利用 MapReduce 模型训练支持向量机的人脸识别方法

童小念,文卫蔚
(中南民族大学 计算机科学学院,武汉 430074)

摘 要 为了在移动互联网中快速识别人脸图像,提出了利用云计算服务端的 MapReduce 模型训练支持向量机(SVM)进行人脸识别的方法. 实验结果表明:该算法在保证人脸识别率的前提下,明显提升了支持向量机的训练速度. 该算法对于移动互联网环境下的人脸识别有一定的实用价值.

关键词 人脸识别;支持向量机;MapReduce 模型;主成分分析
中图分类号 TP391 **文献标识码** A **文章编号** 1672-4321(2013)01-0083-04

A Face Recognition Method for Support Vector Machine Learning by MapReduce Model

Tong Xiaonian, *Wen Weiwei*
(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract In order to identify face images quickly in the mobile Internet, a face recognition approach for support vector machine(SVM) learning by using MapReduce model is proposed. The experimental results show that the learning speed of SVM is improved observably, and the accuracy rate for the identification of face images is ensured. The arithmetic by Using MapReduce to train Support Vector Machine has a practicability to face images recognition in the mobile Internet.

Keywords MapReduce; face recognition; support vector machine; principal component analysis

人脸识别技术是利用计算机对人脸图像进行分析,从中提取有效信息以辨认身份的一门技术,相对于指纹识别、视网膜和虹膜扫描等其他基于生物特征识别的技术,它更直观、更容易鉴别,被广泛地应用于证件识别、门禁监控和视频会议中. 随着移动互联网的发展,移动环境下将人脸识别作为身份认证的应用场合也日益趋多,已成为当前的一个研究课题.

可是,移动互联网中的人脸识别计算受到网络速度以及移动终端电池续航能力等诸多因素的限制,其大量样本数据的复杂计算不适合在移动终端进行. 随着云计算技术的提出,云计算服务端可以提供很好的存储与计算能力,移动互联网中的诸多应用也都趋向于云计算方向发展. 因此,探讨在移动互联网环境下应用云计算方法进行人脸识别,有很好的实用意义.

支持向量机(SVM)是一种基于统计学理论的学习算法[1],它以结构风险最小化原理为基础[2],通过构造最优超平面对未知样本进行分类. 支持向量机能够有效地克服"维度灾难"问题,在文本分类、生物信息、图像分类等领域有很好的应用. MapReduce[3] 是由 Google 提出的一种应用于云计算的编程模型,它能够提供较强的并行处理能力,在机器学习的过程中,可达到减少训练时间的目的. 本文利用 MapReduce 模型训练支持向量机来进行人脸识别,旨在加速支持向量机的人脸学习训练时间,以适应移动环境下的人脸识别要求.

收稿日期 2013-01-07
作者简介 童小念(1954-),女,教授,研究方向:多媒体技术,E-mail:tongxiaonian@yahoo.com.cn
基金项目 湖北省自然科学基金资助项目(2012FFB07404)

1 基于主成分分析(PCA)的人脸特征提取方法

人脸特征表示方法主要有基于几何特征的表示方法与基于统计特征的表示方法[4]. 基于几何特征的表示方法通过对人脸特征例如眼睛、鼻子、嘴巴等部位的相对位置进行提取,其效果的好坏依赖于是否存在满足识别要求的精确的人脸特征检测机制,所以到目前为止对人脸特征的精确检测仍然比较困难. 而基于统计特征的识别方法不仅仅针对人脸的某一具体几何特征,它还从整个人脸的角度利用统计原理从多张人脸中提取它们的共有特征,利用这些特征进行人脸识别. 相对于人脸几何特征识别,人脸统计特征降低了特征描述的困难,因此基于统计特征的人脸表示方法越来越得到重视,并得到了广泛应用. 本文采用主成分分析(PCA)[5]进行人脸特征提取,PCA是一种掌握事物主要矛盾的统计分析方法,它的特点是将主要影响因素从复杂的多元事物中分离,使复杂的问题简化.

PCA由K-L变换[6,7]实现数据降维,对于一幅l×h的人脸图像,PCA将其构造成M=l×h维的列向量,M即为人脸向量的维度,训练样本集的总体散布矩阵如公式(1)所示.

$$S_T = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^{\mathrm{T}}. \tag{1}$$

公示(1)中,μ为样本集的平均向量,N为训练样本总数($N < M$),x_k为第k个训练样本的图像向量,令:

$$X = (x_1 - \mu, x_2 - \mu, \cdots, x_N - \mu). \tag{2}$$

构造矩阵$Q = X^{\mathrm{T}}X$,则矩阵S_t最多有M个特征值与特征向量,矩阵Q最多有N个特征值与特征向量. 根据奇异值分解定理,S_t和Q的前N个最大的特征值是相同的,首先通过Q矩阵计算出其特征值$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$及对应的标准正交特征向量$u_1, u_2, \cdots, u_n$,然后根据公式(3)计算$S_t$中前$N$个最大的特征值对应的标准正交特征向量$v_1, v_2, \cdots, v_N$.

$$v_i = \frac{1}{\sqrt{\lambda_i}} X u_i. \tag{3}$$

对于样本x,主成分特征如公式(4)所示.

$$y = (v_1, v_2, \cdots, v_N)^{\mathrm{T}} x. \tag{4}$$

如此将原图像从M维下降到维,维的投影系数则作为人脸图像的特征向量输入到分类器进行识别.

2 多类支持向量机

2.1 支持向量机基本原理

支持向量机(SVM)是由Corinna Cortes和Vapnik等人首先提出的[8],它由统计学习理论发展而来,是一种新的机器学习算法.

支持向量机方法首先需要构造最优超平面,支持向量机使一组高维向量在超平面作用下进行分隔以达到距离最大化. 在这个过程中,需要通过适当的核函数(Kernel Function)将两组不同类别的向量组在高维空间分散开,然后在这个新空间求最优分类面.

设线性可分训练集样本为$S = ((x_1 \cdot y_1), (x_2 \cdot y_2), \cdots, (x_l \cdot y_l))$,其中$x_i \in R^d$;$y_i \in (1, -1)$,为了使所有样本能够被一个超平面正确分开,必须满足公式(5).

$$y_i [(w \cdot x_i) + b] \geqslant 1, i = 1, 2, \cdots, l. \tag{5}$$

利用拉格朗日优化方法,可将上述问题转化为对偶问题,α_i为原问题中对应每个约束条件的拉格朗日乘子:

$$\min_a \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{i=1}^{i} \alpha_i,$$
$$\text{subject to } \sum_{i=1}^{l} y_i \alpha_i = 0, 0 \leqslant \alpha \leqslant X, i = 1, \cdots, l. \tag{6}$$

其中$C > 0$作为对线性不可分样本的分类错误惩罚因子,求解这个二次规划问题,可以从训练样本中得到一系列对应$\alpha_i \neq 0$的向量,这些向量作为支持向量决定分类面.

2.2 多类支持向量机

传统的支持向量机只能解决两分类的问题,而人脸识别是多分类问题,因此针对人脸识别需要使用多分类的支持向量机. 目前主要有“一对多”分类法与“一对一”分类法.

“一对多”分类法是由Vapnik[9]提出的,它的主要思想是对训练集中的N类样本训练N个支持向量机,在对识别第i类的样本SVM进行训练时,该类样本为正样本,其它不属于第i类样本的样本作为负样本. 在进行分类判别的过程中将待识别的特征输入每个分类器,输出值最大的分类器对应识别结果. “一对多”分类法具有实现简单、训练时间短、容易进行类别扩展等优点,相对“一对一”分类法只需要较少的分类器数量,但是由于训练样本数量间的

不均衡使得“一对多”分类法的泛化能力较弱.

“一对一”分类法在 N 类样本间两两构造分类器，因此该方法对于 N 类的样本总共需要训练 $N(N-1)/2$ 个分类器[10]. 在进行判别的过程中，“一对一”分类法采用投票的方式取出现次数最多的结果作为识别结果. “一对一”分类法中构造单个支持向量机的过程相对比较简单，但是相比“一对多”分类法，其需要构造的总分类器的数目较多，随着类别的增加以平方增长，分类的速度也会随之变慢.

3　基于 MapReduce 的人脸识别方法

利用支持向量机进行人脸识别，降低机器学习的训练时间开销是一个重要的问题. 通过 MapReduce 提供的并行处理能力，能够减少训练时间的开销.

3.1　以 MapReduce 模型训练 SVM

MapReduce 将训练数据集分为多个子集，针对子集开发多级 SVM，即通过指定的 map 函数并行处理各个子集，再通过 Reduce 函数对分块处理的结果按照特定的规则进行归并处理，得到新的 SVM. MapReduce 系统能够自动处理数据的分块、分配与调度等问题，降低了使用难度.

如上所述，MapReduce 由两个执行阶段组成：Map 阶段与 Reduce 阶段. 在 Map 阶段通过用户指定一个 Map 函数将输入的键值对转化为一系列的中间键值对以 < key，value > 的形式输出，具有相同 key 值的键值对会交由相应的 Reduce 函数处理. 在 Reduce 阶段，对接收到的 < key，value > 键值对规约为 < key，list(values) > 键值对的形式，对每个 < key，list(values) > 键值对调用 reduce 方法并输出结果.

为了在 MapReduce 模型下训练支持向量机，考虑对于识别任务最终决定分类面的是支持向量，而且处于两个最优超平面间的样本对于支持向量机的调节具有重要作用，本文首先将训练样本集划分为若干个小的训练样本集，在 Map 任务中针对每个小样本集训练得到支持向量机，然后选取每个支持向量机对应的最优超平面附近的样本，即 $0 < \alpha_i < C$ 的样本数据 (x_i, y_i) 作为 Reduce 的输入，并在 Reduce 阶段训练一个新的支持向量机来作为最终使用的决策函数. 利用 MapReduce 训练支持向量机的算法如图 1 所示.

对于人脸识别这种多分类问题，利用本文算法

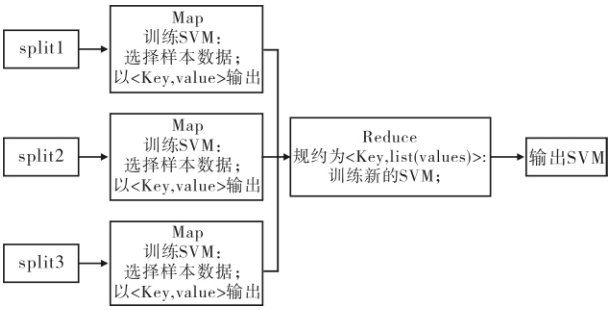

图 1　利用 MapReduce 训练支持向量机

Fig. 1　Using MapReduce to train Support Vector Machine

针对“一对多”方法下训练支持向量机的步骤如下：

Step 1　对含有类训练样本的数据标号并规约为 < key，value > 格式的数据，其中 key 值为样本的类别，value 为样本特征数据.

Step 2　将 < key，value > 格式的数据输入 Map 函数进行处理，在每个 Map 函数中对输入数据求解最优化问题得到个支持向量机，输出格式为 < key，value > 格式的中间数据，其中 key 为支持向量机类别，该类别对应被支持向量机分为正样本的类别，value 为经过标记的支持向量，标记为 1 表示在该支持向量对应的支持向量机中该支持向量对应的训练样本为正样本，标记为 - 1 则表示该支持向量对应的训练样本为负样本.

Step 3　对中间键值对数据进行 Partition 阶段操作将具有相同 key 值的数据发送到相同的 Reduce 节点进行处理.

Step 4　中间键值对数据传送到 Reduce 节点被排序规约为格式为 < key，list(values) > 的数据，其中 key 为支持向量机类别，list(values) 为从中间键值对数据中收集到的所有该类别对应的数据.

Step 5　Reduce 函数对 < key，list(values) > 格式的数据进行处理，通过求解最优化问题得到一个新的支持向量机，该支持向量机即用来识别 key 对应的人脸样本类别. Reduce 阶段执行完毕后，得到新的个支持向量机以 < key，value > 格式输出，其中 key 为支持向量机类别，value 为支持向量机对应的参数.

3.2　实验结果与分析

本文在 4 台 PC 机(CPU 2.2GHz 以上，2GB 内存) 上构建 Hadoop 集群，Hadoop 版本为 Hadoop0.21.0. 实验采用 ORL 标准人脸库作为数据集，ORL 人脸库中共有 40 个人，每人 10 幅图像，图像大小为 112 × 92 像素，共 400 幅人脸图像，每个人的不同人脸图像包含了表情、姿态和面部饰物的变化. 采用直

方图均衡化对图像进行预处理以减小环境、光照等条件的影响,选取数据集中每个人的前6幅图像作为训练集,剩下4张作为测试集进行实验,采用"一对多"分类法进行识别.

本文采用了RBF径向基函数作为SVM的核函数,如公式(7)所示,参数 $g=0.01$,$C=100$. 本文分别在单机以及2Map、3Map和4Map下进行实验. 图2显示了对ORL库中人脸图像进行直方图均衡化的效果.

$$K(x,y)=exp(-g\|x-y\|^2). \quad (7)$$

图2 ORL库中的人脸图像以及直方图均衡化效果

Fig. 2 A face picture of ORL and its histogram equalization

图3显示了不同实验条件下支持向量机的训练时间(设单机环境训练时间为100%为参照)及识别率.

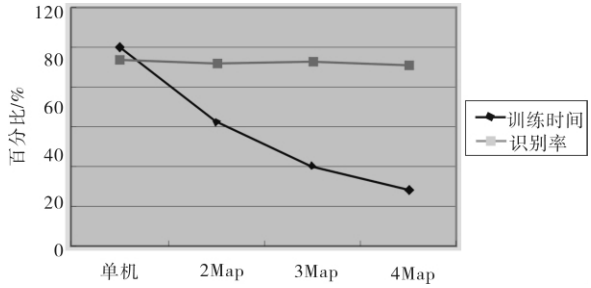

图3 不同环境下支持向量机的训练时间以及识别率

Fig. 3 Train time and recognition rate of Support Vector Machine in different conditions

由图3可见,在机器学习时间方面,采用MapReduce模型训练支持向量机的时间与不使用MapReduce模型的单机环境相比得到了较大幅度的降低,2Map,3Map和4Map情况下训练时间分别是单机情况下的62.2%,40.5%和28.5%. 图3中的识别率指标体现出单机的人脸识别率是93.1%,并行情况下2Map,3Map和4Map的识别率分别为91.9%,92.5%和90.6%. 并行环境下的识别率比单机环境有些微下降,其原因在于原来整体样本中的部分支持向量在分块处理中被丢弃. 本文方法在支持向量机的训练时间方面获得了明显的加速,但付出了1%~2%的识别率代价. 权衡移动互联网环境对人脸识别复杂计算的应用需求,在保证人脸识别率达到90%以上的前提下,本文侧重了机器学习时间的优化.

4 结语

本文利用MapReduce模型训练支持向量机进行人脸识别. 实验结果表明,该算法提升了支持向量机的训练速度,保证了人脸识别的准确率. Hadoop技术与MapReduce模型已经在移动互联网上获得了广泛应用,本文所采用的方法对于移动互联网环境下的人脸识别,特别是对识别效率有较高要求的人脸识别应用具有一定的实用价值. 下一步将分析并行度与识别率之间的关系,研究优化支持向量机性能的策略,力求在加速人脸识别运算的同时,进一步提高其识别率.

参 考 文 献

[1] Vapink V N. 统计学理论的本质[M]. 北京: 清华大学出版社,2000.

[2] 胡正平,张 晔. 结构风险最小化近邻分析解决大规模训练集支持向量机学习问题[J]. 信号处理,2007(1):161-164.

[3] Deng J, Ghemawat S. MapReduce: Simplied data processing on large clusters[C]// USENIX. Proceedings of the 6th USENIX Symposium on Operating System Design and Implementation (OSDI). New York: ACM Press, 2004: 137-150.

[4] 赵武锋. 人脸识别中特征提取方法的研究[D]. 杭州: 浙江大学,2009.

[5] 高全学,潘 泉,梁 彦,等. 基于描述特征的人脸识别研究[J]. 自动化学报,2006,32(3):386-391.

[6] 苏宏涛. 基于统计特征的人脸识别技术研究[D]. 西安: 西北工业大学,2005.

[7] 徐 仲,张凯院,陆 全,等. 矩阵论简明教程[M]. 北京: 科学出版社,2005: 118-123.

[8] Cortes C, Vapink V. Support vector networks[J]. Machine, 1995, 20: 273-297.

[9] Vaprink V N. Statistical learning theory[M]. New York: Wiley, 1998: 493-520.

[10] Krebel U. Pairwise classification and support vector machine[M]. Cambridge, USA: The MIT Press, 1999: 255-268.