ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 4,Jul. 2021,pp. 799－811
DOI:10. 16337/j. 1004-9037. 2021. 04. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于标签分布构造的异构数据集年龄估算

王军祥[1]，吴　伶[2]

(1. 福建船政交通职业学院信息与智慧交通学院,福州 350007;2. 福州大学数学与计算机科学学院,福州 350108)

摘　要：现有年龄估算方法的性能度量主要是基于训练集与测试集独立同分布的假设。为了能更好地符合实际场景以及更好地评估年龄估算方法的泛化性能,提出一种异构数据集评估协议,即在年龄估算时更关注训练集与测试集具有的不同分布和特征情况。此外,为了提高基于卷积神经网络的年龄估算方法的拟合能力,在充分考虑相邻年龄特性的基础上,通过将年龄估算问题建模为基于高斯模型的标签分布学习,提出一种新颖的损失函数。理论分析与实验结果皆说明本文方法的有效性与鲁棒性。

关键词：年龄估算;独立同分布;卷积神经网络;标签分布;损失函数

中图分类号：TP391　　**文献标志码**：A

Age Estimation for Isomerism Data Sets Based on Label Distribution Construction

WANG Junxiang[1], WU Ling[2]

(1. School of Information and Intelligent Transportation, Fujian Chuanzheng Communications College, Fuzhou 350007, China;
2. Department of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: Existing age estimation methods of performance measurement are mainly based on the training set and testing set of independent identically distributed hypothesis. In order to better conform to the actual scene and better assess the age estimation method of generalization performance, a kind of heterogeneous data sets to evaluate agreement is put forward, i.e. paying more attention to the training set and test set with different distribution and characteristics. In addition, in order to improve the age estimation method based on convolution neural network fitting ability, on the basis of fully considering the adjacent age characteristics, a new theory of loss function analysis is proposed through modeling the age estimation problem as the label distribution study based on Gauss model. Theoretical analysis and experimental results show the effectiveness and robustness of the proposed method.

Key words: age estimation; independent homology distribution; convolutional neural network; label distribution; loss function

引　　言

基于面部图像的年龄估算旨在找到一种可将面部图像映射到其相应的年龄标签函数。随着诸如安全控制、社交媒体和人机交互等多种实际需求的不断增长,年龄估算越来越受到人们的关注。但由

基金项目:国家自然科学基金青年基金(62002063)资助项目。
收稿日期:2020-12-20;**修订日期**:2021-03-29

于表情、光照、性别、种族、基因、居住环境与生活方式等许多内在和外在因素的存在，年龄估算问题一直非常具有挑战性。

近年来年龄估算问题研究取得一定的进展，但现有的大多数研究都使用同构数据集年龄估算评估协议，即假定在训练和测试阶段中使用的所有图像都是在相似条件下获得的。但基于同构数据集学习到的模型可能会偏向训练集中图像的特征及分布，从而导致在条件完全不同的情况下年龄估算性能较差。因此在实际的应用中更应该关注异构数据集的年龄估算，即训练集与测试集应具有不同的分布和特征，这样训练后的模型便完全不了解目标数据集的特征，更符合实际场景。

异构数据集评估协议对现有年龄估算方法在实际应用中的效果提出了更高的要求，即要求训练好的模型不仅要在同构数据集下能够准确估算面部图像的年龄，还要在异构数据集下有效地工作。但为了能更好地符合实际场景，并且更好地评估年龄估算方法的泛化性能，本文提出一种异构数据集评估协议。

随着机器学习技术的发展，已有文献提出许多基于深度学习的年龄估算方法。文献[1]提出了一种非线性回归算法，利用分而治之的策略来考虑年龄的位次信息。文献[2]通过训练一系列二元卷积神经网络(Convolutional neural network，CNN)模型获得年龄标签的顺序信息，其中每个二元分类器用来判定输入面部图像的年龄是否超过特定年龄，最终年龄值通过计算CNN输出的总和而得出。尽管基于回归的方法比较直观，但是性能常常不尽如人意。基于分类的方法[3-5]将年龄估算问题建模为一个多类别的分类问题，并将不同的年龄视为独立的类别。在训练阶段，这些方法尝试使用交叉熵(Cross entropy，CE)损失函数来学习判别性特征。文献[6-7]对CE损失函数附加了不同项的正则化，从而惩罚预测年龄与真实年龄之间的差异，其中文献[7]附加了均值方差正则化项，均值正规化惩罚了预测年龄与真实年龄之间的均值差异，而方差正则化项惩罚它们之间的方差差异。文献[8]提出将年龄标签编码为概率分布，以此将年龄标签的局部相关性引入训练过程，这种方法将年龄估计问题建模为分布学习问题。

鉴于损失函数的选择会对域泛化(异构数据集测试)产生巨大影响，本文在分布学习基础上，提出了一种基于分布构造(Distribution construction，DC)的损失函数，以改善在同构和异构数据集场景中模型的泛化能力。在传统的同构数据集评估协议和提出的异构数据集估计协议下，对所提出的损失函数进行实验，在多个数据集下与其他方法进行了比较，结果表明了本文方法的有效性。

1 损失函数对比分析

1.1 问题导入

令$\{(\boldsymbol{x}^n,\boldsymbol{y}^n),n=1,2,\cdots,N\}$表示样本容量为$N$的训练集，其中$\boldsymbol{x}^n$和$\boldsymbol{y}^n$分别代表第$n$张输入图像及其相应的年龄标签。年龄标签$\boldsymbol{y}^n$是属于年龄标签$L=\{l_{\min},\cdots,l_{\max}\}$中的标量值，并且令$l_{\min}=1$，$l_{\max}=K$。年龄估算的目标是学习输入的面部图像$\boldsymbol{x}^n$及其对应的年龄标签$\boldsymbol{y}^n$之间的映射函数。若年龄估算模型使用one-hot编码来表示年龄标签，即标签$\boldsymbol{y}^n$由二进制向量$\boldsymbol{s}^n=[s_1^n,s_2^n,\cdots,s_K^n]^{\mathrm{T}}\in\mathbf{R}^K$来编码，若面部样本$\boldsymbol{x}^n$属于$L$中的第$k$个标签，则$s_k^n=1$，否则$s_k^n=0$。通过此种建模，年龄估算问题就转为了一般的分类问题，其中分类目标是训练CNN模型以在输入人脸图像$\boldsymbol{x}^n$及其对应的年龄标签$\boldsymbol{s}^n$之间找到映射函数$f:\boldsymbol{x}^n\rightarrow\boldsymbol{s}^n$。然而，年龄估算问题不同于一般的模式识别问题，这是由于相仿年龄的面部通常具有非常相似的图像特征。这种语义相关性会导致视觉标签的歧义性[9]。在one-hot年龄标签建模中，通常假定标签是不相关的，而忽略年龄标签的相关性会导致网络训练过程中出现不一致问题[9]。

通过将年龄估算编码为标签分布，文献[8]减轻了训练阶段的年龄标签歧义问题。类似地，本文对

于每个输入样本$\boldsymbol{x}^n$也使用标签分布$\boldsymbol{q}^n=[q_1^n,q_2^n,\cdots,q_K^n]^{\mathrm{T}}\in\mathbf{R}^K$。此时假设$\boldsymbol{q}^n$的每个元素都是在[0,1]范围内的实数,并且约束条件为$\sum_{k=1}^{K}q_k^n=1$。根据此定义,年龄标签服从相同的概率分布,此时q_k^n表示在L中第k个标签的面部样本$\boldsymbol{x}^n$的概率。通过这种类型的建模,年龄估算转化为分布学习问题。此时训练CNN的目标是通过解决式(1)这样一个最小化问题,求解输入面部图像$\boldsymbol{x}^n$与对应标签分布$\boldsymbol{q}^n$之间的映射函数$f:\boldsymbol{x}^n\rightarrow\boldsymbol{p}^n$,即

$$\min_f\sum_{n=1}^{N}L(\boldsymbol{z}^n=f(\boldsymbol{x}^n),\boldsymbol{q}^n) \tag{1}$$

式中:$L(\cdot)$为损失函数;$\boldsymbol{z}^n=f(\boldsymbol{x}^n;\Phi)\in\mathbf{R}^K$表示位于softmax层之前的CNN输出;$\Phi$表示CNN网络的参数。softmax函数将向量$\boldsymbol{z}^n$映射为概率分布$\boldsymbol{p}^n$,即范围为[0, 1]的实数向量,其总和为1。$\boldsymbol{p}^n$的每个元素(即$p_k^n$)表示样本$\boldsymbol{x}^n$属于年龄$k$的概率,即为$p_k^n=\exp(z_n^k)\Big/\sum_i\exp(z_i^n)$。

本文将标签分布定义为高斯分析[10],即对于每个年龄标签均以y^n为中心、以σ为标准差控制年龄分布值的形状(宽度)。对所有年龄段的标签分布,并恒令$\sigma=2$[9]。

此外,不同于大多数年龄估算方法采用的同构数据评价协议,本文提出一种异构数据集评估协议。异构数据集可以看作是域自适应[11]或域泛化[12]。域自适应意味着模型是使用源域中的训练数据而设计的,而工作于与源域完全不同的目标域。一般的机器学习模型便是如此,因为其通常从目标域中获取的有标记数据(即有监督的域自适应)或未标记的数据(即无监督的域自适应)的数量很少。域自适应的方法通常包括将源域中的数据重新映射到目标域,并使用此转换后的源域数据重新设计决策系统。但是在许多场景下,尤其是在年龄估算问题中,采用这种方法是不可行的。例如,假设一个基于域自适应的年龄估算模型正在估算用户上传到其中的测试图像的年龄值,此时若测试图像是从与训练所使用的相同目标域中获得的,则模型将表现良好;然而年龄估算系统要求的是在任何输入图像上都必须表现良好。基于这一重要需求,研究中需要开发一个学习框架,尽管在训练阶段对目标域没有先验知识的了解,但仍然在目标域中表现良好。

与域自适应不同,域泛化指的是预测先前从未了解过的领域中样本的标签,而无需访问目标域。此时便需要在训练过程中获取到任意数量的相关域中的样本,目的是希望所学知识能很好地融合到先前未了解的领域。本文方法基本属于域泛化,差异在于本文方法放宽了在训练阶段必须访问某些相关领域的要求,具体即为本文在一个数据集上训练模型并使用另一个不同的外部环境(例如成像条件等)与内部条件(例如种族等)的数据集进行测试。

1.2 已有损失函数的分析

本文学习算法的主要目标是最大程度地减少预测标签与真实标签分布之间的距离。现有的CE和KL(Kullback Leibler)损失函数已被广泛应用于训练使用one-hot编码和标签分布的基于CNN的年龄估算模型。本节从理论上对这些损失函数进行分析,以说明其在训练基于CNN的年龄估算模型时的局限性。下文为了简洁起见而省略了索引n。

1.2.1 CE损失函数

当年龄标签采用one-hot编码时,CE损失是训练CNN最常使用的损失函数。该损失假设这些类别是独立的,因此对于年龄估算问题而言,采用CE损失函数学习到的网络模型会忽略标签相关性。为了缓解此问题,文献[27]将CE损失函数与均值和方差项结合使用提出CE-MV损失函数,以考虑年龄标签之间的语义相关性,即

$$L_{\text{CE-MV}}(\boldsymbol{p},y)=-\log p_k+\lambda_1(\mu_p-y)^2+\lambda_2\sigma_p^2 \tag{2}$$

式中：λ_1和λ_2为正则化参数；μ_p和σ_p^2分别为估计分布$\boldsymbol{p}$的均值和方差。第1项（交叉熵项）最大化了真实类别的预测概率；第2项（均值项）惩罚了预测向量$\boldsymbol{p}$的平均值μ_p与每个输入样本x的真实年龄标签y之间的差异值；第3项（方差项）最小化预测向量$\boldsymbol{p}$的标准差σ_p。

CE损失所带来的问题是由于其正规化项而导致的参数训练不稳定，即在训练过程中出现异常值时，正则化项会网络导致较大的误差，从而导致梯度爆炸。

1.2.2 KL损失函数

当年龄标签被编码为标签分布时，KL散度是最为常用的训练CNN的损失函数[8,10]。KL损失函数定义为

$$L_{\mathrm{KL}}(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^{K}q_k\log\left(\frac{q_k}{p_k}\right) \tag{3}$$

式中：$\boldsymbol{p}$和$\boldsymbol{q}$为预测的标签分布和实际标签分布，该函数的取值范围为$[0,\infty)$（若两个分布完全匹配，则$L_{\mathrm{KL}}(\boldsymbol{p},\boldsymbol{q})=0$）且KL损失函数越低，$\boldsymbol{p}$与$\boldsymbol{q}$匹配越好。

KL损失函数的问题在于它不是对称的，即$L_{\mathrm{KL}}(\boldsymbol{p},\boldsymbol{q})\neq L_{\mathrm{KL}}(\boldsymbol{q},\boldsymbol{p})$。KL损失函数的这种不对称性会导致：若$q_k$大于$p_k$，则$r_k=q_k\log(q_k/p_k)$值为正；反之，若$q_k$小于$p_k$，则会使$r_k$为负，从而导致在整个年龄范围内匹配过程的不均匀性。

KL损失函数的另一个问题是在反向传播阶段参数的更新规则。KL损失函数相对于z_i的偏导数通过链式法则可以很轻易地得到：$\partial L_{\mathrm{KL}}/\partial z_i=p_i-q_i$。显然，在使用KL散度作为损失函数时，模型参数的更新忽略了其他年龄区间值的贡献而只考虑了涉及到的对应年龄标签之间的差值，即p_i-q_i，这样便会影响到网络收敛后参数的鲁棒性。

与CE损失函数的改进方案类似，伴有正则化项的KL-M函数为[13]

$$L_{\text{KL-M}}(\boldsymbol{q},\boldsymbol{p})=\sum_{k=1}^{K}q_k\log\left(\frac{q_k}{p_k}\right)+\lambda_1\left|\mu_p-y\right| \tag{4}$$

与原始的KL损失函数相比，KL-M损失函数使用期望回归模块作为正则化项对KL损失函数的性能进行改善。期望回归模块对估计的年龄分布平均值与真实年龄标签之间的差异进行惩罚。然而，它也具有与CE-MV损失函数相同的问题。

1.2.3 对称损失函数

本节介绍2个较常使用的对称损失函数，它们解决了之前损失函数的不对称性问题。

（1）f-损失函数

令$f:\mathbf{R}^+\to\mathbf{R}$为凸函数且$f(1)=0$，概率分布$\boldsymbol{p}$与$\boldsymbol{q}$的f-损失函数[14]定义为

$$L_f(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^{K}q_k f\left(\frac{p_k}{q_k}\right) \tag{5}$$

当$f(x)=-\log(x)$时，f-损失函数便成为了KL散度。

当$f(x)=\dfrac{(x-1)^2}{x+1}$时，$\chi^2$统计定义为

$$L_{\chi^2}(\boldsymbol{p},\boldsymbol{q})=\frac{1}{2}\sum_{k=1}^{K}\frac{(p_k-q_k)^2}{q_k} \tag{6}$$

由于此损失函数是不对称的，因此在附加对称性后可得到[14]

$$L_{\chi^2}(\boldsymbol{p},\boldsymbol{q})=\frac{1}{2}\sum_{k=1}^{K}\frac{(p_k-q_k)^2}{p_k+q_k} \tag{7}$$

(2) JS损失函数

当$f(x)=\frac{1}{2}x\log(x)-\frac{1}{2}(x+1)\log\left(\frac{1}{2}(x+1)\right)$,JS损失函数[15]为

$$L_{\mathrm{JS}}(\boldsymbol{p},\boldsymbol{q})=\frac{1}{2}(L_{\mathrm{KL}}(\boldsymbol{p},m)+L_{\mathrm{KL}}(\boldsymbol{q},m)) \tag{8}$$

式中$m=\frac{1}{2}(\boldsymbol{p}+\boldsymbol{q})$。JS损失函数便是KL损失函数的对称版本。

1.3 DC损失函数

由于使用不同的损失函数会影响参数模型的性能优劣,因此设计一个好的损失函数对基于CNN的年龄估算模型来说相当重要。本文基于分布学习提出DC损失函数,定义为

$$L_{\mathrm{DC}}(\boldsymbol{p},\boldsymbol{q})=\frac{1}{2}\sum_{k=1}^{K}\left|q_k^{\alpha}-p_k^{\alpha}\right|^{\frac{1}{\alpha}}=q^k\left|1-\left(\frac{p_k}{q_k}\right)^{\alpha}\right|^{\frac{1}{\alpha}} \tag{9}$$

式中α为介于0到1之间的超参数。当$f(x)=\left|1-x^{\alpha}\right|^{(1/\alpha)}$时,DC损失函数就变为了概率分布$\boldsymbol{p}$与$\boldsymbol{q}$之间的f-损失函数。当$\alpha\to 0.5$时,提出的DC损失函数接近文献[16]中的损失函数。

1.3.1 DC损失函数与用于年龄估算的损失函数的性能比较

对于年龄估算问题,本文提出的DC损失函数具有以下良好特性:

(1)利用链式法则,提出的损失函数L_{DC}对向量$\boldsymbol{z}$的每个元素的导数(具体推导过程略)为

$$\frac{\partial L_{\mathrm{DC}}}{\partial z_i}=p_i\left[\left(1-\left(\frac{q_i}{p_i}\right)^{\alpha}\right)\left|1-\left(\frac{q_i}{p_i}\right)^{\alpha}\right|^{\frac{1-2\alpha}{\alpha}}+\sum_{k=1}^{K}p_k\left(1-\left(\frac{q_k}{p_k}\right)^{\alpha}\right)\left|1-\left(\frac{q_k}{p_k}\right)^{\alpha}\right|^{\frac{1-2\alpha}{\alpha}}\right] \tag{10}$$

从式(10)可以看出,本文提出的损失函数使得模型参数的更新规则取决于$\boldsymbol{p}$和$\boldsymbol{q}$的所有项。与KL损失函数更新规则仅取决于目标项p_i-q_i相比,它更加鲁棒。

KL与DC损失函数在不同高斯分布下的性能对比如图1所示,其中图1(a)为2个原始的高斯分布,图1(b)为KL损失函数与图1(a)的比率,图1(c)为DC损失函数与图1(a)的比率。在图1(b)中,两个分布之间的距离在点A和点B处相等,但是在这些点处的r值却不同。因此,当q_k大于p_k时对总误差的贡献要比q_k较小时对总误差的贡献更大。这也意味点A对总误差的贡献要大于点B。因此可以得出结论:在使用KL损失函数对距离进行最小化后,当q_k大于p_k时,p_k对q_k有更好的拟合度;而当q_k较小时,则反之。

(2)与现有的损失函数相反,对于[0,1]区间中的任何α值,DC损失函数都是对称的(图1(c))。由于这种特性,参数的迭代过程将在整个年龄区间范围内均匀执行。

(3)DC损失函数解决了稳定性问题,即与CE-MV和KL-M损失函数相比,没有任何正则化项,从而避免了参数波动并有助于网络收敛。

(4)与CE-MV和KL-M的关系如下:

令α=0.5,此时本文损失函数可重写为

$$L_{\mathrm{DC}}(\boldsymbol{q},\boldsymbol{p})=\sum_{k=1}^{K}\sqrt{p_k}\left(1-\frac{\sqrt{q_k}}{\sqrt{p_k}}\right)^2=2\left(1-\sum_{k=1}^{K}\sqrt{p_kq_k}\right)=2(1-C) \tag{11}$$

式中:$C=\sum_{k=1}^{K}\sqrt{p_k}\,q_k$,显然$\sum_{k=1}^{K}p_k=1$和$\sum_{k=1}^{K}q_k=1$。使用本文提出的DC损失函数后,网络的softmax层

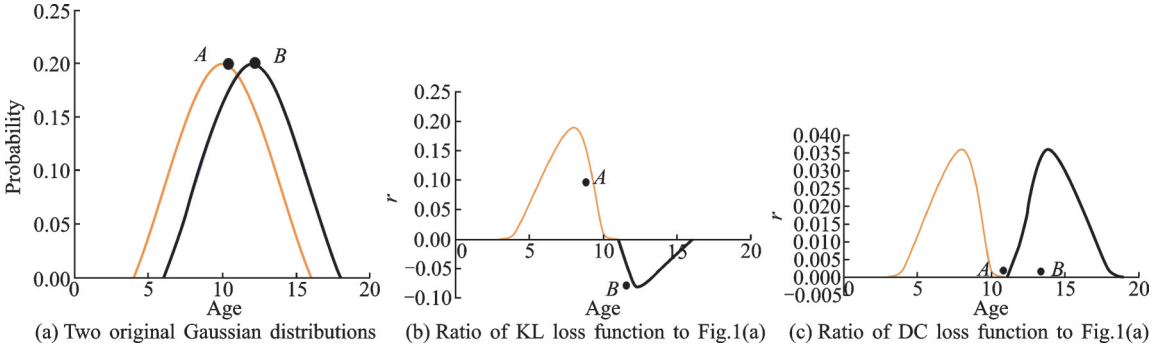

图1 KL与DC损失函数在不同高斯分布下的性能对比

Fig.1 Performance comparison with KL and DC loss functions under two normal distributions

输出为平均值为μ_p和标准偏差为σ_p的高斯分布$\boldsymbol{p}$。此时的C便具有如下闭式表达[17]

$$C=\frac{2\sigma_p\sigma_q}{\sigma_p^2+\sigma_q^2}\exp\left(-\frac{1}{4}\frac{(\mu_p-\mu_q)^2}{(\sigma_p^2+\sigma_q^2)}\right) \tag{12}$$

对式(12)取自然对数求相反值后,DC损失函数的最小值等价于式(13)的最小值。

$$\overline{C}=\frac{1}{4}\frac{(\mu_P-\mu_q)^2}{\sigma_p^2+\sigma_q^2}+\frac{1}{2}\ln\left(\frac{\sigma_p^2+\sigma_q^2}{2\sigma_p\sigma_q}\right) \tag{13}$$

由式(13)可以看出,与CE-MV和KL-M损失函数类似,DC损失函数隐式地惩罚了年龄估算分布与真实年龄分布之间的差异,并且其年龄估算分布与均值周围。此外,式(13)中的$(\mu_p-\mu_q)^2$项使用了方差值进行归一化,也减轻了由异常值引起的同样出现于CE-MV和KL-M损失中的不稳定性问题。

1.3.2 DC损失函数与其他损失函数的性能比较

本节将DC损失函数与JS损失函数和χ^2-统计量进行比较。首先重新构造函数,简单起见下文令α为0.5。通过因式分解和级数展开[18]可以将L_{χ^2},L_{JS}和L_{DC}与$L_f(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^{K}\frac{s_k}{2}G_f\left(\frac{d_k}{s_k}\right)$联系起来,其中$s_k=p_k+q_k$,$d_k=|p_k-q_k|$。此时每个函数$G_f(\cdot)$中的$L_{\chi^2}$,$L_{\mathrm{JS}}$和$L_{\mathrm{DC}}$之间的关系可以推导为

$$\begin{cases}G_{\chi^2}(x)=x^2\\ G_{\mathrm{JS}}(x)=\dfrac{1}{2}((1+x)\log(1+x)+(1-x)\log(1-x))\\ G_{\mathrm{DC}}(x)=1-\sqrt{1-x^2}\end{cases} \tag{14}$$

式中:$s_k\in[0,2]$,$d_k\in[0,1]$。

然后分析不同损失函数之间的比率。对于常数s_k,比率定义为$\frac{G_{\mathrm{DC}}}{G_{\chi^2}}$和$\frac{G_{\mathrm{DC}}}{G_{\mathrm{JS}}}$,对比曲线如图2所示。由图2可见,当误差$d_k$适中时,比率较为平坦;当$d_k\to 1$时恰好达到最大比率;而当$p_k\to q_k$时则达到最小比率。图2结果为DC损失函数的相对性质提供了最直观的解释。当误差较大时,DC损失函数与G_{χ^2}有相似的纠错能力[14]。因此在早期训练阶段,当误差很大时两个函数以相同的方式对误差进行修正。但是,随着训练过程的继续进行和误差越来越小,DC损失函数将减小误差的影响。当误差接近零时,将不再关注估算分布与实际分布中的对应点,这种属性会减少训练过程对不属于分布点的关注,从而可以稳定训练过程。与G_{JS}的关系同样类似,DC损失函数总是提供比G_{JS}更强的响应。从图2可以看

出,DC损失函数在处理不同误差时更具“动态性”[15]。

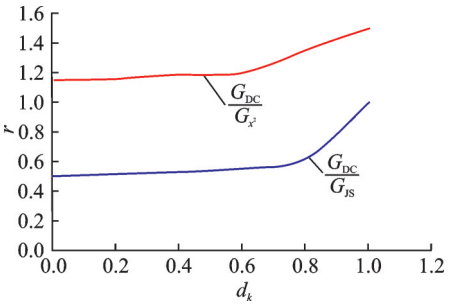

图2 不同损失函数的比率

Fig.2 Ratio of different loss functions

2 实验验证及结果分析

2.1 评价指标

年龄估算的性能主要以2种度量指标进行评价:平均绝对误差(Mean absolute error,MAE)与累计分数(Cumulative score,CS)。MAE定义为$\mathrm{MAE}=\sum_{k=1}^{N}|\hat{l}_k-l_k|/N$,其中$l_k$为测试样本$k$的实际年龄值,$\hat{l}_k$为估计到的年龄值,$N$为测试集的样本容量。CS定义为$\mathrm{CS}(j)=N_{e\leqslant j}/N\times 100\%$,其中$N_{e\leqslant j}$为测试集中的绝对值误差不低于$j$的图像总数,本文中的$j$设置为5。

2.2 数据集

(1) 训练集

本文实验选用IMDB-WIKI[19]数据集作为训练集。IMDB-WIKI是目前最大的可用于年龄估算的开源数据集,它包含523 051张图像,且这些图像的年龄标签介于0~100岁之间。该数据库的图像是直接从网络抓取得到,因此没有经过仔细筛选,所以包含许多不适合年龄估算的图像。虽然文献[5]手动清除了IMDB-WIKI数据库中所有低质量的图像,但是仍然有许多带有错误标签的图像,一定程度上影响了年龄估算的性能。基于文献[5]的结果,本文以一种半监督的方式更加仔细地清理了IMDB-WIKI数据库,并删除了所有不合适的图像。首先,本文使用MTCNN面部检测器[20]检测每个图像中的人脸;然后从数据库中删除了多人图像(因为每个图像只有一个年龄标签)、面部检测器的置信度分数低于检测到的面部阈值(设置为0.9)的图像以及检测到的面部边框尺寸小于一定值(设置为20像素×20像素)的图像。 最终,通过人工逐张地检查剩余的图像,并删除所有低质量的图像以及带有错误标签的图像。此外,还从中筛选出了0~100岁的40 000张图像,构成IMDB-WIKI-40k数据集。相比IMDB-WIKI,IMDB-WIKI-40k的类不平衡性很好地得到了解决。

(2) 测试集

本文实验选用MORPH[21]和FG-NET[22]数据集作为测试集。MORPH数据集包含14~77岁年龄段的13 647名不同种族的55 174张图像,并且其中超过90%的图像是非洲或欧洲人。FG-NET数据集包含1 002张来自82个人的图像,年龄范围是0~70岁,其中的面部图像在姿势、表情和光照条件等方面皆存在巨大差异,因此FG-NET更为符合实际场景当中的面部图像。

2.3 实验设置及数据集预处理

本文采用预训练后的VGG模型[23]作为主干网络进行年龄估算。最后的全连接层2 048被替换为K,其中K为年龄级数且在本文中设置为101;CNN网络的输入是224像素×224像素大小的面部图像;使用mini-batch为80的随机梯度下降算法对整体参数进行优化;动量和权重衰减系数分别设置为0.9和0.000 5;卷积层、前2个全连接层和最后1个全连接层的学习率分别初始化为0.001、0.001和0.01;在每轮中学习率都采用指数下降的模式;采用随机翻转、裁剪和颜色抖动进行数据增强。

训练集和测试集的每张图像都通过以下两个步骤进行预处理:(1)采用MTCNN面部检测器[20]用于检测每个图像中的5个面部标志(眼睛的左右中心、鼻尖、左右角);(2)通过文献[24]中提出的方法,使用面部关键点来使每个面部居中并对齐;(3)将标准化后的脸部调整为256像素×256像素。在测试时,对估计到的年龄分布$\boldsymbol{p}$中通过$\hat{y}=\mathrm{argmax}_k\, p_k$计算来获得预测年龄。

2.4 评估协议

2.4.1 同构数据集评估协议

本文采用了随机分割协议和S_1-S_2-S_3交叉验证协议来评估MORPH数据集上的年龄估算性能。MORPH数据集中的图像在性别与种族中的分布非常不平衡,其中男女比例约为5.5,白人与黑人的比例约为4。借鉴文献[25-29],将MORPH数据集分为3个不重叠的子集S_1、S_2和S_3,以缓解这种不平衡分布。这种划分方式使男女比例约为3,而白人与黑人比例约为1。由此进行了两个实验:(1)使用S_1进行训练,使用S_2+S_3进行测试;(2)使用S_2进行训练,使用S_1+S_3进行测试。本文实验也使用上述2个协议及其平均值评价性能,将这2个协议分别定义为S_1/S_2+S_3协议和S_2/S_1+S_3协议。

2.4.2 异构数据集评估协议

现有的年龄估算方法主要遵循同构数据集评估协议,即训练和测试集来自同一数据集。例如在随机分割协议中,随机选择数据集的80%图像进行训练,其余的用于测试。使用同构数据集评估协议而训练的模型可能会对训练集具有有偏性,并在面对未知信息的面部图像时会提供不可靠的年龄结果。在许多实际场景中,测试和训练集中的分布和特征完全不同。因此,同构数据集评估协议在评估年龄估算方法的泛化性能方面始终具有局限性。

本文提出了一种新颖的年龄估算评估协议,称为异构数据集评估协议(图3),它使得年龄估算方法的性能评价更有意义。该协议主要考虑的是训练后的模型应完全不了解测试数据集的分布及特征,这意味着不应使用测试数据集中的任何图像来训练网络。此外,训练集和测试集也不应该使用同一个人的不同条件下的面部图像。这便是与同构数据集协议的不同之处,在同构数据集协议中,测试图像与训练图像来源于同一数据集。在本文的评估协议下,可以更可靠地评估训练模型的泛化能力。

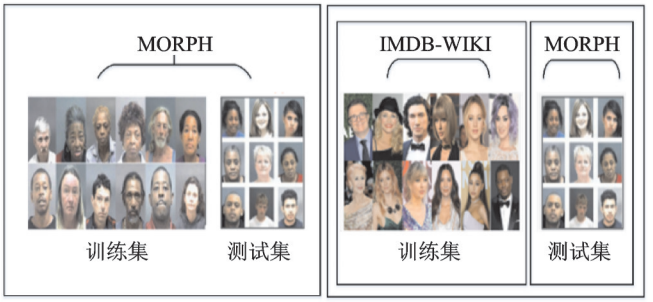

图3 两种数据集层面的评估协议对比

Fig.3 Comparison of different assessment protocols

2.5 结果分析

2.5.1 超参数对性能的影响

本文提出的损失函数中唯一的超参数是α,因此本节评估其对年龄估算性能的影响。将α值从0.1更改为1,并使用VGG模型在IMDB-WIKI数据集中训练,然后在FG-NET数据集中进行验证,得到的不同α值情况下的MAE,如图4所示。图4结果表明,当$0.4\leqslant\alpha\leqslant0.55$时,$\alpha$值对年龄估算的结果影响较小;当使用较大或较小的$\alpha$值时,年龄估算的性能会降低。基于此分析,本文将$\alpha$的值固定设置为0.5。

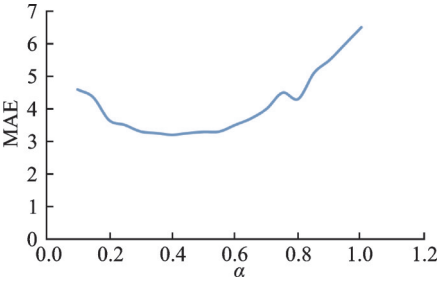

图4 不同超参数对MAE的影响

Fig.4 Performance difference among different hyperparameters

2.5.2 采用同构数据集评估协议的实验对比

本节在同构数据集评估协议下对本文年龄估算方法的性能与其他方法进行比较。随机分割、S_1/S_2+S_3和S_2/S_1+S_3协议下的MAE值和CS分值如表1、2所示。

表1 MORPH中采用随机分割协议的性能对比

Table 1 Performance comparison under random splitting protocol in MPORH

方法	MAE/年	CS/%
文献[2]*	2.96	50.7
文献[1]*	2.75	—
文献[30]	2.42	—
文献[7]	2.41	90.1
文献[7]*	2.17	—
文献[31]	2.14	91.3
文献[13]	1.97	—
本文方法	**1.87**	**94.27**
本文方法*	**1.84**	**94.74**

注:*为经过预训练的方法。

表1中文献[2]将年龄估算分类问题转换为排序问题进行解决;文献[1]提出一种紧凑型级联的基于上下文的年龄估算模型;文献[30]使用带有标签分布编码的KL和KL-M损失函数;文献[7]采用CE-MV作为损失函数,并采用one-hot年龄标签编码;文献[31]提出可微的深度随机森林;文献[13]采用基于标签分布的年龄估算方法。从表1中可以看出,本文方法在MORPH数据集上的随机分割协议下达到了最优的性能。当直接在MORPH数据集上对模型进行微调时(此时没有在IMDB-WIKI数据集上对网络进行预训练),MAE值可以达到1.87。为了进一步提高性能,在IMDB-WIKI数据集上对网络进行了预训练,可以看出,本文方法得到的MAE为1.84,这相比于其他方法中的最优方法提高了0.13年。

表2 MORPH中采用S_1/S_2+S_3和S_2/S_1+S_3协议的性能对比

Table 2 Performance comparison under S_1/S_2+S_3 and S_2/S_1+S_3 protocols in MPORH

方法	S_1/S_2+S_3		S_2/S_1+S_3		平均性能	
	MAE/年	CS/%	MAE/年	CS/%	MAE/年	CS/%
文献[27]	3.84	—	3.87	—	3.85	—
文献[32]	—	—	—	—	2.90	82.60
文献[33]	3.07	—	2.77	—	2.95	—
文献[34]	2.80	—	2.81	—	2.81	—
文献[28]*	2.82	—	2.58	—	2.70	—
文献[35]*	2.74	—	**2.51**	—	**2.63**	86.00
本文方法	2.80	87.21	2.81	85.15	2.80	86.18
本文方法*	**2.73**	**87.44**	2.74	**85.47**	2.73	**86.42**

注:*为经过预训练的方法。

表2中文献[27]采用基于“结构”的年龄估算方法;文献[32]进行年龄差异的估算;文献[33]引入辅助人口统计信息进行年龄估算;文献[34]先采用相关的属性对年龄进行分类,继而采用排序方法进行年龄估算;文献[28]采用组编码与解码进行年龄估算;文献[35]采用软排序对标签进行建模。根据表2结果可以看出,本文方法无需在任何其他年龄相关的面部数据集上进行预训练即可实现2.8的平均

MAE。与其他方法中最优的结果相比，CS分值提高了0.18%。同样地，在采用IMDB-WIKI数据集进行预训练后进一步提高了性能：MAE为2.73，CS为86.42%。从表2中还可以看出，其他方法在S_1/S_2+S_3和S_2/S_1+S_3协议下的结果之间存在明显差距，证实了现有方法对训练数据集的特征与分布变化（如性别，肤色等）较为敏感，而本文方法在两种协议上的性能相差无几。值得一提的是，文献[34]通过采用多任务学习策略来提高训练模型对性别、种族等不同面部属性特征的鲁棒性，从而达到与本文方法相似的性能。然而，本文方法是在没有利用其他属性（例如性别和肤色等）的基础上提高了训练模型的鲁棒性。

在同构数据集评估协议下仿真的常见问题是测试到的结果带有有偏性，而这种有偏性在S_1/S_2+S_3和S_2/S_1+S_3评估协议下体现得更加淋漓尽致。因此在该协议下，无法衡量年龄估算方法在对未知面部图像的泛化能力。

2.5.3 采用异构数据集评估协议的实验对比

考虑到在同构数据集评估协议下可能无法准确度量年龄估算方法的性能，本节在本文提出的异构数据集评估协议中进行了更为有效的实验。为了公平地进行比较，本文在IMDB-WIKI数据集训练了这些模型，此外还用JS对称损失函数和χ^2-统计量训练VGG模型以表明所提出的损失函数的有效性。采用异构模型在不同目标域的性能对比如表3、4所示。

表3 采用异构数据集评估协议在不同目标域中的性能对比（训练集：IMDB-WIKI）

Table 3 Performance comparison under proposed protocol and different testing sets (training set: IMDB-WIKI)

方法	MORPH		FG-NET		平均性能	
	MAE/年	CS/%	MAE/年	CS/%	MAE/年	CS/%
文献[36]	5.48	60.25	3.74	78.50	4.61	69.38
文献[2]	5.50	60.34	3.20	82.14	4.35	71.24
文献[28]	5.40	60.95	3.15	82.98	4.28	71.97
文献[35]	5.28	62.55	3.12	83.80	4.20	73.18
文献[10]	5.27	62.34	3.08	83.83	4.18	73.09
文献[7]	5.22	61.31	3.07	83.23	4.15	72.27
文献[13]	4.95	64.95	3.06	82.83	4.01	73.89
χ^2统计量	4.76	66.49	5.35	59.28	5.06	62.89
JS散度	4.81	65.83	2.99	83.53	3.90	74.68
本文方法	**4.63**	**66.03**	**2.93**	**84.43**	**3.78**	**75.23**

表3中文献[36]使用排序CNN进行年龄估算；文献[10]提出带有标签歧义性的深度标签分布方法；文献[28]进行人口统计方面的统计。从表3、4结果可以看出：

（1）文献[10]和文献[13]方法的年龄估计准确性高于文献[2]和文献[7]（CE-MV）方法，这表明标签分布有助于改善年龄估算的性能。这是因为在训练过程中，基于one-hot编码方式的损失函数并未考虑标签模糊性（ambiguity）的影响[10]。

（2）文献[35]的MAE和CS与文献[18]方法比较接近，这是因为文献[35]和文献[10]的算法具有线性关系[13]。应当强调的是，诸如CE-MV和文献[13]之类的方法分别在CE和KL损失函数中加入了正则项，正是由于这些正则化参数的大小不同，导致训练网络的最终对正则化参数的选择相当敏感。

表4 采用异构数据集评估协议在不同目标域中的性能对比(训练集:IMDB-WIKI-40k)

Table 4 Performance comparison under proposed protocol and different testing sets (training set: IMDB-WIKI-40k)

方法	MORPH		FG-NET		平均性能	
	MAE/年	CS/%	MAE/年	CS/%	MAE/年	CS/%
文献[2]	6.54	53.38	3.57	78.94	5.06	66.16
文献[28]	6.40	53.97	3.53	79.78	4.96	66.87
文献[10]	6.01	57.36	3.24	81.54	9.25	69.54
文献[7]	6.22	55.60	3.34	80.44	4.78	68.02
文献[13]	5.80	57.30	3.35	81.44	4.58	69.37
JS散度	5.64	58.17	3.33	79.74	4.49	68.96
本文方法	**5.63**	**59.13**	**3.21**	**81.59**	**4.42**	**70.36**

与之不同,本文提出的损失函数没有任何正则化超参数,因此可以有效地缓解此问题,从而在年龄预测时保持良好的性能。

(3)从表3的下半部分可以推断出,采用本文DC损失函数进行年龄估算时的准确性显著高于其他对称损失函数(例如χ^2统计量和JS散度)所获得的预测准确性,这些结果也为第1节中的理论分析增加了实验支撑。因此可以得出,本文的方法在异构数据集测试中具有良好的泛化能力,因此可以处理未知场景。

本文还进一步使用MORPH数据集对模型进行训练以对异构数据集评估协议进行进一步效果分析,结果如表5所示。可以看出,采用MORPH作为训练数据集时,所有方法的性能都会下降,但本文方法仍然是性能最好的方法。

表5 采用异构数据集评估协议在FG-NET中的性能对比(训练集:MORPH)

Table 5 Performance comparison under proposed protocol and FG-NET (training set: MORPH)

方法	MAE/年	CS/%
文献[2]	5.74	58.31
文献[18]	5.45	62.76
JS散度	5.34	63.47
本文方法	5.29	63.7

当比较同构数据集评估协议(表1、2)和异构数据集评估协议(表3、4)在MORPH上的实验结果时,发现在同构数据集评估协议下达到较高的年龄估算准确性并不能保证在面对未知目标域时仍具有良好的性能。很显然,未知情况会导致所有年龄估算方法的性能下降,这可能因为它们并没有学习到在不同性别、不同光线变化、不同姿势和不同面部表情变化的不同特征与分布的图像特征。但是,无论在哪种评估协议中,本文方法仍然得到最优的MAE和最高的CS,这也进一步说明了本文方法的有效性和鲁棒性。

3 结束语

本文考虑年龄的标签值通常呈现局部相关性,将年龄估算问题建模为一个以真实年龄值为中心的高斯分布学习问题。为了使得模型参数学习到这种分布特征,提出了一个损失函数用于提供估计分布与真实分布之间的迭代逼近。然后,为了更好地评估年龄估算方法的泛化性能,与此前通常假设训练集与测试集中数据同特征同分布的同构数据集评估协议不同,提出了更符合年龄估算方法的实际用例场景的异构数据集评估协议。理论分析和实验结果皆证明了本文方法的有效性。

参考文献：

[1] ZHANG C, LIU S, XU X, et al. C3AE: Exploring the limits of compact model for age estimation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019.

[2] CHEN S, ZHANG C, DONG M. Deep age estimation: From classification to ranking[J]. IEEE Transactions on Multimedia, 2018, 20(8): 2209-2222.

[3] EIDINGER E, ENBAR R, HASSNER T. Age and gender estimation of unfifiltered faces[J]. IEEE Transactions on Information Forensics and Security, 2014,9(12): 2170-2179.

[4] LIU H, LU J, FENG J, et al. Label-sensitive deep metric learning for facial age estimation[J]. IEEE Transactions on Information Forensics and Security, 2017 (2): 1.

[5] ZHANG K, GAO C, GUO L, et al. Age group and gender estimation in the wild with deep RoR architecture[J]. IEEE Access, 2017, 5: 22492-22503.

[6] HU Z, WEN Y, WANG J, et al. Facial age estimation with age difference[J]. IEEE Trans Image Process, 2017, 26(7): 3087-3097.

[7] PAN H, HAN H, SHAN S, et al. Mean-variance loss for deep age estimation from a face[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2018.

[8] GENG X, YIN C, ZHOU Z H. Facial age estimation by learning from label distributions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35: 2401-2412.

[9] GENG X, JI R. Label distribution learning[C]//Proceedings of IEEE International Conference on Data Mining Workshops. [S.l.]: IEEE, 2013.

[10] GAO B B, XING C, XIE C W, et al. Deep label distribution learning with label ambiguity[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2825-2838.

[11] 黄兵, 郭继昌. 基于Gabor小波与LBP直方图序列的人脸年龄估计[J]. 数据采集与处理, 2012, 27(3): 340-345.
HUANG Bing, GUO Jichang. Face age estimation based on Gabor wavelet and LBP histogram sequence[J]. Journal of Data Acquisition and Processing, 2012, 27(3): 340-345.

[12] 章超. 基于深度学习的图像有序性估计研究[D]. 成都:电子科技大学, 2019.
ZHANG Chao. Research on image ordnance estimation based on deep learning [D]. Chengdu: University of Electronic Science and Technology of China, 2019.

[13] GAO B B, ZHOU H Y, WU J, et al. Age estimation using expectation of label distribution learning[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018). Sweden: [s.n.], 2018.

[14] ÖSTERREICHER F. Csiszár's f-divergences-Basic properties[J]. Rgmia Res Rep: Coll, 2002: 1-13.

[15] IGAL S. On f-Divergences: Integral representations, local behavior, and inequalities[J]. Entropy, 2018, 20(5): 383.

[16] CHA S H. Comprehensive survey on distance/similarity measures between probability density functions[J]. International Journal of Mathematical Models & Methods in Applied ENCES, 2007, 1(4): 300-307.

[17] AHERNE F J, THACKER N A, ROCKETT P I. The Bhattacharyya metric as an absolute similarity measure for frequency coded data[J]. Kybernetika, 1998, 34: 363-368.

[18] SCOTT A. Geometric comparison of popular mixture-model distances[J]. Journal of Modern Mathematics and Frontier, 2010, 4: 1-12.

[19] ROTHE R, TIMOFTE R, GOOL L V. DEX: Deep expectation of apparent age from a single image[C]//Proceedings of IEEE International Conference on Computer Vision Workshop. [S.l.]: IEEE, 2016.

[20] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

[21] RICANEK K , TESAFAYE T . MORPH: A longitudinal image database of normal adult age-progression[C]//Proceedings of International Conference on Automatic Face & Gesture Recognition. [S.l.]: IEEE, 2006.

[22] PANIS G, LANITIS A, TSAPATSOULIS N, et al. Overview of research on facial ageing using the FG-NET ageing database[J]. IET Biometrics, 2016, 5(2): 37-46.

[23] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep face recognition[C]//Proceedings of British Machine Vision Conference. Swansea, UK: BMVA Press, 2015.

[24] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[M]. [S.l.]: Springer International Publishing, 2016.

[25] LI K, XING J, HU W, et al. D2C: Deep cumulatively and comparatively learning for human age estimation[J]. Pattern Recognition, 2017, 66: 95-105.

[26] LI W, LU J, FENG J, et al. Bridgenet: A continuity-aware probabilistic network for age estimation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019.

[27] LIU K, LIU T. A structure-based human facial age estimation framework under a constrained condition[J]. IEEE Transactions on Image Processing, 2019, 28(10): 5187-5200.

[28] TAN Z, WAN J, LEI Z, et al. Efficient group-*n* encoding and decoding for facial age estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2610-2623.

[29] YI D, LEI Z, LI S Z. Age estimation by multi-scale convolutional network[C]//Proceedings of Asian Conference on Computer Vision. [S.l.]: IEEE, 2015.

[30] GAO B B, XING C, XIE C W, et al. Deep label distribution learning with label ambiguity[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2825-2838.

[31] SHEN W, GUO Y, WANG Y, et al. Deep differentiable random forests for age estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 404-419.

[32] SHEN W, GUO Y, WANG Y, et al. Deep regression forests for age estimation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018.

[33] WAN Jun, TAN Zichang, LEI Zhen, et al. Auxiliary demographic information assisted age estimation with cascaded structure [J]. IEEE Transactions on Cybernetics, 2018, 48(9): 2531-2541.

[34] XIE J C, PUN C M. Chronological age estimation under the guidance of age-related facial attributes[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2500-2511.

[35] ZENG X, HUANG J, DING C. Soft-ranking label encoding for robust facial age estimation[J]. IEEE Access, 2020 (99): 1.

[36] CHEN S, ZHANG C, DONG M, et al. Using ranking-CNN for age estimation[C]//Proceedings of IEEE Conference on Computer Vision & Pattern Recognition. [S.l.]: IEEE, 2017.

作者简介:

王军祥(1975-),**通信作者**,男,副教授,研究方向:人工智能技术应用、网络与信息安全、软件设计与开发,E-mail: jx_wang1975@163.com。

吴伶(1985-),女,博士,讲师,硕士生导师,研究方向:深度学习、数据挖掘与智能信息处理。

(编辑:张黄群)