ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.3, March 2008, pp.621−629
DOI: 10.3724/SP.J.1001.2008.00621

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于最大-最小相似度学习方法的文本提取*

付 慧[1,2], 刘峡壁[1,3+], 贾云得[1]

[1](北京理工大学 计算机科学与技术学院 智能信息技术北京市重点实验室,北京 100081)

[2](北京林业大学 信息学院,北京 100083)

[3](中国科学院 自动化研究所 模式识别国家重点实验室,北京 100080)

Text Extraction Based on Maximum-Minimum Similarity Training Method

FU Hui[1,2], LIU Xia-Bi[1,3+], JIA Yun-De[1]

[1](Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[2](School of Information Technology, Beijing Forestry University, Beijing 100083, China)

[3](National Laboratory of Pattern Recognition, Institute of Automation, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-86343158, Fax: +86-10-68940955, E-mail: liuxiabi@bit.edu.cn, http://www.mcislab.org.cn

Fu H, Liu XB, Jia YD. Text extraction based on maximum-minimum similarity training method. *Journal of Software*, 2008,19(3):621−629. http://www.jos.org.cn/1000-9825/19/621.htm

Abstract: This paper proposes a maximum-minimum similarity training algorithm to optimize the parameters in the effective method of text extraction based on Gaussian mixture modeling of neighbor characters. The maximum-minimum similarity training (MMS) methods optimize recognizer performance through maximizing the similarities of positive samples and minimizing the similarities of negative samples. Based on this approach to discriminative training, it defines the objective function for text extraction, and uses the gradient descent method to search the minimum of the objective function and the optimum parameters for the text extraction method. The experimental results of text extraction show the effectiveness of MMS training in text extraction. Compared with the maximum likelihood estimation of parameters from expectation maximization (EM) algorithm, the training results after MMS has the performance of text extraction improved greatly. The recall rate of 98.55% and the precision rate of 93.56% are achieved. The experimental results also show that the maximum-minimum similarity (MMS) training behaves better than the commonly used discriminative training of the minimum classification error (MCE).

Key words: text extraction; Gaussian mixture modeling; discriminative training; maximum-minimum similarity training; minimum classification error training

摘 要: 应用最大-最小相似度(maximum-minimum similarity,简称 MMS)学习方法,对基于高斯混合模型的文本区

* Supported by the National Natural Science Foundation of China under Grant No.60473049 (国家自然科学基金); the National Basic Research Program of China under Grant No.2006CB303105 (国家重点基础研究发展计划(973)); the Excellent Young Scholars Research Fund of Beijing Institute of Technology of China under Grant No.2006Y1202 (北京理工大学优秀青年教师资助计划)

Received 2006-04-18; Accepted 2006-10-10

域提取方法中的有关参数进行优化.该学习方法通过最大化正样本相似度和最小化反样本相似度获得最佳分类能力.根据这种判别学习思想,建立了相应的目标函数,并利用最速梯度下降法寻找目标函数最小值,以得到文本区域提取方法的最优参数集合.文本区域提取实验结果表明:在用期望最大化(expectation maximization,简称 EM)算法获得参数的极大似然估计值后,使用最大-最小相似度学习方法,使文本提取综合性能明显提高,开放实验的召回率和准确率分别达到 98.55%和 93.56%.在实验中,最大-最小相似度学习方法的表现还优于常用的判别学习方法——最小分类错误(minimum classification error,简称 MCE)学习方法.

关键词: 文本提取;高斯混合模型;判别学习;最大-最小相似度学习;最小分类错误学习

中图法分类号: TP391 **文献标识码**: A

随着计算机技术以及通信技术的飞速发展,越来越多的文字信息以图像的形式出现.因此,获取图像和视频中的文字逐渐成为研究热点之一,在基于内容的图像检索、自动视频记录、光学字符识别(optical character recognition,简称 OCR)等许多方面起着关键性的作用.图像中的文字在字体、大小、颜色和排列上变化多端,背景也千差万别,使得从图像中自动提取文字非常困难.

已有的文本提取方法主要包括基于区域的方法和基于纹理的方法两大类[1].基于区域的方法利用文本区域与背景区域在颜色或灰度特性上的不同来进行提取,通常采用自底向上的策略,即首先辨识子结构,然后通过子结构的合并来标记文本区域[2−4].基于纹理的方法利用文本区域与背景区域在纹理特性上的不同来进行提取,可以采用 Gabor 变换、小波变换等方法实现[5].目前,文字在字体、颜色、排列等方面的变化对文本提取结果有较大影响,同时也缺乏描述多语种文本对象的统一模型,提取不同语种文本须采用各自不同的特征和方法.

近年来,人们提出基于相邻字符区域统计模型的文本提取方法,显示出良好的发展前景.Zhang 和 Chang[6]提出了字符串的马尔可夫随机场(Markov random field,简称 MRF)模型,用于场景图像中颜色一致的英文文本提取,取得了较好的效果.本文将 3 个相邻字符之间的关系特征建模为高斯混合模型(Gaussian mixture model,简称 GMM),并根据该模型区分字符与非字符.在此基础上,提出一种多语种文本提取方法[7],可以提取颜色不一致的文本和呈曲线排列的文本.与 Zhang 和 Chang[6]的方法相比,该方法中所采用的 GMM 模型与 MRF 模型一样有效,但学习和提取效率更高.同时,利用 Voronoi 分割确定图像区域之间的邻接关系,避免了经验性的邻域选择方法,具有更高的鲁棒性.该方法中,GMM 模型参数通过期望最大化算法(expectation maximization,简称 EM)学习得到,使文本提取准确率较高,但召回率偏低.EM 属于数据学习方法,对分类的优化是间接的.与数据学习不同,以最小分类错误学习(minimum classification error,简称 MCE)为代表的判别学习方法直接以最优化分类能力为目标,目前已在文字识别、语音识别等领域得到成功应用[8−12].

本文采用一种新的判别学习方法——最大-最小相似度(maximum-minimum similarity,简称 MMS)学习[13],在 EM 算法所得参数结果的基础上继续学习,优化有关参数.最大-最小相似度学习方法通过最大化正样本相似度和最小化反样本相似度获得最佳分类能力.基于这种判别学习的思想,我们建立了相应的目标函数,并利用最速梯度下降法寻找目标函数最小值,以得到文本区域提取方法的最优参数集合.文本区域提取实验结果显示:使用最大-最小相似度学习方法后,文本提取准确率基本不变,但召回率由 80.39%提高到 98.55%,明显改善了方法的综合性能,提高了方法的鲁棒性.在实验中,MMS 学习方法的表现还优于 MCE 学习方法.与 MCE 学习方法相比,在封闭测试中,MMS 学习方法对应的准确率下降 0.8%,但召回率增长 13.2%;在开放测试中,准确率下降 4%,但召回率增长 24%.实验结果表明,用于文本提取的最大-最小相似度学习方法是有效的.

1 基于高斯混合模型的文本提取方法

文本提取的关键在于如何区分字符区域与非字符区域.我们利用相邻字符区域之间的关系特征来解决这个问题.首先,与非字符区域比较,相邻字符区域具有以下关系特征:

x_1:字符重心间距的一致性.图像中的大多数文本串,无论是呈直线排列还是呈曲线排列,相邻字符之间的距离都是近似相等的.设$\{A,B,C\}$表示 3 个相邻字符的重心,不失一般性,我们假设$\|A-B\|\leq\|A-C\|$,并且

$||B-C||\leq||A-C||$,则 3 个相邻字符之间的间距一致性可以度量为

$$x_1=\frac{\|A-B\|}{\|B-C\|}.$$

x_2:字符区域面积的一致性.与相邻字符重心间距的一致性相似,相邻字符区域的面积常常也是近似相等的.设 $Area_A$,$Area_B$ 和 $Area_C$ 分别表示 3 个相邻字符的区域面积,则 3 个相邻字符之间的面积一致性可以度量为

$$x_2=\frac{\max(Area_A,Area_B,Area_C)}{\min(Area_A,Area_B,Area_C)}.$$

x_3:区域密度.字符区域的密度与非字符区域的密度之间存在着差异.我们计算 3 个相邻字符的平均区域密度作为第三维特征

$$x_3=\frac{(density_A,density_B,density_C)}{3}.$$

综合起来,采用 $\boldsymbol{x}=\{x_1,x_2,x_3\}$作为 3 个相邻区域对应的特征矢量.我们假设 3 个相邻字符区域的 $\boldsymbol{x}$ 服从高斯混合分布[14].设 C 表示 3 个字符相邻的情况,其中,K 是高斯混合模型的高斯成分个数.$\boldsymbol{\mu}_k$ 是特征向量 $\boldsymbol{x}$ 的均值向量,$\boldsymbol{\Sigma}_k$ 是特征向量 $\boldsymbol{x}$ 的协方差矩阵,w_k 是第 k 个高斯成分在高斯混合模型中的权值,d 是特征向量 $\boldsymbol{x}$ 的维数,则有

$$p(x\mid C)=\sum_{k=1}^{K}w_kN(x\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

其中, $N(x\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=(2\pi)^{-\frac{d}{2}}\mid\boldsymbol{\Sigma}_k\mid^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_\kappa^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)$.

为了简化计算,这里认为$\boldsymbol{\Sigma}_k$ 是对角阵,例如 $\boldsymbol{\Sigma}_k=[\sigma_{kj}]_{j=1}^3$.根据贝叶斯公式,我们进一步得到

$$P(C\mid\boldsymbol{x})=\frac{p(\boldsymbol{x}\mid C)P(C)}{p(\boldsymbol{x})}\tag{1}$$

显然,用概率 $P(C|\boldsymbol{x})$来判别 $\boldsymbol{x}$ 是否属于 3 个字符相邻的情况是合理的.在式(1)中,$p(\boldsymbol{x}|C)$可以从 3 个相邻字符的样本中估计出来.但反例(非文本)的情况很复杂,因此,学习 $P(C)$和 $p(\boldsymbol{x})$很困难.然而,如果假设 $\boldsymbol{x}$ 的分布是均匀的,则 $P(C)$和 $p(\boldsymbol{x})$对于所有的 $\boldsymbol{x}$ 来说都是恒定的,于是有

$$P(C|\boldsymbol{x})\propto p(\boldsymbol{x}|C)\tag{2}$$

根据式(2),可以用一个以 $p(\boldsymbol{x}|C)$为自变量的,光滑、单调递增且值域在[0,1]之间的函数来模拟 $P(C|\boldsymbol{x})$,我们将这样的函数值称为伪概率.以下函数符合计算伪概率的要求,其中,β为一个正数,

$$\rho(p(\boldsymbol{x}|C))=1-\exp(-\beta p(\boldsymbol{x}|C))\tag{3}$$

于是,对于任意3个相邻区域,计算其特征矢量$\boldsymbol{x}=\{x_1,x_2,x_3\}$,然后用式(3)计算相应的伪概率,如果伪概率的值大于0.5,则认为这3个相邻区域均为字符区域;否则认为其中至少存在一个非字符区域,并不再对其中某一区域的特性进行判断.因此,上述 GMM 模型需用于由 3 个及 3 个以上字符所组成的文本.

为了将上述相邻字符区域的 GMM 模型用于文本提取,需要获得图像中每个字符对应的区域.因此,首先对输入图像进行文本初始定位和二值化.这里采用了基于边缘像素聚类的文本区域初始定位和二值化方法[15],可以保证二值图像中文本的完整性.然后,对二值图像做形态学闭运算,使二值图像中的每个字符对应一个单独的连通成分.考虑到字符在图像中可能出现的多种形态,我们使用具有不同大小和方向的结构元素执行闭运算,从而获得多个新二值图像.在这些新二值图像和原始二值图像中,基于相邻字符区域的 GMM 模型将连通成分标注为字符或非字符,从而完成文本提取.如果新二值图像中的连通成分被标注为字符,那么,在原始二值图像中确定该连通成分对应的区域,并将原始二值图像中这一区域内的所有连通成分都标注为字符.关于上述基于高斯混合模型的文本提取方法的更多细节,参见文献[7].

2 用最大-最小相似度学习优化文本提取方法

在第 1 节所述文本区域提取方法中,未知参数的集合为

$$\boldsymbol{\Lambda}=\{w_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\beta\},\ k=1,\ldots,K.$$

我们用期望最大化算法获得 GMM 模型中有关参数的极大似然估计值,用实验方法确定其他参数,得到的文本区域提取方法具有较高的准确率,但召回率偏低,仅为 80.39%.为了提高方法综合性能,我们利用最大-最小相似度学习方法[13]在 EM 算法所得到的极大似然估计值基础上继续学习,优化有关参数.

最大-最小相似度学习属于判别学习,通过最大化正样本相似度和最小化反样本相似度,获得最佳分类能力.正样本是指属于某一类别的样本,反样本是指不属于某一类别的样本.对于一个输入对象,如果该对象与同类模型的相似度为 1,与不同类模型的相似度为 0,则分类能力最好.因此,在最大-最小相似度学习方法中,学习目标是使相似度在同类时趋近于 1,在异类时趋近于 0.本文中,每个样本是 3 个相邻连通成分的集合.正样本是指文本样本,其中 3 个相邻连通成分均为字符.反样本是指非文本样本,其中 3 个相邻连通成分中至少有 1 个不是字符.设 $\hat{\boldsymbol{X}}$ 表示正样本, $\overline{\boldsymbol{X}}$ 表示反样本,$f(\boldsymbol{X};\boldsymbol{\Lambda})$表示计算相似度的函数,则 MMS 学习方法的目标函数为

$$F(\boldsymbol{\Lambda})=\Sigma[f(\hat{\boldsymbol{X}};\boldsymbol{\Lambda})-1]^2+\Sigma[f(\overline{\boldsymbol{X}};\boldsymbol{\Lambda})]^2 \tag{4}$$

最优参数集合为

$$\boldsymbol{\Lambda}^*=\arg\min_{\boldsymbol{\Lambda}}F(\boldsymbol{\Lambda})=\arg\min_{\boldsymbol{\Lambda}}(\Sigma[f(\hat{\boldsymbol{X}};\boldsymbol{\Lambda})-1]^2+\Sigma[f(\overline{\boldsymbol{X}};\boldsymbol{\Lambda})]^2).$$

如第 1 节所述,本文中,

$$f(\boldsymbol{x};\boldsymbol{\Lambda})=1-\exp\left(-\beta\left(\sum_{k=1}^{K}w_kN(\boldsymbol{x}\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)\right) \tag{5}$$

式(5)中,部分参数取值具有约束,因此,我们对参数进行变换[8],将约束优化问题变成无约束优化问题.参数变换过程为

(1) 因为$\sum w_k=1$,所以$w_k\to\tilde{w}_k$: $w_k=\dfrac{e^{\tilde{w}_k}}{\sum e^{\tilde{w}_k}}$,

(2) 因为$\beta>0$,所以$\beta\to\tilde{\beta}$: $\tilde{\beta}=\ln\beta$,

(3) 因为$\sigma_{kj}>0$,所以$\sigma_{kj}\to\tilde{\sigma}_{kj}$: $\tilde{\sigma}_{kj}=\ln\sigma_{kj}$.

最后,采用最速梯度下降法[16]求解式(4)的最小值,得到最优参数集合.最速梯度下降法沿函数的梯度方向迭代更新参数.设 s_t 为第 t 次迭代时的步长因子,$\boldsymbol{\Lambda}_t$ 为第 t 次迭代时的参数集合,$\nabla F(\boldsymbol{\Lambda})$表示 $F(\boldsymbol{\Lambda})$对$\boldsymbol{\Lambda}$的偏导,则

$$\boldsymbol{\Lambda}_{t+1}=\boldsymbol{\Lambda}_t-s_t\nabla F(\boldsymbol{\Lambda}_t)=\boldsymbol{\Lambda}_t-s_tg_t \tag{6}$$

根据式(6),具体学习流程是:

步骤 1. 利用所有文本样本(正样本)和非文本样本(反样本)计算目标函数对各个参数的偏导数.设ψ表示参数集合中的任意参数,则计算公式为

$$\frac{\partial F}{\partial\psi}=2\Sigma\left[((f(\hat{\boldsymbol{X}};\boldsymbol{\Lambda})-1)+f(\overline{\boldsymbol{X}};\boldsymbol{\Lambda}))\frac{\partial F(\boldsymbol{\Lambda})}{\partial\psi}\right] \tag{7}$$

这里,因为 $\dfrac{\partial f(\hat{\boldsymbol{X}};\boldsymbol{\Lambda})}{\partial\psi}$ 与 $\dfrac{\partial f(\overline{\boldsymbol{X}};\boldsymbol{\Lambda})}{\partial\psi}$ 的计算公式相同,所以统一表示为 $\dfrac{\partial f(\boldsymbol{\Lambda})}{\partial\psi}$,具体计算公式见第 2.1 节.

步骤 2. 根据偏导数计算步长因子.

步骤 3. 按式(6)更新参数.

步骤 4. 重复以上步骤直到收敛或达到最大迭代次数为止.设ε表示极小值,则收敛条件为

$$\|g_t\|=\left(\Sigma\left(\frac{\partial F}{\partial\psi}\right)^2\right)^{\frac{1}{2}}\le\varepsilon.$$

2.1 参数偏导的计算公式

在第 1 节所述文本提取方法中,高斯成分个数为 K,则针对每个具体参数,式(7)中 $\dfrac{\partial f(\boldsymbol{\Lambda})}{\partial\psi}$ 的计算公式分别列出,其中,式(5)中的 $N(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$简化为 $N_k(\boldsymbol{x})$,k 为 GMM 中高斯成分的序列号,j=1,2,3 是在第 1 节中介绍的特征

向量的元素成分的序列号:

$$\frac{\partial f}{\partial \tilde{w}_k}=\beta N_k(\boldsymbol{x})w_k(1-w_k)\mathrm{e}^{\left(-\beta\left(\sum_{i=1}^{K}w_iN_i(\boldsymbol{x})\right)\right)},$$

$$\frac{\partial f}{\partial \tilde{\beta}}=\left(\sum_{i=1}^{K}w_iN_i(\boldsymbol{x})\right)\mathrm{e}^{\left(-\beta\left(\sum_{i=1}^{K}w_iN_i(\boldsymbol{x})\right)+\tilde{\beta}\right)},$$

$$\frac{\partial f}{\partial \mu_{kj}}=\beta w_kN_k(\boldsymbol{x})\left(\frac{x_j-\mu_{kj}}{\sigma_{kj}}\right)\mathrm{e}^{\left(-\beta\left(\sum_{i=1}^{K}w_iN_i(\boldsymbol{x})\right)\right)},$$

$$\frac{\partial f}{\partial \tilde{\sigma}_{kj}}=\beta w_kN_k(\boldsymbol{x})\left(\frac{(x_j-\mu_{kj})^2}{2\sigma_{kj}^2}\right)\mathrm{e}^{\left(-\beta\left(\sum_{i=1}^{K}w_iN_i(\boldsymbol{x})\right)+\tilde{\sigma}_{kj}\right)}.$$

2.2 步长s_t的选取

步长因子的计算方法采用 Armijo-Goldstein 不精确搜索方法[16],具体步骤如下.为简洁起见,均省略表示参数更新序号的下标 t.

步骤 1. 选定初始点 s_0=1,给出 $\rho\in\left(0,\frac{1}{2}\right)$,$l>1$,令 a_0=0;b_0=+∞;k=0.

步骤 2. 令 $\varphi(s)=F(\boldsymbol{\Lambda}-sg)$,如果 $\varphi(s_k)\le\varphi(0)-\rho s_k\|g\|^2$,则转步骤 3;否则,令 $a_{k+1}=a_k,b_{k+1}=s_k$,转步骤 4.

步骤 3. 如果 $\varphi(s_k)\ge\varphi(0)-(1-\rho)s_k\|g\|^2$,则停止迭代,输出 s_k;否则,令 $a_{k+1}=s_k,b_{k+1}=b_k$,
若 $b_{k+1}<+\infty$,则转步骤 4;否则,令 $s_{k+1}=ls,k=k+1$,转步骤 2.

步骤 4. 取 $s_{k+1}=\frac{a_{k+1}+b_{k+1}}{2}$,令 $k=k+1$,转步骤 2.

3 实验结果与分析

我们应用本文所述文本提取方法进行了从图像中提取中英文文本的实验,并在实验中将最大-最小相似度学习方法与极大似然学习方法和最小分类错误学习方法进行了比较.极大似然学习是经典的数据学习方法,本文中采用期望最大化算法来实现;最小分类错误学习[8]则是一种常用的判别学习方法.实验中共采用了 250 幅图像,其中 220 幅用于训练,另外 30 幅用于测试.除了 10 幅测试图像来自于 ICDAR 2003 测试集[17]以外,其余的训练和测试图像均来自于我们自己建立的数据库.在测试图像集中包含 5 幅不含文本区域的图像,以验证算法的鲁棒性.

我们从训练图像集中手工提取出 427 个英文训练样本、531 个中文训练样本和 5 973 个非文本训练样本,用于训练方法中的有关参数.所有样本均由 3 个连通成分组成,其中,文本样本中 3 个连通成分均为字符,非文本样本中至少有 1 个不是字符.首先根据英文和中文训练样本集,采用期望最大化算法获得 GMM 模型中有关参数的极大似然估计值,同时,通过实验方法设定 β=1.020 和 K=3.然后根据所有训练样本,包括文本样本和非文本样本,分别利用最小分类错误学习方法和最大-最小相似度学习方法继续学习 GMM 模型参数和 β.这样,共获得 3 组参数.我们分别在这 3 组参数的基础上进行了文本提取的封闭和开放实验,相应实验结果在表 1 中列出.需要指出的是,由于 EM 算法只能利用文本样本,因此,表 1 中与 EM 算法对应的训练图像的个数为 200,而不是 220.

表 1 采用召回率 R 和准确率 P 评估文本提取算法的性能.召回率是指提取出的文本区域个数与图像中实际文本区域个数的比例,它反映了算法漏检文本区域的情况,召回率越高,漏检率越低.准确率是指提取出的文本区域个数与提取出的所有区域个数的比例,它反映了算法将非文本区域误检为文本区域的情况,准确率越高,误检率越低.设 $\boldsymbol{L}$ 表示图像中字符对应的连通成分集合,$\boldsymbol{I}$ 表示提取方法所确定的对应于字符的连通成分集合,则召回率和准确率计算公式为[18]

$$P=\frac{|\boldsymbol{L}\cap\boldsymbol{I}|}{|\boldsymbol{I}|},\ R=\frac{|\boldsymbol{L}\cap\boldsymbol{I}|}{|\boldsymbol{L}|}.$$

Table 1 The different experimental results after EM, MCE & MMS training

表 1 EM,MCE 和 MMS 学习后的不同实验结果

Parameters / Learning method (Image)	$\|\boldsymbol{L}\|$	$\|\boldsymbol{I}\|$	$\|\boldsymbol{L}\cap\boldsymbol{I}\|$	$\|\boldsymbol{NT}\|$	P (%)	R (%)
EM (200 training)	1 296	1 096	1 053	59	96.08	81.25
EM (30 test)	413	351	332	116	94.59	80.39
MCE (220 training)	1 417	1 276	1 228	59	96.24	86.66
MCE (30 test)	413	337	328	116	97.33	79.42
MMS (220 training)	1 417	1 457	1 390	152	95.40	98.09
MMS (30 test)	413	435	407	116	93.56	98.55

此外,表 1 中的$|\boldsymbol{NT}|$表示非文本块的个数.

表 1 中的数据表明,从文本提取的综合性能看,MMS 方法的训练结果明显优于 EM 算法和 MCE 方法的训练结果.虽然准确率略有下降,但召回率大幅提升,提高了方法的鲁棒性.与 EM 算法相比,在封闭测试中,MMS 学习对应的准确率下降了 0.7%,但召回率增长了 20.7%;在开放测试中,准确率下降了 1.1%,但召回率增长了 22.6%.EM 算法是一种数据学习方式,对文本提取的优化是间接的,并且它只能根据正样本学习 GMM 模型参数,因此,使文本提取准确率较高而召回率偏低.MMS 是一种判别学习方法,学习目标直接服务于文本提取,并且能够从正、反两类样本中学习,同时,EM 算法中不能学习到的参数,如β,在 MMS 中也能进行学习.因此,在采用 MMS 学习算法以后,准确率和召回率趋于平衡,文本提取方法的综合性能明显提高.作为一种判别学习方法,MCE 方法的训练结果稍好于 EM 算法的训练结果,但 MMS 方法的训练结果更好.与 MCE 学习相比,在封闭测试中,MMS 学习对应的准确率下降了 0.8%,但召回率增长了 13.2%;在开放测试中,准确率下降了 4%,但召回率增长了 24%.MCE 方法与 MMS 方法的分类思想不同.MCE 方法是从样本的角度出发进行学习,试图尽可能地区分某一样本对应于真实类别的判别值与该样本对应于其他类别的判别值.MMS 方法则是从类的角度出发进行学习,试图尽可能地区分某一类的所有样本对应于该类别的判别值与其他类的所有样本对应于该类别的判别值.这种学习策略上的不同,造成了 MMS 学习方法对训练样本的依赖更小,从而在实验中表现出更好的性能.事实上,如前所述,非文本训练样本的个数明显多于文本训练样本的个数,而 MCE 学习后得到了很高的准确率和较低的召回率,反映出非文本样本对训练结果的影响较大.

图 1 显示了对应于 3 种不同学习方法的部分实验结果,图 1(a)~图 1(c)分别为 EM 算法、MCE 方法和 MMS 方法学习后的提取结果,提取出的文本区域均用矩形框来表示.这些例子也展示出 MMS 学习明显提高了文本提取的召回率.但经过 MMS 学习后,也使部分非文本被误认为文本,图 2 显示了一些例子,其中箭头所指示的矩形框表示误认为文本的非文本区域.这也是我们下一步的研究重点,预计可以通过增加学习样本和迭代次数得到解决.

4 结 论

本文应用最大-最小相似度学习方法对基于相邻字符区域高斯混合模型的文本区域提取方法进行学习,优化其中的参数.最大-最小相似度学习属于判别学习,通过最大化正样本相似度和最小化反样本相似度获得最佳分类性能.在这种学习思想基础上,我们建立了用于优化文本区域提取方法中有关参数的目标函数,然后用最速梯度下降法求解.实验结果表明,在通过期望最大化算法获得参数的极大似然估计值后,应用最大-最小相似度学习方法继续学习和优化有关参数,使文本提取召回率大幅度提升,同时准确率基本不变,明显改善文本提取综合性能.开放实验对应的召回率和准确率分别达到了 98.55%和 93.56%.在实验中,最大-最小相似度学习方法的表现不仅优于期望最大化算法,还优于常用的判别学习方法-最小分类错误学习方法.与最小分类错误学习相比,在封闭测试中,最大-最小相似度学习对应的准确率下降了 0.8%,但召回率增长了 13.2%;在开放测试中,准确率下降了 4%,但召回率增长了 24%.实验结果表明,基于最大-最小相似度学习方法的文本提取是有效的.

常用的判别学习方法除了 MCE 以外,还包括最大互信息方法(maximum mutual information,简称 MMI)[19]

和支持向量机(support vector machines,简称 SVM)[20].MCE 和 MMI 是从样本的角度考虑类别可分性,试图尽可能区分每一个样本对应于真实类别的判别值与该样本对应于其他类别的判别值.而 MMS 则是从类的角度考虑类别可分性,试图尽可能地区分每一类的所有样本对应于该类别的判别值与其他类的所有样本对应于该类别的判别值.这种类别区分策略的不同使得 MMS 对训练数据的依赖更小,这一点已在本文所述实验中得到了一定的证实.类似于支持向量机,MMS 也试图寻找类之间的最大间隔,但二者实现这一目标的途径不同.SVMs 在输入空间或高维特征空间中构造最优分界面,但 MMS 考虑最优相似性度量,这使得 MMS 可用于任何具有可微相似性函数的分类器.由于“相似性”是进行分类的自然量度,其他常用量度(如后验概率、距离等)可转化为相应的相似性,因此,与 SVMs 相比,MMS 的应用范围更广.如在结构识别方法中,可应用 MMS,但难以应用 SVMs.

(a) Corresponding with the EM algorithm (a) 对应于 EM 算法 (b) Corresponding with the MCE algorithm (b) 对应于 MCE 算法 (c) Corresponding with the MMS algorithm (c) 对应于 MMS 算法

Fig.1 The illustration of different text extraction results corresponding with three training methods (extracted texts are contained in the rectangles)

图 1 对应于 3 种学习算法的不同文本提取结果示例(结果用矩形框来表示)

Fig.2 The examples of false text extraction after MMS training (false results are indicated by arrowheads)

图 2 MMS 学习后文本提取误检示例(箭头所示为误检区域)

References:

[1] Jung K, Kim KI, Jain AK. Text information extraction in images and video: A survey. Pattern Recognition, 2004,37(5):977−997.

[2] Chen YX, Liu CS, Ding XQ. Character extraction in complex color document images. Journal of Chinese Information Processing, 2003,17(5):55−59 (in Chinese with English abstract).

[3] Liu Y, Xue XY, Lu H, Guo YF. A video text detecting method based on edge detection and line features. Chinese Journal of Computers, 2005,28(3):427−433 (in Chinese with English abstract).

[4] Jain AK, Yu B. Automatic text location in images and video frames. Pattern Recognition, 1998,31(12):2055−2076.

[5] Wu V, Manmatha R, Riseman EM. TextFinder: An automatic system to detect and recognize text in images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1999,21(11):1224−1229.

[6] Zhang DQ, Chang SF. Learning to detect scene text using a higher-order MRF with belief propagation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW 2004). Washington: IEEE, 2004. 101−108. http://portal.acm.org/citation.cfm?id=1032637.1032992&coll=GUIDE&dl=GUIDE&CFID=10222107&CFTOKEN=81753630

[7] Fu H, Liu XB, Jia YD. Gaussian mixture modeling of neighbor characters for multilingual text extraction in images. In: Proc. of the IEEE Int'l Conf. on Image Processing 2006 (ICIP 2006). Atlanta: IEEE, 2006. 3321−3324. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4107281

[8] Juang BH, Chou W, Lee CH. Minimum classification error rate methods for speech recognition. IEEE Trans. on Speech and Audio Processing, 1997,5(3):257−265.

[9] Yu H, Gao JF, Bu FL. One new discriminative training method for language modeling. Chinese Journal of Computers, 2005,28(10):1708−1715 (in Chinese with English abstract).

[10] Sun GL, Liu JF, Tang XL, Shi DM, Zhao W. Active discriminant function for handwriting recognition. Journal of Software, 2005, 16(4):523−532 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/523.htm

[11] Han JQ, Gao W. Robust speech recognition method based on discriminative environment feature extraction. Journal of Computer Science and Technology, 2001,16(5):458−464.

[12] Zhang R, Ding XQ. Minimum classification error training for handwritten character recognition. In: Proc. of the 16th Int'l Conf. on Pattern Recognition, Vol.1. Quebec: IEEE Computer Society, 2002. 580−583. http://portal.acm.org/citation.cfm?id=839290.842596&coll=GUIDE&dl=GUIDE

[13] Liu XB, Jia YD, Chen XF, Deng Y, Fu H. Max-Min posterior pseudo-probabilities estimation of posterior class probabilities to maximize class separability. Technical Report, BIT-CS-TP-20060002. 2006. http://www.mcislab.org.cn/technical%20reports/index.asp

[14] Moerland P. A comparison of mixture models for density estimation. In: Proc. of the Int'l Conf. on Artificial Neural Networks (ICANN '99), vol.1. IEE, 1999. 25−30.

[15] Fu H, Liu XB, Jia YD. Edge-Pixels clustering for text area extraction. Journal of Computer-Aided Design and Computer Graphics, 2006,18(5):729−734 (in Chinese with English abstract).

[16] Yuan YX, Sun WY. Optimization Theory and Methods. Beijing: Science Press, 2003 (in Chinese).

[17] Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R. Icdar 2003 robust reading competitions. In: Proc. of the 7th Int'l Conf. on Document Analysis and Recognition. Edinburgh: IEEE, 2003. 682−687. http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/8701/27545/01227749.pdf

[18] Karatzas D, Antonacopoulos A. Text extraction from Web images based on a split-and-merge segmentation method using colour perception. In: Proc. of the 17th Int'l Conf. on Pattern Recognition (ICPR 2004). Cambridge: IEEE, 2004. 634−637. http://portal.acm.org/citation.cfm?id=1018428.1020791

[19] Vertanen K. An overview of discriminative training for speech recognition. Technical Report, 2004. http://www.inference.phy.cam.ac.uk/kv227/papers/Discriminative_Training.pdf

[20] Byun H, Lee SW. Applications of support vector machines for pattern recognition: A survey. In: Proc. of the 1st Int'l Workshop on Pattern Recognition with Support Vector Machines. Niagara Falls: Springer-Verlag, 2002. 213−236. http://portal.acm.org/citation.cfm?id=647230.719394

附中文参考文献:

[2] 陈又新,刘长松,丁晓青.复杂彩色文本图像中字符的提取.中文信息学报,2003,17(5):55−59.

[3] 刘洋,薛向阳,路红,郭跃飞.一种基于边缘检测和线条特征的视频字符检测算法.计算机学报,2005,28(3):427−433.

[9] 于浩,高剑峰,步丰林.一种新的语言模型判别训练方法.计算机学报,2005,28(10):1708−1715.

[10] 孙广玲,刘家锋,唐降龙,石大明,赵巍.基于主动判别函数的手写体识别.软件学报,2005,16(4):523−532. http://www.jos.org.cn/1000-9825/16/523.htm

[15] 付慧,刘峡壁,贾云得.用于文本区域提取的边缘像素聚类方法.计算机辅助设计与图形学学报,2006,18(5):729−734.

[16] 袁亚湘,孙文瑜.最优化理论与方法.北京:科学出版社,2003.

付慧(1978－),女,吉林农安人,博士,讲师,主要研究领域为图像处理,模式识别.

刘峡壁(1972－),男,博士,讲师,主要研究领域为模式识别,机器学习,多媒体信检索.

贾云得(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机视觉,人工智能,智能系统.