)	73	(15.49,0.758,0.992)	40
Conference room	(17.72,1.238)	79	(11.65,1.040,0.675)	42
Forest path	(29.85,2.590)	116	(24.55,1.230,0.772)	76
Highway	(75.20,1.193)	176	(17.28,1.036,0.631)	95
Kitchen	(47.85,0.647)	161	(11.20,0.948,0.791)	59
Mountain snowy	(19.22,0.375)	73	(14.56,0.752,0.980)	46
Playground	(12.22,4.336)	53	(7.06,1.692,0.554)	13
Vegetable garden	(22.36,1.949)	101	(13.49,1.385,0.669)	60

3.2.2 分类准确率比较

为验证文中算法的分类性能,使用 3.2.1 中的 10 个数据类别,采用复制的方法构造了包含 10 000 张场景图像的数据集,将其中 30% 的图像作为测试数据集,对不同数据类别的测试准确率如图 2 所示。

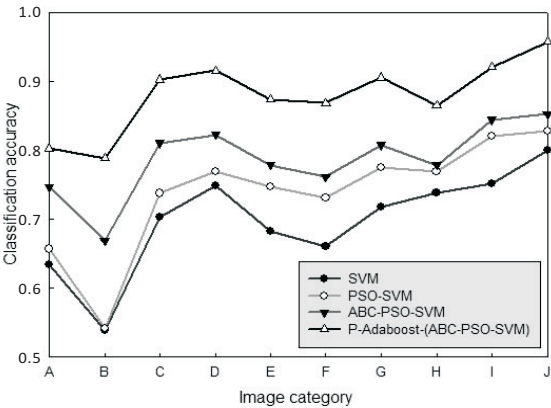


图 2 不同类别图像测试准确率对比

图中 A—Airport terminal, B—Beach, C—Bedroom, D—Conference room, E—Forest path, F—Highway, G—Kitchen, H—Mountain snowy, I—Playground, J—Vegetable garden。

从图 2 可以看出,文中算法的分类准确率明显高于其他单机平台上的 SVM、PSO-SVM 和 ABC-PSO-SVM 算法。这主要是因为:一方面是基于组合优化的思想使用 Adaboost 算法对混合优化参数后的 SVM 分类器进行了加强,构建强分类器提高了分类精度;另一方面采用 MapReduce 并行编程模型对算法进行了并行化设计,增强了计算能力,训练的分类模型更优,进一步提高了分类准确率。

另外,为进一步验证文中算法的有效性,又构造了不同规模的数据集,使用不同的算法做了分类对比实验(训练样本:测试样本=7:3),如表 2 所示。

表 2 不同数据规模下的分类性能对比

图像规模	分类算法	训练准确	测试准确
		率/%	率/%
2 000	SVM	0.841	0.832
	PSO-SVM	0.860	0.859
	ABC-PSO-SVM	0.912	0.912
	P-Adaboost-(ABC-PSO-SVM)	0.938	0.936
5 000	SVM	0.825	0.814
	PSO-SVM	0.851	0.846
	ABC-PSO-SVM	0.906	0.903
	P-Adaboost-(ABC-PSO-SVM)	0.929	0.929
10 000	SVM	0.788	0.785
	PSO-SVM	0.832	0.828
	ABC-PSO-SVM	0.893	0.882
	P-Adaboost-(ABC-PSO-SVM)	0.919	0.915
25 000	SVM	0.713	0.698
	PSO-SVM	0.744	0.738
	ABC-PSO-SVM	0.802	0.802
	P-Adaboost-(ABC-PSO-SVM)	0.881	0.880
50 000	SVM	0.663	0.571
	PSO-SVM	0.658	0.634
	ABC-PSO-SVM	0.732	0.713
	P-Adaboost-(ABC-PSO-SVM)	0.879	0.876

从表 2 中的数据可以看到,在不同规模的数据集下,P-Adaboost-(ABC-PSO-SVM)算法的训练准确率和测试准确率都要高于其他 3 种单机平台架构下的算法;而且,随着场景图像数据规模的不断增大,虽然各种算法的分类准确率都在下降,但文中算法的下降趋势较为平缓,而其他 3 种算法的下降趋势很明显,尤其是当场景图像规模超过 10 000 时;另外,对于不同的

数据规模,文中算法的测试准确率和训练准确率非常接近。这都充分说明了 MapReduce 并行编程模型和 Hadoop 平台分布式处理的计算能力的强大,在这种集群环境下训练得到了最优的 SVM 分类模型,使得分类精度不会随着数据集的增大而急剧下降,同时测试准

确率接近于训练准确率。

### 3.2.3 训练耗时对比

为验证文中算法在处理大规模数据集时的时间性能,对不同分类器的训练时间进行了实验比较,如表 3 所示。

表 3 不同算法在不同规模数据集下的训练时间对比

分类算法	不同规模图像数据集下的训练时间/s				
	2 000	5 000	10 000	25 000	50 000
SVM	11	26	78	1 902	6 575
PSO-SVM	12	26	80	1 920	6 579
ABC-PSO-SVM	14	27	80	1 919	6 578
P-Adaboost-(ABC-PSO-SVM)	0.78	1.0	2.3	31	98

表 3 的数据明显表明,集群环境下设计的并行算法训练时间要比单机平台下的算法训练时间少得多;而且,随着图像规模的不断增大,单机平台下的算法训练时间会急剧增加,而文中算法基于集群环境设计,能够进行分布式并行处理,因此训练时间不会增加太多,实现了短时间内对海量数据的处理。

## 4 结束语

对集群环境下融合混合优化和组合思想的大规模场景图像分类算法做了深入探讨。对使用 ABC 算法和 PSO 算法混合优化 SVM 的参数进行了研究,将 Adaboost 算法组合 SVM 的分类结果构建强分类器,并利用 Hadoop 平台下的 MapReduce 并行编程模型对提出的算法进行了并行化设计,将其应用于大规模场景图像的分类问题。在 SUN Database 场景图像库上的对比实验结果表明,该算法能快速优化 SVM 的参数,分类准确率高,训练耗时少,构建的 Hadoop 集群环境能充分利用各节点计算机的资源,发挥其计算能力,提高分类器的训练和预测的速度和精度,获得最优的分类模型。相对于单机平台下的传统算法,系统性能良好,很好地体现了分布式并行处理集群环境的强大运算能力,拓宽了大数据技术在数字图像理解领域的应用。

### 参考文献:

[1] 韩伟,张学庆,陈 旻. 基于 MapReduce 的图像分类方法[J]. 计算机应用,2014,36(4):1600-1603.

[2] 张慧娜,李裕梅,傅莺莺. 基于 Haar-CNN 模型的自然场景图像分类的研究[J]. 四川师范大学学报:自然科学版,2017,40(1):119-126.

[3] 刘应东,牛惠民. 基于 k-最近邻图的小样本 KNN 分类算法[J]. 计算机工程,2011,37(9):198-200.

[4] 邓桂骞,赵跃龙,刘 霖,等. 一种优化的贝叶斯分类算法[J]. 计算机测量与控制,2012,20(1):199-201.

[5] 章 毅,郭 泉,王建勇. 大数据分析的神经网络方法[J].

工程科学与技术,2017,49(1):9-18.

[6] 储茂祥,王安娜,巩荣芬. 一种改进的最小二乘孪生支持向量机分类算法[J]. 电子学报,2014,42(5):998-1003.

[7] CAO J F, CHEN L C, SHI H, et al. A parallel Adaboost-back-propagation neural network for massive image dataset classification[J]. Scientific Reports, 2016, 6:38201.

[8] 刘志强,顾 荣,袁春风,等. 基于 SparkR 的分类算法并行化研究[J]. 计算机科学与探索,2015,9(11):1281-1294.

[9] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113.

[10] 高汉松,肖 凌,许德玮,等. 基于云计算的医疗大数据挖掘平台[J]. 医学信息学杂志,2013,34(5):7-12.

[11] 涂金金,杨 明,郭丽娜. 基于 MapReduce 的基因数据密度层次聚类算法[J]. 中国科学技术大学学报,2014,44(7):537-543.

[12] 唐子民. Hadoop 在移动云计算中的应用研究[J]. 山东通信技术,2012,32(4):5-9.

[13] 匡芳君,徐蔚鸿,张思扬. 基于改进混沌粒子群的混合核 SVM 参数优化及应用[J]. 计算机应用研究,2014,31(3):671-674.

[14] 姜建国,叶 华,马亚华. 一种采用抽样策略的 PSO 算法[J]. 控制与决策,2015,30(10):1779-1784.

[15] 张新明,魏 峰,牛丽平,等. 混合排名映射概率和混沌搜索的 ABC 算法[J]. 计算机科学,2014,41(2):102-106.

[16] 廖雨婷,王慧琴,柴 茜,等. Adaboost 算法在图像型火灾探测中的应用研究[J]. 计算机应用与软件,2015,32(4):153-155.

[17] WHITE T. Hadoop: the definitive guide[M]. 3rd ed. Sebastopol: O'Reilly Media, 2012.

[18] DAVID G. Distinctive image features from scale-invariant keypoints[J]. International Journal in Computer Vision, 2004, 60(2):91-110.

[19] LIU Yu, LIU Shuping, WANG Zengfu. Multi-focus image fusion with dense SIFT[J]. Information Fusion, 2015, 23:139-155.