基于特征子集与特征区分度的生物认证方法

王 娜,李劲松,姚明海
(渤海大学 信息科学与技术学院,辽宁 锦州 121013)

摘 要:生物认证是信息安全领域研究的热点问题,已经成为社会安全各个领域用于身份识别的重要技术手段。随着数字图像获取技术和采集设备的快速发展,生物认证图像数据在采集过程中往往会出现高维度、高冗余现象。为了解决生物认证数据在计算过程中出现的维度高、冗余信息多、计算复杂度高的问题,在生物数据处理过程中构建了基于特征子集与特征区分度的特征选择方法。该方法首先利用改进的随机子空间方法和费舍尔得分法分别对特征排序;然后,将两种方法选择的特征结果进行加权融合得到全新的特征排序;最后,利用顺序前向搜索策略进行特征选择。为验证方法的有效性,将该方法与传统方法分别在五个经典的生物认证数据库上进行了比较。实验结果证明该方法获得了非常高的识别准确度。

关键词:特征选择;随机子空间;费舍尔得分;生物认证;特征融合
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)12-0051-05
doi:10.3969/j.issn.1673-629X.2020.12.009

Biometric Authentication Method Based on Feature Subset and Feature Discrimination

WANG Na,LI Jin-song,YAO Ming-hai
(School of Information Science and Technology,Bohai University,Jinzhou 121013,China)

Abstract:Biometric authentication is a hotspot issue in the information security field,which has become an important technical means for identity recognition in various fields of social security. With the rapid development of digital image acquisition technology and shooting equipment,high dimension and high redundancy often appear in the process of biological image data acquisition. In this paper,the biometric methods based on feature subset and feature discrimination is proposed to solve the problem of the biometric authentication data high dimensionality and redundancy. Firstly,modified random subspace method and Fisher score method are employed to pre-rank the feature. Then,the new feature ranking is obtained by fusing the feature selection results. Finally,sequential forward search method is utilized to select the most significant feature subset. In order to verify its effectiveness,the proposed method is compared with the traditional method on five classic biometric authentication databases. The experiment shows that the proposed method has high recognition accuracy.

Key words:feature selection;random subspace;Fisher score;biometric authentication;feature fusion

0 引 言

生物认证方法就是指智能系统通过人体自身具有唯一性的生物或行为特征来验证人的身份。由于人体生物特征具有唯一、可靠、安全的特点,已经受到广大科研人员的广泛关注。利用人体特征进行身份识别的方法已经成为了社会安全和网络安全等领域进行身份识别的重要手段之一。基于人体生物特征的身份识别在社会医疗、案件侦破、金融服务、网络销售、公司考勤等领域都有广泛应用[1]。但随着图像处理技术的快速发展,数据样本的采集也变得非常便捷,但是数字图像技术的发展也使得采集的数据样本的维度会非常高,高维样本数据在运算中很容易产生维数灾难[2]。

特征选择方法就是从采集到的数据样本中挑选出少量且具有代表性的数据,实现原始数据维数的缩减,去掉冗余和干扰信息,提高预测准确率,进而加强对学习结果的理解等。近年来,特征选择方法在模式识别[3]、生物认证[4]、数字图像处理[5]等领域受到广大科研工作者的广泛关注。近年来,国内外学者提出各种特征选择方法,大致可分过滤式、封装式和启发嵌入式[6]。过滤式方法通过对特征重要性打分来进行特征

收稿日期:2019-12-25 **修回日期**:2020-04-24
基金项目:辽宁省自然科学基金项目(2019-ZD-0503);辽宁省教育科学技术项目(LQ2017004)
作者简介:王 娜(1981-),女,讲师,硕士,研究方向为模式识别、图像处理。

选择,方法简单、快速与学习算法无关。但是这种方法忽视了特征间的相关性。封装式方法通过训练和测试选定的分类器寻找特征子集,这种方法考虑了特征子集和分类器间的相互作用,但也需要付出较高的计算代价,容易出现过拟合。启发嵌入式方法将特征选择方法融入到学习模型构建过程中。因为封装式方法和嵌入式方法考虑到了和分类器的交互,因此在准确率上普遍优于过滤式方法,但过滤式方法具有简单、计算快速等特点,所以过滤式方法在特征选择中也占有很重要的位置。

通过对大量文献的分析和总结,在众多学者研究结果的基础上,提出基于特征子集与区分度的特征选择方法。首先利用随机子空间(random subspace method,RSM)和 Fisher 得分方法计算出特征排序,然后对其融合获得新的特征排序,最后根据顺序前向搜索方法筛选能够代表样本数据原始表达的特征子集。该方法既具备过滤式特征选择方法的简单、快速的特点,又具有封装式特征选择识别率高的特点;同时还考虑不同方法对特征进行打分后的融合策略。

1 相关方法

1.1 随机子空间特征选择方法

基于随机判别理论的随机子空间方法[7]采用随机抽样方式从原始特征数据空间中获得特征子集,被广泛应用到聚类分析、特征选择、降维等领域。RSM 通过随机构建特征子空间,在构建的结果中发现最优结果。

假设给定一组具有 n 个样本的训练集 $X = \{(x_i, y_i)\}_{i=1}^{n}$,其中 $x_i \in \Re$,表示第 i 个样本具有 D 维特征;$y_i \in \{+1, -1\}$ 表示第 i 个样本对应的类标签。$y_i = +1$ 表示正样本,$y_i = -1$ 表示负样本。设随机产生维数为 q 的随机子空间 T 个,满足大于阈值 th 条件的子空间 t 个,则 RSM 算法流程如下所示:

初始化:$i \leftarrow 0$, $t \leftarrow 0$, $C \leftarrow 0_{1\times D}$, th, T

do$i \leftarrow i + 1$

$f_i \leftarrow$ 随机产生 q 维子空间,s. t. $\sum_{j=1}^{p} f_{i,j} = q$

计算子空间 f_i 的预测准确率 s_i

如果 $s_i >$th 并且 $f_{i,j} = 1$

则 $C_j \leftarrow C_j + 1$,$t \leftarrow t + 1$,直到 $i = T$

$C_j \leftarrow C_j / t$

算法结束

输出:特征权重向量 C

输出结果 C 表示随机子空间算法得到的特征权重向量,C_j 越大说明该特征被选择的频率越高。

1.2 Fisher score

基于 Fisher 得分的算法是一种发现具备最好区分度的特征子集的有监督选择方法[8],其定义如式(1)所示:

$$F_j = \frac{n_{y=+1}(\mu_{y=+1}^{j} - \mu^{j})^2 + n_{y=-1}(\mu_{y=-1}^{j} - \mu^{j})^2}{n_{y=+1}(\sigma_{y=+1}^{j})^2 + n_{y=-1}(\sigma_{y=-1}^{j})^2} \tag{1}$$

其中,$\mu_{y=+1}^{j}$、$\sigma_{y=+1}^{j}$、$\mu_{y=-1}^{j}$ 和 $\sigma_{y=-1}^{j}$ 分别是正负样本第 j 个特征的均值和标准差,$n_{y=+1}$ 和 $n_{y=-1}$ 是正负样本的数量,μ^{j} 是全体样本第 j 个特征的均值。F_j 值越大区分能力就越强。

1.3 顺序前向搜索算法

顺序前向搜索算法(sequential forward search,SFS)[9]是一个前向搜索算法,其核心思想是每次增加一个能使识别率得到提升的特征,直到识别率不再发生改变。

2 基于特征子集与区分度的特征选择方法

该文提出的特征选择方法,分别利用随机子空间 RSM 和 Fisher 得分方法给出两个不同的特征排序。然后对特征数据被选中的频率和特征数据的 Fisher 得分进行有效融合,产生一个新的特征数据的排序,最后利用 SFS 方法选出最终的特征子集。

算法流程如图 1 所示。

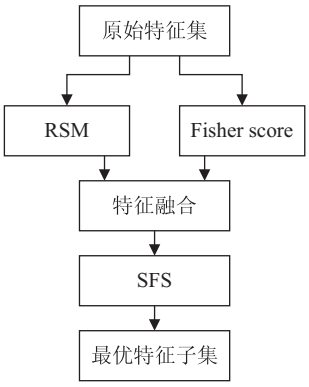

图 1 算法流程

融合公式如式(2)所示:

$$W_j = \widetilde{F_J} * \widetilde{C_J} \tag{2}$$

其中,$\widetilde{F_J}$ 和 $\widetilde{C_J}$ 分别表示对特征计算 Fisher 得分和 RSM 的频率进行归一化处理的结果,W 表示通过计算得到的权重向量。

经过了融合后,每一个特征都会拥有一个权重,根据权重可以得到一个初步的排序结果。权重越高说明该特征越重要,但是这些高权重的特征也有可能含有冗余信息,因此采用顺序搜索方法来剔除数据中的冗余信息,获得维度低、预测准确率高的特征。原始顺序搜索方法理论上也能够获得最优特征,但是原始顺序搜索算法的计算效率相对较低,不易实现。该文通过

对特征先预排序,在此基础上采用顺序前向搜索算法可以明显地提高算法的搜索效率。

3　实验结果与分析

为了验证文中方法的有效性,实验中的数据采用生物认证领域中常用于算法验证的五个生物识别数据库,并在实验前数据库数据进行预处理。实验中将文中提出的方法同多种特征选择方法在选择出的维度最高不超过 200 维的前提下进行对比。为了验证特征选择方法的实际使用效果,采用 K-nn 分类算法来验证。使用样本预测的准确率(predictive accuracy,PR)作为评价算法有效性的标准,具体计算方法如式(3)所示。为满足统计规律中覆盖样本数量的要求,全部实验中都采用 10 次随机取样的方法对算法有效性的验证。每次的测试都使用 50% 的样本用于训练分类模型,剩余的 50% 样本作为测试样本进行分类模型的测试。经实验统计 10 次的随机采样已经基本覆盖了 99% 的实验数据都参与了分类模型的训练和测试过程,计算获得的平均 PR 为最后结果。

$$PR = \frac{RP}{Num} \times 100\% \tag{3}$$

其中,Num 为测试样本个数,RP 为正确识别的样本个数。

3.1　在 FERET 数据库上的实验结果

FERET 数据库[10]是由美国国防部发起的人脸识别项目(face recognition technology,简称 FERET)数据库,在 1993 年到 19997 年创建,是生物认证领域普遍使用的算法验证数据库之一。FERET 库共有 1 428 个采集样本的 14 051 幅面部灰度图像。对比实验中选择了来自 72 个人的 432 幅图像,每个人选取了 6 幅不同姿态的图像,实验前对这 432 幅图像进行了预处理,将图像大小调整为 32×32 像素。图 2 展示了部分实验用图。

图 2　FERET 库中的部分人脸图像数据

由表 1 可以看出,文中提出的 IFS 方法在维数仅为 100 的前提下识别准确率就达到了 80.4%,明显高于其他方法。

3.2　在 ORL 人脸数据库上的实验结果

ORL 数据库[11]中包含了 400 幅人脸图像,这 400 幅图像是来自于 40 个人的不同面部表情图像。ORL 库中的图像具有表情和轻微的姿态变化,是人脸识别算法验证实验中经常使用的标准数据库。对比实验中将 ORL 库中的图像进行了预处理,实验中将数据库中的人脸图像进行处理,图像大小调整为 44×36 像素,图 3 展示了部分实验用图。

表 1　在 FERET 数据库上的实验对比结果

方法	预测准确率/%	选择维度
NonFS	76.94	1 024
MRMR	36.62	10
ChiSquare	70.83	70
T-test	73.94	200
Relief	73.89	200
InforGain	70.83	70
Gini	60.14	200
Kruskal-Wallis	60.14	200
Fisher	76.44	200
文中方法	80.4	100

图 3　ORL 库中的部分人脸图像数据

由表 2 可以看出,文中提出的 IFS 在维数仅为 100 时就具有较好的预测准确率。虽然其他方法也取得了较高的预测准确率,但是在维度选择上 IFS 方法要明显低于其他方法。

表 2　在 ORL 数据库上的实验对比结果

方法	预测准确率/%	选择维度
NonFS	93.70	1 584
MRMR	36.50	10
ChiSquare	89.15	180
T-test	82.20	200
Relief	84.40	200
InforGain	87.80	200
Gini	77.20	200
Kruskal-Wallis	77.30	200
Fisher	86.85	200
文中方法	88.10	100

3.3　在 CMU PIE 人脸数据库上的实验结果

CMU PIE 数据库[12]中包含了 41 368 幅人脸图像,这些图像是来自于 68 个人的不同面部表情图像。

CMU PIE 数据库中的图像包括了在不同姿态、光照和表情的轻微改变,是生物认证研究领域非常重要的测试数据库。文中采用文献[13]的方法对数据进行预处理,每个样本选取相同姿势、相同表情和有差异性光照的 21 幅进行实验,实验前对这些图像进行了预处理,将图像大小调整为 32×32 像素,图 4 展示了部分实验用图。

图 4 CMU PIE 库中的部分人脸图像数据

分析实验结果可以看出,所有方法的实验效果都很好,这是由于该数据库中人脸图像自身的问题,全部特征选择方法的识别率均达到了 90% 以上,个别算法达到 100% 。但文中提出的特征选择方法在选取维数相对较少时就取得了较好的实验效果(见表 3)。

表 3 在 CUMPIE 数据库上的实验对比结果

方法	预测准确率/%	选择维度
NonFS	99.92	1 024
MRMR	93.31	10
ChiSquare	99.83	160
T-test	99.58	200
Relief	99.98	200
InforGain	99.98	190
Gini	99.71	70
Kruskal-Wallis	100.00	190
Fisher	100.00	110
文中方法	99.97	100

3.4 在扩展的 YaleB 人脸数据库上的实验结果

扩展的 YaleB 库[14]中共有 38 人的 2 432 幅人脸图像,平均每个样本约 64 幅图像,扩展的 YaleB 库中的图像也包括面部表情差异和光照差异。实验前对这些图像进行了预处理,将图像大小调整为 32×32 像素。图 5 展示了部分实验用图。

图 5 扩展的 YaleB 库中的部分人脸图像数据

表 4 列出了不同方法的最高平均准确率,可以看到文中提出的特征选择方法在维数相对较低时就具有最高的识别准确率。

表 4 在扩展 YaleB 数据库上的实验对比结果

方法	预测准确率/%	选择维度
NonFS	63.04	1 024
MRMR	50.25	10
ChiSquare	73.74	200
T-test	67.41	100
Relief	70.58	200
InforGain	74.06	200
Gini	66.23	200
Kruskal-Wallis	65.22	200
Fisher	71.13	120
文中方法	81.8	100

3.5 在 CASIA 虹膜数据库上的实验结果

CASIA 虹膜库是由中国自主创建的用于生物识别验证的数据库,CASIA 虹膜库包含了 108 只眼睛的 756 幅虹膜图像,图 6 展示了部分实验用图。CASIA 虹膜库是生物认证领域应用最广泛的全公开数据库,已有全球 800 多家科研机构申请使用该数据库,近些年 CASIA 虹膜库已成为世界生物认证领域重要的数据支撑。对比实验中采用文献[15]中的数据处理方法对数据库中的图像进行了处理,提取了感兴趣的区域来验证实验效果。

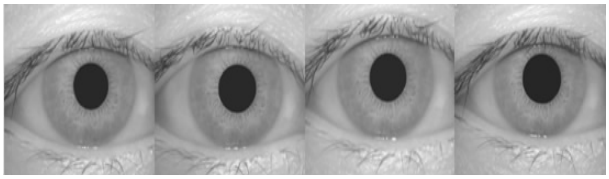

图 6 CASIA 库中的部分人脸图像数据

通过表 5 的实验可以看出,所有方法在选取 200 维特征数据的前提下识别率都不是很高。出现这一情况的主要原因在于实验前虹膜图像的系列预处理,这些预处理操作包括了在图像中定义感兴趣区域、压缩图像比例等,数据维数变为了原来的 1/75。由于数据预处理的效果较好,图像中的噪声数据和冗余数据已经基本被去除,这使得其他方法出现过收敛现象。即使这样文中方法与其他方法相比仍然取得了较好的实验效果。

表 5 在 CASIA 数据库上的实验对比结果

方法	预测准确率/%	选择维度
NonFS	93.73	2 048
MRMR	66.20	10
ChiSquare	88.09	200
T-test	87.65	200
Relief	88.92	200
InforGain	89.04	200

续表 5

方法	预测准确率/%	选择维度
Gini	77.47	180
Kruskal-Wallis	79.60	200
Fisher	88.02	200
文中方法	88.6	200

4　结束语

通过对比实验可以看出,提出的基于特征子集与特征区分度的生物认证方法适用于不同类型的数据库,并且在所有的对比实验中都在较低维数下取得了非常好的预测准确率。但在实际应用中还应当针对不同的实际问题进行详细分析。在该方法中特征权重的计算采用的是 Fisher 得分法,在今后的研究工作中应该对特征选择采用自适应的评价算法,相信会进一步提高算法的预测效果。

参考文献:

[1] 付　波,徐　超,赵熙临,等. 基于最值平均的人脸识别 LBP 算法[J]. 计算机应用与软件,2019,36(9):209-213.

[2] JAIN A K,DUIN R P W,MAO J. Statistical pattern recognition:a review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(1):4-37.

[3] 刘万军,孙　虎,姜文涛. 自适应特征选择的相关滤波跟踪算法[J]. 光学学报,2019,39(6):234-247.

[4] RAI Himanshu, YADAV Anamika. Iris recognition using combined support vector machine and Hamming distance approach[J]. Expert Systems with Applications,2014,41(2):588-593.

[5] 姚丽娟,李冬冬,王　喆. 基于 Relief 特征选择的心衰死亡率预测[J]. 计算机工程与应用,2018,54(23):125-130.

[6] LIU H, YU L. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4):491-502.

[7] KLEINBERG E. On the algorithmic implementation of stochastic discrimination[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(5):473-490.

[8] 俞景丽,胡恩良,张　涛. 一种新的 L1 度量 Fisher 线性判别分析研究[J]. 计算机工程与应用,2018,54(4):128-134.

[9] 闫光辉,李战怀. 两阶段无监督顺序前向分形属性规约算法[J]. 计算机研究与发展,2008,45(11):1955-1964.

[10] PHILLIPSP J,MOON H,RIZVI S A,et al. The FERET evaluation methodology for face recognition algorithms[J]. IEEE Transactions on Pattern Analysis and Machine Learning, 2000,22(10):1090-1104.

[11] SAMARIA F S, HARTER A C. Parameterisation of a stochastic model for human face identification[C]//Proceedings of the second IEEE workshop on applications of computer vision. Sarasota,USA:IEEE,1994:138-142.

[12] SIM T,BAKER S,BSAT M. The CMU pose,illumination, and expression database[J]. IEEE Transactions on Pattern Analysis and Machine Learning,2003,25(12):1615-1618.

[13] HE X,CAI D,NIYOGI P. Laplacian score for feature selection[J]. Advances in Neural Information Processing Systems,2006(18):507-514.

[14] LEE K C, HO J, KRIEGMAN D J. Acquiring linear subspaces for face recognition under variable lighting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(5):684-698.

[15] QI M, LU Y, LI J, et al. User-specific iris authentication based on feature selection[C]//International conference on computer science and software engineering. Wuhan: IEEE, 2008:1040-1043.