一种基于前向无监督卷积神经网络的人脸表示学习方法

朱 陶[1] 任海军[2] 洪卫军[1]
(中国人民公安大学警务信息工程学院 北京 100038)[1] (重庆大学软件学院 重庆 400044)[2]

摘 要 当前基于深度卷积神经网络的人脸表示学习方法需要利用海量的有标注的人脸数据。在实际应用中,精确标注人脸的身份非常困难。因此,提出了一种基于前向无监督卷积神经网络的人脸表示学习方法。其中,基于K-means聚类获取训练样本虚拟标签,再利用线性判别分析进行卷积核学习。提出的网络结构简单有效,训练阶段不需要反向传递,训练速度显著优于有监督的深度卷积神经网络。实验结果表明,提出的方法在真实条件下的人脸数据集LFW和经典的Feret数据集上取得了优于当前主流的无监督特征学习方法和局部特征描述子的性能。

关键词 无监督学习,卷积神经网络,人脸识别,表示学习
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.060

Forward and Unsupervised Convolutional Neural Network Based Face Representation Learning Method

ZHU Tao[1] REN Hai-jun[2] HONG Wei-jun[1]
(College of Police Information Engineering, People's Public Security University of China, Beijing 100038, China)[1]
(College of Software Engineering, Chongqing University, Chongqing 400044, China)[2]

Abstract The existing face representation learning methods based on deep convolutional neural network demand massive labeled face dataset. In real-world application, it is difficult to precistly annotate the labels of face dataset. In this paper, an unsupervised forward convolutional neural network based face representation learning algorithm was proposed. By design, virtual labels of training samples were abtained based on K-means clustering and then convolution kernels were learnt by linear discriminant analysis. The network architecture in this paper is simple and effecitve and does not need back propagation during training, so its training speed is much quicker than supervised deep convolution neural network. The experimental results demonstrate that the proposed method in this paper out performs the state-of-the-art unsupervised feature learning algorithm and local feature descriptors in both real-world Labeled Face in the Wild (LFW) dataset and the classical controlled Feret dataset.

Keywords Unsupervised learning, Convolutional neural network, Face recognition, Representation learning

1 引言

自动人脸识别是一种非接触式的生物特征识别技术,涵盖了计算机视觉、机器学习、数字图像处理和神经感知理论等多个技术领域[1,20]。与指纹识别、虹膜识别以及掌纹识别技术不同,人脸识别技术具有非接触和不易篡改的特点,因此广泛应用于访问控制、人脸考勤、安防监控、智能相册和各类互动娱乐应用。2007年以来,随着真实场景下的人脸识别评测数据库LFW(Labeled Face in the Wild)的提出,真实场景下的鲁棒人脸识别问题成为当前的研究热点[12]。

一般而言,人脸识别技术分为人脸表示和建立在人脸表示基础上的测度学习两个步骤。其中,鲁棒的人脸表示是实现高精度人脸识别的基础。主流的人脸表示学习方法可以分为两大类:基于局部特征的人脸表示方法[4-7]和基于学习的人脸表示方法[8,9,15,16]。

一直以来,局部特征表示是人脸表示领域的经典方法,其基本思想是通过人工设置的规则在图像中每个像素的局部区域抽取特征表示。其中较有代表性的方法包括:Gabor Face[4],即基于Gabor小波的人脸表示方法;LBP Face[5],即基于局部二值特征的人脸表示方法;HOG Face[6],即基于梯度方向直方图的人脸表示方法;SIFT Face[7],即基于旋转不变特征的人脸表示方法等。局部特征表示相比像素特征具有更好的鲁棒性,但特征的设计也高度依赖于特征设计者的经验。

基于学习的人脸表示方法可进一步分为有监督的人脸表示学习方法和无监督的人脸表示学习方法。在有监督的人脸表示学习方法中,最具代表性的方法是基于深度卷积神经网络(DCNN)[13]的人脸表现学习方法,例如DeepFace[15]和DeepID2[16],在真实场景下的人脸识别评测数据集LFW上取得了当前最好的性能。在无监督的人脸表示学习方法中,代

到稿日期:2015-05-27 返修日期:2015-07-04 本文受国家高技术研究发展计划(863计划)(2013AA014604 2014)资助。
朱 陶(1984-),女,博士生,主要研究方向为电子物证、网络安全、计算机视觉,E-mail:zhutaotao721@163.com;任海军(1978-),男,博士,副教授,主要研究方向为电力系统自动化、计算机视觉;洪卫军(1958-),男,博士生导师,主要研究方向为安全防范系统、模式识别。

表性的方法包括 K-means Net[8] 和 PCANet[9]。其中 K-means Net 利用 K-means 聚类[10]进行卷积核学习，PCANet 利用主成分分析方法(PCA)[2]进行卷积核学习。目前，基于学习的人脸表示方法是人脸表示学习领域中最热门的研究方向。

基于深度卷积神经网络的人脸表示学习方法需要大量的有标注的人脸数据来进行训练。对于很多人脸识别场景而言，采集海量的人脸样本相对容易，但是精确地标记训练集合中人脸的身份标签则非常困难。为了利用海量的无标注的人脸数据，围绕无监督的人脸表示学习方法进行研究。文中提出了一种基于前向无监督卷积神经网络的人脸表示学习方法。其基本思想为，利用人脸图像的局部性特点对人脸进行分块，在每个人脸块上进行无监督的卷积核学习。代表性的无监督人脸表示学习方法 K-Means Net 和 PCANet 通过无监督学习算法进行卷积核学习，无法保证所提取的人脸特征的判别性。为了提高卷积特征的判别性，本方法首先通过 K-Means 聚类对人脸块上提取的人脸小块进行聚类，赋予其虚拟标签，通过虚拟标签进一步利用线性判别分析学习判别投影作为卷积核，从而能够在无标注的人脸数据上实现卷积核的判别学习。通过卷积得到响应图后，再利用经典的卷积神经网络中的 Sigmoid 操作进行归一化，进一步利用 Pooling 操作获得不变性，在向量化后采用白化主成分分析(WPCA)进行降维。

2 基于前向无监督卷积网络的人脸表示学习方法

本节首先介绍背景技术 K-means 聚类和线性分析，然后介绍提出的基于前向无监督卷积网络的人脸表示学习方法。

2.1 相关技术

本节介绍人脸表示学习的背景技术：K-means 聚类[9]和线性判别分析[3]。

(1)K-means 聚类[9]

K-means 聚类是一种无监督的机器学习方法，其基本思想在于通过迭代将输入样本集合分为若干个子类(cluster)，从而揭示数据在特征空间中的结构。假设样本集合 $\{x_1, x_2, \cdots, x_m\} \in R^{m*n}$，K-means 算法的基本步骤如下：随机选择 K 个聚类中心(cluster centroids)为：$\mu_1, \mu_2, \cdots, \mu_K \in R^n$，重复以下步骤直到收敛。

对于每一个样本 $x^{(i)}$，计算其从属于哪一个类别：

$$c^{(i)} := \arg\min_j \| x^{(i)} - \mu_j \|^2 \tag{1}$$

对于每一个类别 j，重新计算该类的聚类中心：

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \tag{2}$$

最后输出所有样本的类别隶属：$c^{(i)}$，$i=1,2,\cdots,m$。

(2)线性判别分析[3]

线性判别分析(Linear Discriminant Analysis，LDA)[3]是一种经典的有监督判别投影学习方法。与无监督投影学习方法 PCA[2]的不同之处在于，LDA 的优化目标是提高投影的判别性，即在投影后的特征空间里同类样本更加靠近，而不同类样本相互远离。

给定 N 个 m 维特征的训练样例 $X=\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\} \in R^{m*N}$，每个样本对应一个类标签 $y^{(i)}$，假设类别数为 C。

首先计算类内散度矩阵：

$$S_W = \sum_{i=1}^{C} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^{\mathrm{T}} \tag{3}$$

其中，μ_i 表示第 i 个类别所有样本的均值。然后计算类间散度矩阵：

$$S_B = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}} \tag{4}$$

其中，μ 表示所有样本的均值。求解判别投影：

$$W = \arg\max_W \frac{|W^{\mathrm{T}} S_B W|}{|W^{\mathrm{T}} S_W W|} \tag{5}$$

上式可以转化为一个泛化特征值问题 $S_W^{-1} S_B w_i = \lambda w_i$ 求解，最后输出判别投影 w_i，$i=1,2,\cdots,C-1$。

2.2 本文方法

所提出的基于前向无监督卷积神经网络的人脸表示学习方法的流程如图 1 所示。本方法的步骤包括：训练集设置与归一化、卷积核学习、人脸特征抽取和相似度计算。

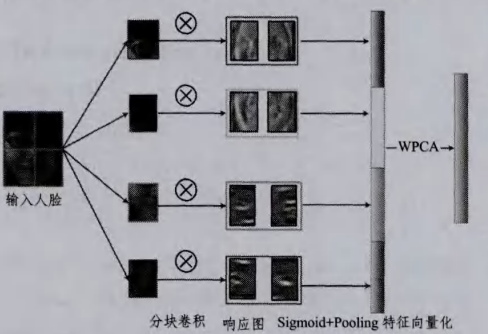

图 1 基于前向无监督卷积神经网络的人脸表示学习流程

(1)训练集设置与归一化

假设训练集中有 N 张人脸图像，将训练集记为 $T=[x_1, x_2, \cdots, x_N]$，为保证最终人脸表示的鲁棒性，训练集合中的图像应体现姿态、光照、表情和分辨率的变化。

(2)卷积核学习

卷积核学习可以看作是学习特征提取器。其基本思想为对每个人脸块提取人脸小块，通过聚类从数据中学习分布并赋予样本虚拟标签，进一步进行线性判别特征学习，获得卷积核。完整的流程如下：

将对齐后的人脸图像划分为 2×2 个相同大小 $width \times height$ 的块，对于每一个人脸块，首先在每个人脸块上提取 $k\times k$ 大小的小块，步长为 1，记为：

$$X_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m*n}] \in R^{k*k*m} \tag{6}$$

设置 $k=9$，其中：

$m = width - k + 1$，$n = height - k + 1$

减去所有人脸小块的均值，得到：

$$\overline{X_i} = [\overline{x_{i,1}}, \overline{x_{i,2}}, \cdots, \overline{x_{i,m*n}}] \in R^{k*k*m} \tag{7}$$

依次处理训练集中的每一幅图像，得到：

$$X = [\overline{X_1}, \overline{X_2}, \cdots, \overline{X_N}] \in R^{k*k*(m*N)} \tag{8}$$

通过 K-means 聚类，将所有的人脸小块聚类为 H 个类，通过线性判别分析算法学习判别投影，保留前 L 个判别投影作为卷积核。

$$D = [d_1, d_2, \cdots, d_L] \in R^{(k*k)*L} \tag{9}$$

其中，d_l 表示第 l 个判别投影，每个判别投影可以看作是一个独立的卷积核。分别按上述步骤对所有的 2×2 个人脸块进行卷积核学习。

(3)人脸特征提取

人脸的特征提取分为 4 个步骤：卷积、Sigmoid、最大值 Pooling 和 WPCA 降维。

1)卷积操作

对于输入的人脸图像，每个人脸块分别经过 L 个卷积核卷积操作之后，共得到 $4L$ 个响应图像。

2)Sigmoid 操作

通过式(10)所示的 Sigmoid 操作将卷积层输出的响应图中的响应值进行归一化，通过 Sigmoid 操作引入了特征的非线性特性。

$$Sigmoid(x)=\frac{1}{1+e^{-x}} \tag{10}$$

3)Pooling 操作

通过最大值 Pooling 操作来提取不变特征。首先，将每个卷积图像划分为 4×4 大小的不重叠的网格，取每个网格的最大值，得到最大值 Pooling 后的响应图。

对于输入的人脸 x_i，将 Sigmoid 与 Pooling 操作后的所有响应图特征拉直后得到的人脸表示记为 f_i。

4)WPCA 降维

利用白化主成分分析(WPCA)[2]对拉直后的人脸特征进行降维处理。首先计算训练集特征的协方差矩阵：

$$S_T=\sum_{k=1}^{N}(f_k-\mu)(f_k-\mu)^{\mathrm{T}} \tag{11}$$

其中，μ 表示所有特征的均值。

$$W_{pca}=\arg\max_{W}|W^{\mathrm{T}}S_TW|=[W_1,W_2,\cdots,W_m]\in R^{N*m} \tag{12}$$

其中，$\{W_i|i=1,2,\cdots,m\}$ 表示 S_T 的前 m 个最大特征值[e_1，e_2，…，e_m]所对应的特征向量，将特征值开根号后按以下形式组织为对角矩阵：

$$E=\begin{bmatrix}\frac{1}{\sqrt{e_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{e_n}}\end{bmatrix} \tag{13}$$

对于输入的特征 f_i，经过 WPCA 降维后得到：

$$y_i=E*W_{pca}^{T}*(f_i-\mu) \tag{14}$$

训练集中的所有特征经过降维后记为 $Y=\{y_1,y_2,\cdots,y_N\}$，再对特征进行二范数归一：

$$y=\frac{y}{\|y\|} \tag{15}$$

(4)相似度计算

基于 cosine 距离计算相似度。由于特征已经进行二范数归一，因此只需计算特征的内积：

$$sim(y_1,y_2)=y_1*y_2 \tag{16}$$

文中提出的基于前向无监督卷积神经网络的人脸表示方法，可以利用海量的无监督人脸数据学习鲁棒的、紧凑的人脸表示，同时网络结构简单，不需要反馈训练，因此训练速度显著优于有监督的深度卷积神经网络。

3 实验与分析

本节对所提方法进行实验与分析。首先介绍实验数据集和参数设置，然后给出实验结果与分析。

3.1 实验数据集

(1)LFW 人脸数据集[12]

LFW 的全称是 Labeled Face in the Wild，是目前最广泛采用的真实条件下的人脸识别测试数据集。LFW 中的人脸数据来自于雅虎新闻图片，其中包含来了来自 5749 个不同个体的共 13233 张图片，每个个体的图片数最少为一张，最多可达 500 多张。LFW 中的图像包含了丰富的姿态、光照、表情与老化，部分人脸图像示例如图 2 所示。

图 2 LFW 中人脸图像示例

LFW 的测试协议分为受控协议(Restricted Protocol)和非受控协议(Unrestricted Protocol)两种。其中非受控协议可利用训练数据完整的 Label 信息，而受控协议只能利用成对的训练样本。本文采用被广泛用于无监督人脸表示学习方法评测的受控协议。受控协议中共有 6000 对人脸，含 3000 对正例和 3000 对反例，所有人脸对被提取划分为十折，每折 300 对正例样本和 300 对反例样本。算法测试为十折交叉测试，将其中任意一折作为测试，剩余九折 2700 对正例和 2700 对反例进行训练，测试方法为判断测试集中一对人脸为正例或者反例，用十折平均精度作为算法性能的评价指标。

(2)Feret 数据集[16]

Feret 数据集是一个经典的实验室条件下的人脸数据集，包含来自 1199 人的 14126 张图像。常用于准正面人脸识别评测的数据集包括 Fb、Fc、Dup-I 和 Dup-II 这 4 个子集合，其中 Fb 包含 3 种人脸表情变化，Fc 集合中包含不同的光照条件，Dup-I 在 3～4 个月的采集周期内再次采集，Dup-II 至少在一年半的周期后再次采集。Feret 采用在 4 个给定集合上的首先识别率作为性能评价指标，示例人脸图像如图 3 所示。

图 3 Feret 数据集图像示例

3.2 参数设置

(1)人脸归一化。基于两眼中心位置，对人脸进行对齐，对齐后的人脸图像如图 4 所示，对齐后的人脸大小为 128 * 80 像素，从而分块之后每个人脸块的大小为 64 * 40 像素。

图 4 对齐后的人脸图像示例

(2)方法参数。本文设置卷积核的大小为 9×9,卷积核个数 $L=16$,WPCA 降维参数为 $PCA\_Dim=m=1000$。

3.3 实验结果与分析

(1)LFW 数据集实验

表 1 给出了本方法与经典的局部特征描述子 LBP[4]、Gabor[5]、SIFT[6]以及当前主流的无监督特征学习方法 PCANet[8]、K-means Net[9]和 SRFD[11]的实验对比,其中 Sqrt 加括号表示对特征取平方根。LBP、Gabo、SIFT 和 PCANet 均采用白化主成分分析方法进行进一步降维。为了对比的公平性,所有方法均采用相同的归一化为 120 * 80 像素的人脸。实验结果表明,提出的人脸表示方法取得了显著优于主流局部特征描述子的性能,并且与当前主流的无监督特征学习方法的性能可比。

表 1 本算法与对比方法在 LFW 数据集上的性能对比

方法	十折平均精度(%)
LBP[4]	79.50
LBP(Sqrt)[4]	80.51
Gabor[5]	82.30
Gabor(Sqrt)[5]	83.09
SIFT[6]	80.30
SIFT(Sqrt)[6]	80.79
PCANet[8]	84.56
K-means Net[9]	84.23
SRFD[11]	84.40
本文方法	84.90

进一步地,表 2 给出了聚类中心个数对最终算法性能的影响。设置聚类中心个数为 20 到 100,步长为 10。实验结果表明,本算法对聚类中心个数不敏感,且当聚类中心个数设置为 50 时,本文算法可达到最优性能。

表 2 不同聚类中心个数下本文算法在 LFW 数据集上的性能对比

聚类中心个数	十折平均精度(%)
20	84.50
30	84.75
40	84.75
50	**84.90**
60	84.70
70	84.63
80	84.58
90	84.50
100	84.45

(2)Feret 数据集实验

表 3 给出了本方法与经典的局部特征描述子 LBP[4]、P-LBP[18]、POEM[19]以及当前主流的无监督特征学习方法 PCANet[8]在 Feret 数据库上的实验对比。实验结果表明,本文提出的人脸表示方法的性能优于经典的局部描述子和当前主流的无监督特征学习方法 PCANet 的性能,其对表情、光照和时间跨度均具有较好的鲁棒性。

表 3 本文算法与对比方法在 Feret 数据集上的首选识别率对比(%)

Probe Sets	Fb	Fc	Dup-I	Dup-II	Avg.
LBP[4]	93.00	51.00	61.00	63.75	63.75
P-LBP[18]	98.10	98.50	81.60	83.20	89.60
POEM[19]	99.60	99.50	88.80	85.00	93.20
PCANet[8]	99.33	99.48	88.92	84.19	92.98
本文方法	**99.60**	**99.50**	**90.20**	**89.30**	**94.65**

结束语 为了利用海量的无标注人脸数据,围绕无监督人脸表示学习问题展开研究,提出了一种基于前向无监督卷积神经网络的人脸表示学习方法。其中,设计了一种基于 K-means 聚类获取训练样本虚拟标签,再利用线性判别分析进行卷积核学习的方法,提高了卷积核的判别性。本方法网络结构简单,训练速度快。实验结果进一步表明,提出的人脸表示学习方法在最具代表性的真实条件下的人脸数据集 LFW 和经典的实验室条件下的数据集 Feret 上均取得了优于当前主流的无监督特征学习方法和局部特征描述子的性能,可以有效用于无监督人脸表示学习。

参 考 文 献

[1] Zhao W, Chellappa R, Phillips P J, et al. Face recognition: A literature survey[J]. ACM Computer Surveys (CSUR), 2003, 35(4): 399-458

[2] Turk M, Pentland A. Eigenfaces for recognition[J]. Journal of cognitive neuroscience, 1991, 3(1): 71-86

[3] Belhumeur P N, Hespanha J P. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720

[4] Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face[J]. IEEE Transactions on Image Processing, 2002, 11(4): 467-476

[5] Ahonen T, Hadid A, Pietikainen M. Face Description with Local Binary Patterns: Application to Face Recogniton[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12): 2037-2041

[6] Vu N S, Caplier A. Enhanced patterns of oriented edge magnitudes for face revognition and image matching[J]. IEEE Transactions on Image Processing, 2011, 21(3): 1352-1365

[7] M Bicego, A Lagorio, E Grosso, et al. On the use of SIFT features for face authentication[C]// Computer Vision and Pattern Recognition Workshop, 2006(CVPRW'06). 2006: 35

[8] Chan T H, Jia K, Gao S, et al. PCANet: A Simple Deep Learning Baseline for Image Classification? [J]. IEEE Transactions on Image Processing, 2014, 24(12): 5017-5032

[9] Coates A, Ng A Y. Learning Feature Representation Using K-means[M]. Springer Berlin Heidelberg, 2012

[10] Hartigan J A, Wong M A. Algorithm AS 136: A K-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100-108

[11] Cui Z, Li W, Xu D, et al. Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild [C]// Proc. of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR: IEEE, 2013: 3554-3561

[12] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[C]// Proc. of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Marseille, France: Springer Berlin Heidelberg, 2008: 864-877

[13] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]// Proc. of the Advances in Neural Information Processing Systems. 2014: 1988-1996

[14] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Proc. of the Advances in Neural Information Processing Systems. 2012: 1097-1105

[15] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]// Proc. of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH: IEEE, 2014: 1701-1708

[16] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]// Proc. of the Advances in Neural Information Processing Systems. 2014: 1988-1996

[17] Phillips P J, Moon H, Rauss P J, et al. The FERET evaluation methodology for face recognition algorithms[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1997: 137-143

[18] Tan X, Triggs B. Enhanced local texture feature sets for face recognition under difficult lighting conditions[J]. IEEE TIP, 2010, 19(6): 1635-1650

[19] Vu N S, Caplier A. Enhanced patterns of oriented edge magnitudes for face recognition and image matching[J]. IEEE TIP, 2012, 21(3): 1352-1368

[20] Xie P, Wu X J. Modular Multilinear Principal Component Analysis and Application in Face Recognition[J]. Computer Science, 2015, 42(3): 274-279(in Chinese)
谢佩, 吴小俊. 分块多线性主成分分析及其在人脸识别中的应用研究[J]. 计算机科学, 2015, 42(3): 274-279

[21] Tian Hua, Pu Tian-yin. Improved ASM localization method for human facial features[J]. Journal of Chongqing Unversity Posts and Telecommunications (Natural Science Edition), 2014, 26(1): 124-130(in Chinese)
田华, 蒲天银. 一种改进的 ASM 人脸特征点定位方法[J]. 重庆邮电大学学报(自然科学版), 2014, 26(1): 124-130

(上接第 282 页)

的大小来区分变量间的差异在分布不均衡的数据集上效果无法让人满意。Vote 数据集也是一个比较特殊的数据集，所有的属性都是二值属性，由于属性变量分布太过简单，所以 3 种相似度算法的效果几乎没有区别。

与 SMS 算法相比，HDS 算法的准确度提高最少的是在 Vote 数据集上，只有 0.07%，提高最多的是在 Breast cancer 数据集上，提高了 17.85%，6 个数据集平均提高了 9.64%。与 ADD 算法相比，HDS 算法只在 Vote 数据集上的结果降低了 1%，在 Adult 数据集上最多提高 33%，平均提高10.63%。

在概率中，方差用来度量随机变量与其数学期望的之间的偏离程度，即数据波动幅度的大小，方差越小，说明算法的结果越稳定。表 4 所列为 3 种相似度算法在各数据集上聚类结果的方差，单位为 10^{-4}。容易看出，除了 Soybean 数据集外，HDS 聚类结果的方差都是最小的，说明在稳定性方面，HDS 优于 ADD 和 SMS。

表 4 聚类精度方差的比较(10^{-4})

相似度算法 \ 数据集	Adult	Breast cancer	CMC	Heart Discase	Soybean	Vote
HDS	11	71	2.1917	70	277	0.0130
SMS	63	351	8.8700	70	243	0.5206
ADD	140	92	18	164	323	0.0133

结束语 本文提出了一种基于 Hellinger 距离的分类变量相似度算法，它适用于有监督和无监督学习且支持混合变量数据集。实验结果证明，相比传统的 MVDM, SMD, ADD 等相似度算法，本方法不但适用范围更广，而且在准确度、有效性和稳定性方面都有较大提高。下一步的计划是将该算法应用于其他对混合数据集支持不足的机器学习领域。

参 考 文 献

[1] Han J, Kamber M, Pei J. Data mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2000, 5(4): 1-18

[2] Anderberg M R. Cluster Analysis for Applications[M]// Probability and Mathematical Statistics: A Serices of Monographs and Textbooks. 1973: ibc1-ibc2

[3] Gan G, Ma C, Wu J. Data clustering: theory, algorithms, and applications[M]// Data Clustering: theory, algorithms, and applications. Society for Industrial and Applied Mathematics, American Statistical Association, 2007: 44-51

[4] Hanneman R A, Riddle M. Introduction to social network methods[D]. Department of Sociology, University of California Riverside, 2005

[5] Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation [J]. Proceedings of the 2008 SIAM International Conference on Data Mining, 2008, 30(2): 243-254

[6] Huang Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining[C]// DMKD. 1998: 1-8

[7] Stanfill C, Waltz D. Toward memory-based reasoning [J]. Communications of the ACM, 1986, 29(12): 1213-1228

[8] Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with symbolic features[J]. Machine Learning, 1993, 10(1): 57-78

[9] Wilson D R, Martinez T R. Improved heterogeneous distance functions [J]. Journal of Artificial Intelligence Research, 1997, 6: 1-34

[10] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data [J]. Data & Knowledge Engineering, 2007, 63(2): 503-527

[11] Wang C, Cao L, Wang M, et al. Coupled nominal similarity in unsupervised learning [C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 973-978

[12] Liang J Y, Bai L, Cao F Y. K-Modes Clustering Algorithm Based on a New Distance Measure [J]. Journal of Computer Research and Development, 2010, 47(10): 1749-1755(in Chinese)
梁吉业, 白亮, 曹付元. 基于新的距离度量的 K-Modes 聚类算法[J]. 计算机研究与发展, 2010, 47(10): 1749-1755

[13] Cao F, Liang J, Li D, et al. A dissimilarity measure for the k-Modes clustering algorithm [J]. Knowledge-Based Systems, 2012, 26: 120-127

[14] Csiszaár I. Information-type measures of difference of probability distributions and indirect observations[M]. Studia Sci. Math. Hungar., 1967: 299-318

[15] Morimoto T. Markov processes and the H-theorem[J]. Journal of the Physical Society of Japan, 1963, 18(3): 328-331

[16] Ali S M, Silvey S D. A general class of coefficients of divergence of one distribution from another[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1966, 28(1): 131-142