

DOI:10.11918/202306063

用于人脸情感识别的低冗余稀疏性集成剪枝

陈星, 李丹杨, 唐玉梅, 黄仕松, 吴义青

(贵州大学 大数据与信息工程学院, 贵阳 550025)

摘要: 为降低噪声和异常值对集成剪枝的影响, 并鲁棒性地选择更稀疏的基分类器子集, 从而提升人脸情感识别的性能, 提出了一种具有依赖分数的鲁棒性稀疏低冗余集成剪枝方法用于人脸情感识别。首先, 该方法将样本实例的预测结果视为基分类器特征, 分别利用互信息和熵来评估成对基分类器之间的依赖性和它们之间的优先级。其次, 将优先级依赖关系添加到基于回归的目标方程中实现冗余基分类器的修剪, 此目标方程使用 $l_{2,1}$ 范数来增加分类器子集的鲁棒性从而提升算法的泛化性能。然后, 将内积正则化项引入到目标方程中, 通过计算分类器特征系数向量内积的绝对值的和去选择稀疏和低冗余的基分类器。最后, 使用大多数投票法对选择的基分类器子集进行集成从而得到最终的识别结果。结果表明: 本文提出的方法在 FER2013、JAFPE、CK+ 和 KDEF 4 个公共人脸情感数据集上的识别准确率, 比所有基分类器进行集成得到的准确率分别高 3.29%、10.39%、1.76% 和 4.89%, 表明该方法可以选择出识别效果更好、冗余度更低的分类器子集, 提高集成剪枝的泛化能力。

关键词: 人脸情感识别; 集成剪枝; $l_{2,1}$ 范数; 内积正则化项; 依赖分数

中图分类号: TP3 **文献标志码:** A **文章编号:** 0367-6234(2024)05-0152-10

Sparsity and low-redundancy ensemble pruning for facial expression recognition

CHEN Xing, LI Danyang, TANG Yumei, HUANG Shisong, WU Yiqing

(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

Abstract: To reduce the impact of noise and outliers on ensemble pruning and robustly select a more sparse subset of base classifiers to improve the performance of facial expression recognition, a robust sparse and low-redundancy ensemble pruning method with dependency scores is proposed in this paper for facial expression recognition. First, the method treats the prediction results of sample instances as base classifier features, and evaluates the dependency and priority between pairs of base classifiers using mutual information and entropy, respectively. Second, the priority dependency is added to the regression-based objective equation to prune redundant base classifier, which uses the $l_{2,1}$ -norm to increase the robustness of the classifier subset and thus improve the generalization performance of the algorithm. Then, an inner product regularization term is introduced into the target equation to select sparse and low-redundant base classifiers by computing the sum of the absolute values of the inner products of the classifier feature coefficient vectors. Finally, the majority voting method is used to integrate the selected subset of base classifiers to obtain the final recognition results. The recognition accuracy of the proposed method on four public facial expression datasets, FER2013, JAFPE, CK+, and KDEF, is 3.29%, 10.39%, 1.76%, and 4.89% higher than those obtained by integrating all base classifiers, respectively, indicating that the method can select a subset of classifiers with better recognition results and lower redundancy, and improve the ensemble pruning generalization ability.

Keywords: facial expression recognition; ensemble pruning; $l_{2,1}$ -norm; inner product regularization term; dependency score

人脸情感识别是极具挑战性的机器学习任务, 图片完整程度、图片亮度、被拍摄者的肤色和种族以及拍照角度等因素的差异都会对情感识别形成干扰^[1-2]。因此, 如何提升情感识别的正确性是当前的国内外相关研究中的难点与重点。为了有效提升

情感识别准确率, 研究者多采用集成多基分类器的方式增强系统识别能力与鲁棒性^[3-5]。然而, 由于集成学习不可避免地引入了一些低识别率和冗余的基分类器, 这些“低质量”的基分类器不仅会降低最终集成的性能, 还会增加存储空间和运行时间的负

收稿日期: 2023-06-19; 录用日期: 2023-07-18; 网络首发日期: 2023-11-06

网络首发地址: <http://kns.cnki.net/kcms/detail/23.1235.T.20231103.1151.008.html>

基金项目: 贵州省科技计划项目(黔科合平台人才[2018]5781)

作者简介: 陈星(1998—), 男, 硕士研究生

通信作者: 李丹杨, danyangcl@163.com

担。因此,研究人员提出了集成剪枝方法^[6-7],这类方法将弱分类器和冗余分类器从基分类器池中剔除,保留最优能力的基分类器进行集成。相比传统集成学习,集成剪枝不仅降低了算法的时空复杂度,还提升了最终的识别效果。例如,文献[4]提出了一种聚类软标签的方法进行样本选择,进而进行集成剪枝,这种方法有效地提高了人脸情感的识别性能并降低了算法整体的时空复杂度。然而,如何设计一种方法剔除弱分类器和冗余分类器仍是一个需要不断研究的问题。

集成剪枝旨在选择“好而不同”的基分类器,以构建高质量的基分类器子集,这就意味着,选择的基分类器不仅是高识别性能,而且还是低冗余的。在集成剪枝领域,基分类器高识别性能和低冗余分别用准确率和多样性来表示,它们的值越高,选择的基分类器子集识别性能越高、冗余度更低。文献[8]证明了同时考虑基分类器的准确率和多样性能够获得更优秀、泛化能力更强的基分类器子集,从而提高集成模型的效果。此外,文献[7]提出了一种将频谱聚类和结构稀疏性集成于一个回归联合框架中的方法,可以无监督地选择具有较低冗余度和良好识别性能的基分类器。然而,这些方法都没有考虑到分类器噪声和异常值对集成剪枝的影响,也没有充分考虑基分类器之间的冗余关系。为了解决上述问

题,本文提出了一种基于回归的目标方程的方法,称为具有依赖分数的鲁棒性稀疏低冗余集成剪枝(robust sparse and low-redundancy ensemble pruning with dependencies score, RSLDS)。该方法使用 $l_{2,1}$ 范数表示损失函数,以实现基分类器进行鲁棒性选择。同时,它使用互信息和熵分别计算成对基分类器之间的依赖关系和优先级,并将它们结合起来,得到基分类器的依赖分数,从而更加准确地表示基分类器之间的依赖关系。此外,该方法使用内积化正则项来选择数量更少、更冗余的基分类器子集,从而进一步提高基分类器子集的识别性能和效率。

1 具有依赖分数的鲁棒稀疏低冗余集成修剪

本文所提出的 RSLDS 方法框架见图 1。此方法主要可以分为 3 个部分。首先, RSLDS 通过训练多个深度神经网络模型来创建基分类器池。其次, RSLDS 设计了一种基于回归优化的集成剪枝方法,以平衡基分类器的准确率和多样性,并对分类器进行排序。最后,该方法从最新的基分类器序列中选择前 k 个基分类器,并使用多数投票对基分类器子集的预测标签进行集成,在测试集上测试该方法的效果。

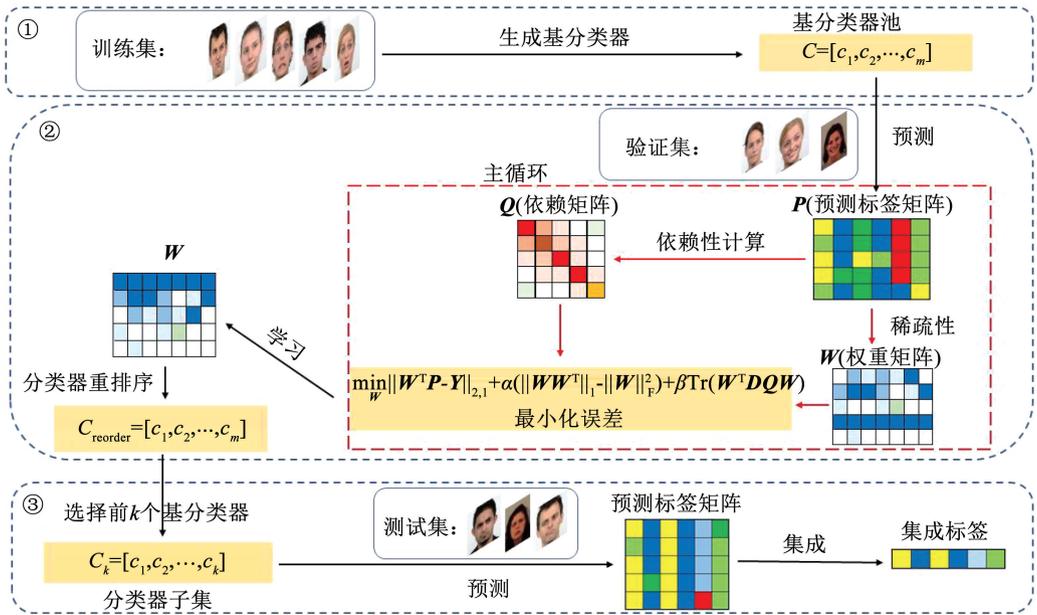


图 1 RSLDS 的框架

Fig. 1 Framework of RSLDS

1.1 符号

在整篇文章中向量 A_i 和 $A_{.j}$ 分别表示矩阵 A 的行和列, A_{ij} 表示矩阵 A 在 i 行 j 列的元素, 方阵 A 的迹和转置分别表示为 $\text{Tr}(A)$ 和 A^T , $\|A\|_1 =$

$\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|$ 、 $\|A\|_{2,1} = \sum_{i=1}^m \sqrt{\sum_{j=1}^n A_{ij}^2}$ 和 $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$ 分别表示 A 的 l_1 范数、 $l_{2,1}$ 范数和 Frobenius 范数。

在本文中,用 $V = \{x_i, y_i\}_{i=1}^n$ 表示一组标签实例, $X = \{x_1, x_2, \dots, x_n\} \in R^{d \times n}$ 是一个表示所有实例特征空间的 $d \times n$ 维的矩阵, $P \in R^{m \times n}$ 表示 m 个基分类器对具有 n 个实例的预测标签矩阵。基本分类器 c_i 是一个映射: $X \rightarrow Y$ 从特征空间 X 到标签空间 Y 。集成剪枝生成一组基本分类器 $C = \{c_i\}_{i=1}^m$, 它们被集成以产生最终的预测。 $Y \in R^{n \times c}$ 表示 c 维标签空间中有 n 个实例的标签矩阵, 其中当第 i 个实例与第 j 个标签相关联时, Y_{ij} 表示 1, 否则 Y_{ij} 表示 0。

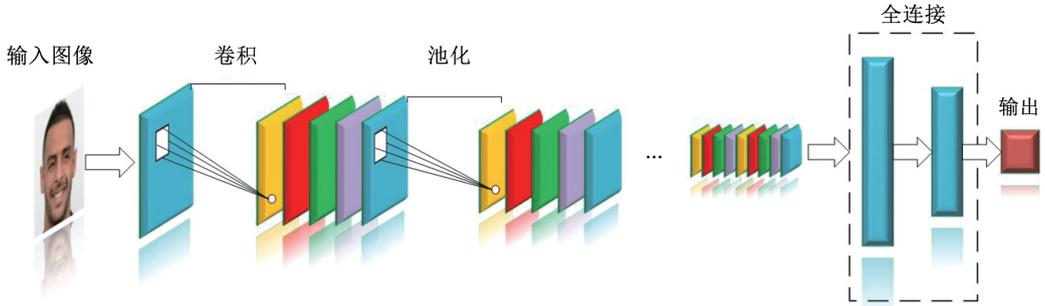


图 2 卷积神经网络的结构框图

Fig. 2 Block diagram of structure of convolutional neural network

表 1 生成的卷积神经网络模型参数范围

Tab. 1 Range of generated convolutional neural network model parameters

参数	层的数量	卷积核的大小	池化核的大小	池化步长	随机失活率	特征映射的数量	学习率	维度
范围	3~9	从[4 4]到[8 8]	从[2 2]到[5 5]	2	0.4~1	16,32,48,64,80,96	0.01,0.03,0.05	7,1 024,2 048,3 072

基分类器对样本的预测结果更加能反映基分类器的学习能力^[3], 而不是样本的特征。因此, 本文使用基分类器对样本的预测值作为基分类器的特征来进行分类器选择。本文用 $c_j(x_i)$ 表示第 j 个分类器对第 i 个样本实例的一个预测值, 则第 m 个基分类器对第 n 个样本实例的预测矩阵 $P \in R^{m \times n}$ 可以表达为

$$P = \begin{bmatrix} c_1(x_1) & \dots & c_1(x_n) \\ \vdots & & \vdots \\ c_m(x_1) & \dots & c_m(x_n) \end{bmatrix} \quad (1)$$

1.3 目标方程

传统基于回归的集成剪枝的目标方程一般形式为

$$\min_w J(W^T P, Y) + \alpha \Omega(W) \quad (2)$$

式中: $J(W^T P, Y)$ 是损失函数; $W \in R^{m \times c}$ 是权重矩阵, $\Omega(W)$ 是正则化项, 迫使 W 行稀疏性, $\alpha \geq 0$ 表示一般正则化项 $\Omega(W)$ 的正则化参数; c 是类别的数量; $Y = [y_1, \dots, y_n] \in \{0, 1\}^{c \times n}$ 是热编码标签矩阵, 其中, $y_i \in \{0, 1\}^c$ 表示样本实例 x_i 的标签向量。

传统的方法对式(2)中的损失函数 $J(W^T P, Y)$ 采用了最小二乘回归模型, 正则化项 $\Omega(W)$ 通常使用 $l_{2,1}$ 范数^[9], 则式(2)可表示为

1.2 构建基分类器特征

集成剪枝的有效性在很大程度上取决于基分类器的准确性和多样性。卷积神经网络在人脸情感识别中具有很强的识别效果^[2], 其结构框图见图 2。为了保证最终实验具有良好的识别性能, 本文使用卷积神经网络作为基础分类器, 通过修改卷积神经网络的参数和网络结构来构建一系列基分类器 $C = \{c_i\}_{i=1}^m$ 。这些卷积神经网络模型的参数值范围见表 1。

$$\min_w \|W^T P - Y\|_F^2 + \alpha \|W\|_{2,1} \quad (3)$$

此目标函数旨在找到最小化投影数据 $W^T P$ 和标签 Y 之间差异的分类器, 同时通过对权重矩阵 W 施加 $l_{2,1}$ 范数使 W 强制稀疏性, 根据权重矩阵 W 判断基分类器优先级, 从而实现基分类器子集的选择。使用最小二乘回归模型对噪声和异常值敏感, 可能会导致选择错误的基分类器, 文献[10]进一步证明了使用 $l_{2,1}$ 范数的损失函数具有更强的鲁棒性。因此, 本文将式(3)中的损失函数 $J(W^T P, Y)$ 改为 $l_{2,1}$ 范数, 提升基分类器选择的鲁棒性, 从而增加基分类器子集的泛化性能。式(3)可以重新表达为

$$\min_w \|W^T P - Y\|_{2,1} + \alpha \|W\|_{2,1} \quad (4)$$

为了进一步学习最优权重, 考虑冗余的基分类器对集成剪枝的影响, 互信息能够捕捉随机变量之间的非线性统计依赖^[11], 可采用互信息来评估分类器之间的依赖关系。依赖分类器又称为冗余分类器。2 个分类器 c_i 和 c_j 的预测标签之间的互信息 $MI(c_i, c_j)$ 定义为 $MI(c_i, c_j) = H(c_i) + H(c_j) - H(c_i, c_j)$, 其中 $H(c_i) = -\sum P(c_i) \log P(c_i)$ 是联合熵, 它使用联合概率函数 $P(c_i)$ 来衡量 c_i 中包含的信息。本文使用互信息来构建多分类器预测结果的依赖矩阵

\mathbf{Q} 。此矩阵的每个元素定义为

$$\mathbf{Q}_{i,j} = \begin{cases} \text{MI}(c_i, c_j) & \text{if } i \neq j \\ \sum_{k \neq i} \mathbf{Q}_{i,k} & \text{if } i = j \end{cases} \quad (5)$$

将依赖矩阵 \mathbf{Q} 合并到式(4)中,可得

$$\min_{\mathbf{W}} \|\mathbf{W}^T \mathbf{P} - \mathbf{Y}\|_{2,1} + \alpha \|\mathbf{W}\|_{2,1} + \beta \text{Tr}(\mathbf{W}^T \mathbf{Q} \mathbf{W}) \quad (6)$$

$$\text{Tr}(\mathbf{W}^T \mathbf{Q} \mathbf{W}) = \text{Tr}[\mathbf{W}^T \text{diag}(\mathbf{Q}) \mathbf{W}] + \text{Tr}[\mathbf{W}^T \text{diag}(\mathbf{Q})' \mathbf{W}] =$$

$$\sum_{i=1}^{m-1} \sum_{j=i+1}^m (\langle \mathbf{W}_i, \mathbf{W}_i \rangle + \langle \mathbf{W}_j, \mathbf{W}_j \rangle) \text{MI}(c_i, c_j) + \sum_{i=1}^{d-1} \sum_{j=i+1}^d 2 \langle \mathbf{W}_i, \mathbf{W}_j \rangle \text{MI}(c_i, c_j) =$$

$$\sum_{i=1}^{m-1} \sum_{j=i+1}^m \langle \mathbf{W}_i + \mathbf{W}_j, \mathbf{W}_i + \mathbf{W}_j \rangle \text{MI}(c_i, c_j) = \sum_{i=1}^{m-1} \sum_{j=i+1}^m \|\mathbf{W}_i + \mathbf{W}_j\|_2^2 \text{MI}(c_i, c_j) \quad (7)$$

式中: $\text{diag}(\mathbf{Q})$ 表示 \mathbf{Q} 的对角线部分元素,而 $\text{diag}(\mathbf{Q})'$ 表示 \mathbf{Q} 的非对角线部分元素。用 $\|\mathbf{W}_i\|_2$ 表示 i 维基分类器的重要性。根据式(7),当 $\text{MI}(c_i, c_j)$ 较大时, $\|\mathbf{W}_i + \mathbf{W}_j\|_2$ 较小,由于 \mathbf{W}_i 和 \mathbf{W}_j 都是向量,较小的 $\|\mathbf{W}_i + \mathbf{W}_j\|_2$ 并不能保证 \mathbf{W}_i 和 \mathbf{W}_j 同时较小。因此,该方法存在次优解,为了增加冗余基分类器的优先级,将熵纳入式(7)中,新的优化问题为

式中 $\beta > 0$ 也是一个正则化参数。对于式(6)的第3项,文献[12]发现了该表达式存在2个缺点。首先,当2个特征之间的互信息较高时,并不能保证它们的优先级较低,无法判断哪个特征更重要。其次,它平等地考虑了所有的特征。 $\text{Tr}(\mathbf{W}^T \mathbf{Q} \mathbf{W})$ 可以表述为

$$\min_{\mathbf{W}} \|\mathbf{W}^T \mathbf{P} - \mathbf{Y}\|_{2,1} + \alpha \|\mathbf{W}\|_{2,1} + \beta \sum_{i=1}^{m-1} \sum_{j=i+1}^m (\text{H}(c_j) \|\mathbf{W}_i\|_2 + \text{H}(c_i) \|\mathbf{W}_j\|_2) \text{MI}(c_i, c_j) \quad (8)$$

该公式的第3项表明,如果 c_i 和 c_j 之间的互信息较大,那么 c_i 和 c_j 不应该同时重要。如果 $\text{H}(c_i)$ 更大,那么 c_i 具有更多的不确定性或多样性,应该比 c_j 更重要。式(8)不包含 $\text{MI}(c_i, c_i)$,因此 $\text{MI}(c_i, c_i)$ 可以定义为0,式(8)的第3部分可以重写为

$$\sum_{i=1}^{m-1} \sum_{j=i+1}^m (\text{H}(c_j) \|\mathbf{W}_i\|_2 + \text{H}(c_i) \|\mathbf{W}_j\|_2) \text{MI}(c_i, c_j) =$$

$$\frac{1}{2} \sum_{i,j=1}^m (\text{H}(c_j) \|\mathbf{W}_i\|_2 + \text{H}(c_i) \|\mathbf{W}_j\|_2) \text{MI}(c_i, c_j) =$$

$$\frac{1}{2} \left[\sum_{i,j=1}^m \text{H}(c_j) \|\mathbf{W}_i\|_2 \text{MI}(c_i, c_j) + \sum_{i,j=1}^m \text{H}(c_i) \|\mathbf{W}_j\|_2 \text{MI}(c_i, c_j) \right] =$$

$$\frac{1}{2} \left[\sum_{i,j=1}^m \text{H}(c_j) \|\mathbf{W}_i\|_2 \text{MI}(c_i, c_j) + \sum_{j,i=1}^m \text{H}(c_j) \|\mathbf{W}_i\|_2 \text{MI}(c_i, c_j) \right] =$$

$$\sum_{i,j=1}^m \text{H}(c_j) \|\mathbf{W}_i\|_2 \text{MI}(c_i, c_j) = \sum_{i=1}^m \|\mathbf{W}_i\|_2 \sum_{j=1}^m \text{H}(c_j) \text{MI}(c_i, c_j) =$$

$$\sum_{i=1}^m \|\mathbf{W}_i\|_2 \left[\text{H}(c_i) \text{MI}(c_i; c_i) + \sum_{j \neq i} \text{H}(c_j) \text{MI}(c_i, c_j) \right] =$$

$$\sum_{i=1}^m \|\mathbf{W}_i\|_2 \sum_{j \neq i} \text{H}(c_j) \text{MI}(c_i; c_j) = \text{Tr}(\mathbf{W}^T \mathbf{D} \mathbf{Q} \mathbf{W}) \quad (9)$$

式中 $\mathbf{D} \in R^{m \times m}$ 和 $\mathbf{Q} \in R^{m \times m}$ 是对角矩阵,并且矩阵 \mathbf{D} 和 \mathbf{Q} 的 i_{th} 对角元素分别为 $\frac{1}{2} \|\mathbf{W}_i\|_2^{-1}$ 和 $\sum_{j \neq i} \text{H}(c_j) \text{MI}(c_i, c_j)$ 。式(8)可以被重新表达为

$$\min_{\mathbf{W}} \|\mathbf{W}^T \mathbf{P} - \mathbf{Y}\|_{2,1} + \alpha \|\mathbf{W}\|_{2,1} + \beta \text{Tr}(\mathbf{W}^T \mathbf{D} \mathbf{Q} \mathbf{W}) \quad (10)$$

式(10)的第1项迫使本文选择识别性能好的基分类器,第2项使用 $l_{2,1}$ 范数迫使权重矩阵 \mathbf{W} 行稀疏性,而第3项使用互信息和熵表示基分类器依赖分数,进而评估分类器的冗余程度。3项约束条件同

时作用于权重矩阵 \mathbf{W} ,并且通过限制 \mathbf{W} 中行的分布情况来进行基分类器的选择。然而,该式的第2项使用具有 $l_{2,1}$ 范数约束的正则化项在基分类器选择过程中忽略了基分类器与高冗余信息基分类器之间的相关性,从而导致容易陷入次优解。文献[13]设计了一个新的范数,称为内积正则化项,它通过计算分类器特征系数向量内积的绝对值的和,从而选择数量更少和冗余度更低的基分类器子集。本文将这个范数应用于权重矩阵 \mathbf{W} 上,内积正则化项的公式为

$$\begin{aligned}\Omega(\mathbf{W}) &= \sum_{i=1}^m \sum_{j=1, i \neq j}^m | \langle \mathbf{W}_{i \cdot}, \mathbf{W}_{j \cdot} \rangle | = \\ &= \sum_{i=1}^m \sum_{j=1}^m | \langle \mathbf{W}_{i \cdot}, \mathbf{W}_{j \cdot} \rangle | - \sum_{i=1}^m | \langle \mathbf{W}_{i \cdot}, \mathbf{W}_{j \cdot} \rangle | = \\ &= \| \mathbf{W} \mathbf{W}^T \|_1 - \text{Tr}(\mathbf{W} \mathbf{W}^T) = \| \mathbf{W} \mathbf{W}^T \|_1 - \| \mathbf{W} \|_F^2\end{aligned}\quad (11)$$

式中 $\Omega(\mathbf{W})$ 越小, 去除的无效和冗余基分类器越多, 即如果 $\mathbf{W}_{i \cdot}$ 和 $\mathbf{W}_{j \cdot}$ 之间的冗余足够高, 则 $\mathbf{W}_{i \cdot}$ 和 $\mathbf{W}_{j \cdot}$ 之间的内积很大, 在不考虑这部分冗余分类器后, 使得 $\Omega(\mathbf{W})$ 更小。从而提高了正则化项的惩罚强度, 保证了基分类器之间的低冗余度。将式(11)与式(10)结合可以得到

$$\min_{\mathbf{W}} \| \mathbf{W}^T \mathbf{P} - \mathbf{Y} \|_{2,1} + \alpha (\| \mathbf{W} \mathbf{W}^T \|_1 - \| \mathbf{W} \|_F^2) + \beta \text{Tr}(\mathbf{W}^T \mathbf{D} \mathbf{Q} \mathbf{W}) \quad (12)$$

1.4 优化

为了解决式(12)中的第 1 项在求解时 $l_{2,1}$ 范数难以计算的问题, 本文引入了一个辅助变量 \mathbf{K} , 式(12)被重新表述为

$$\min_{\mathbf{W}} \| \mathbf{K} \|_{2,1} + \alpha (\| \mathbf{W} \mathbf{W}^T \|_1 - \| \mathbf{W} \|_F^2) + \beta \text{Tr}(\mathbf{W}^T \mathbf{D} \mathbf{Q} \mathbf{W}) \quad (13)$$

式中 $\mathbf{K} = \mathbf{W}^T \mathbf{P} - \mathbf{Y}$ 。这个公式是凸的, 因此, 可以使用 ADMN^[14] 来解式(13), 式(13)相应的增广拉格朗日方程为

$$\begin{aligned}L(\mathbf{W}, \mathbf{K}, \mathbf{Y}) &= \| \mathbf{K} \|_{2,1} + \\ &+ \alpha (\| \mathbf{W} \mathbf{W}^T \|_1 - \| \mathbf{W} \|_F^2) + \beta \text{Tr}(\mathbf{W}^T \mathbf{D} \mathbf{Q} \mathbf{W}) + \\ &+ \text{Tr}[\mathbf{E}^T (\mathbf{K} - \mathbf{W}^T \mathbf{P} + \mathbf{Y})] + (\mu/2) \| \mathbf{K} - \mathbf{W}^T \mathbf{P} + \mathbf{Y} \|_F^2\end{aligned}\quad (14)$$

式中 \mathbf{E} 和 μ 分别为拉格朗日乘子和惩罚项。

1.4.1 更新 \mathbf{W}

式(14)相对于 \mathbf{W} 的最小化表示为

$$\begin{aligned}\min_{\mathbf{W}} \alpha (\| \mathbf{W} \mathbf{W}^T \|_1 - \| \mathbf{W} \|_F^2) + \\ \beta \text{Tr}(\mathbf{W}^T \mathbf{D} \mathbf{Q} \mathbf{W}) + (\mu/2) \| \mathbf{K} - \mathbf{W}^T \mathbf{P} + \mathbf{Y} - \mathbf{E}/\mu \|_F^2\end{aligned}\quad (15)$$

当式(15)相对于 \mathbf{W} 的导数设置为零时, 可以得到

$$\begin{aligned}2\alpha (\mathbf{I}_{dd} \mathbf{W} - \mathbf{W}) + 2\beta \mathbf{D} \mathbf{Q} \mathbf{W} - \mu \mathbf{P} \mathbf{K}^T + \\ \mu \mathbf{P} \mathbf{P}^T \mathbf{W} - \mu \mathbf{P} \mathbf{Y}^T + \mathbf{P} \mathbf{E}^T = 0\end{aligned}\quad (16)$$

从而有

$$\begin{aligned}\mathbf{W} = [2\alpha (\mathbf{I}_{dd} - \mathbf{1}_{dd}) + \\ 2\beta \mathbf{D} \mathbf{Q} + \mu \mathbf{P} \mathbf{P}^T]^{-1} \mathbf{P} (\mu \mathbf{K}^T + \mu \mathbf{Y}^T - \mathbf{E}^T)\end{aligned}\quad (17)$$

1.4.2 更新 \mathbf{K}

式(14)相对于 \mathbf{K} 的最小化表示为

$$\min_{\mathbf{K}} \| \mathbf{K} \|_{2,1} + (\mu/2) \| \mathbf{K} - \mathbf{J} \|_F^2 \quad (18)$$

式中 $\mathbf{J} = \mathbf{W}^T \mathbf{P} - \mathbf{Y} + \mathbf{E}/\mu$ 。关于 \mathbf{K} 的优化与关于 \mathbf{W} 的优化相同。因此可以快速得到:

$$\begin{aligned}2\mathbf{M} \mathbf{K} + \mu \mathbf{K} - \mu \mathbf{J} = 0 \\ \Rightarrow \mathbf{K} = (2\mathbf{M} + \mu \mathbf{I})^{-1} \mu \mathbf{J}\end{aligned}\quad (19)$$

式中 \mathbf{M} 是一个对角矩阵且对角元素为 $M_{ii} =$

$$\frac{1}{2 \| \mathbf{K}_{i \cdot} \|_2^2}$$

1.4.3 更新 \mathbf{E} 和 μ

在更新其他变量后, 根据 ADMN 的规则和文献[10]更新拉格朗日乘数 \mathbf{E} 和惩罚参数 μ , 即:

$$\mathbf{E} = \mathbf{E} + \mu (\mathbf{K} - \mathbf{W}^T \mathbf{P} + \mathbf{Y}) \quad (20)$$

$$\mu = \min(\rho \mu, \mu_{\max}) \quad (21)$$

式中: ρ 是收敛条件控制参数; μ_{\max} 是最大惩罚参数, 通常设置为 10^6 。

使用 ADMN 方法求解式(14)满足一般条件的收敛定理^[14]。在优化 RSLDS 的目标函数时, 最耗时的操作是计算矩阵 \mathbf{W} , 时间复杂度为 $O(cmn)$ 。RSLDS 标准化过程的伪代码见算法 1。

算法 1: PSLDS

输入: 一组经过训练的基分类器 $C = \{c_i(x)\}_{i=1}^m$; 验证集数据 $V = \{(x_i, x_j)\}_{i=1}^n$; 测试集数据 $T = \{(x_i, x_j)\}_{i=1}^n$; 选择基分类器的数量 k

输出: k 个基分类器, 集成剪枝的准确率

1. 初始化: $\mathbf{W} \in R^{m \times c}$, $\mathbf{K} \in R^{c \times n}$, $\mathbf{D} \in R^{m \times m}$, $\mathbf{M} \in R^{c \times c}$, $\mathbf{E} \in R^{c \times n}$, parameters: α, β, ρ, μ

2. 使用式(1)生成验证集预测标签矩阵 $\mathbf{P} \in R^{m \times n}$

3. 生成验证集热编码标签矩阵 $\mathbf{Y} \in R^{c \times n}$

4. 使用 $\sum_{j \neq i}^m H(c_j) \text{MI}(c_i, c_j)$ 生成验证集预测标签依赖矩阵 $\mathbf{Q} \in R^{m \times m}$

5. 重复

6. $\mathbf{K} \leftarrow (2\mathbf{M} + \mu \mathbf{I})^{-1} \mu (\mathbf{W}^T \mathbf{P} - \mathbf{Y} + \mathbf{E}/\mu)$

7. $\mathbf{E} \leftarrow \mathbf{E} + \mu (\mathbf{K} - \mathbf{W}^T \mathbf{P} + \mathbf{Y})$

8. $\mu \leftarrow \min(\rho \mu, \mu_{\max})$

$$9. \quad \mathbf{W} \leftarrow [2\alpha(I_{mm} - 1_{mm}) + 2\beta\mathbf{D}\mathbf{Q} + \mu\mathbf{P}\mathbf{P}^T]^{-1} \mathbf{P}(\mu\mathbf{K}^T + \mu\mathbf{Y}^T - \mathbf{E}^T)$$

10. 检查收敛条件 $\|\mathbf{W}^T \mathbf{P} - \mathbf{Y} - \mathbf{E}\|_{\infty} \leq \varepsilon$

11. 直到收敛

12. 通过计算 $\|\mathbf{W}_i\|_2$ 并按降序对所有基本分类器进行排序,选择排名前 k 个基本分类器在测试集中使用多数投票进行集成

1.5 分类器融合

由于本文主要研究基分类器的选择,因此分类器融合策略采用大多数投票方法。多数投票的过程可以表示如下:

$$C_i = \arg \max_e \sum_{j=1}^k \varphi(c_j(x_i) = e) \quad (22)$$

式中 $e = 1, 2, \dots, 7$; 如果基分类器 j 对样本实例 i 的预测标签 $c_j(x_i)$ 等于 e , 则 $\varphi(\cdot) = 1$, 否则, $\varphi(\cdot) = 0$ 。

为了更加清晰地展示集成剪枝在现实中是如何进行人脸情感识别,本文使用一个例子来进行说明。首先,在通用人脸情感数据集训练基分类器模型;然后,使用 RSLDS 选择最优能力的基分类器子集;其次,使用摄像头采集人物图片,再使用 OpenCV 等工具识别出相关人脸区域;最后,基分类器子集对人脸图片进行情感识别,并使用多数投票对多个模型的预测进行集成从而得到最终预测的情感分类。当基分类器子集的大多数模型都预测错误时,就会导致最终集成的预测出现误差,因此可以使用识别能力

更好的基分类器组建分类器池从而提高最终的分类性能。

2 实验结果和分析

2.1 实验环境与数据集

本文所有实验均在 Windows 10 专业版下完成,电脑处理器为 11th Gen Intel (R) Core (TM) i7-11800H@2.30 GHz 2.30 GHz,运行内存大小为 16 GB,显卡为 NVIDIA GeForce RTX 3060 Laptop GPU,本文使用 TensorFlow-GPU2.5 完成神经网络的搭建,相关软件为 Anaconda3.2、PyCharm2021.3 和 Matlab R2022b。

本文使用 4 个真实的人脸情感数据集进行实验,分别是 FER2013^[15]、JAFPE^[3]、CK +^[16] 和 KDEF^[17]。本文使用 Opencv 工具识别图片中的人脸区域,并将其裁剪缩放为 48×48 ,所有数据集的情感类别统一为 7 类,每种情感的详细信息见表 2。

表 2 上述 4 个数据集中每种情感的详细信息

Tab.2 Detailed information of each emotion in four datasets

数据集	生气	厌恶	害怕	开心	悲伤	惊喜	中性	总计
FER2013	4 953	547	5 121	8 989	6 077	4 002	6 198	35 887
JAFPE	30	29	32	31	31	30	30	213
CK +	135	177	75	207	84	249	593	1 520
KDEF	70	70	70	70	70	70	70	490

2.2 实验设置

FER2013 数据集可分为训练集、验证集和测试集 3 个部分。本文使用训练集训练 231 个卷积神经网络基分类器,然后使用 RSLDS 方法在验证集中选择合适的基分类器子集,最后在测试集、JAFPE、CK + 和 KDEF 上测试基分类器子集的效果。

本文使用准确率 (Accuracy, A_{cc}) 和分类器子集的数量作为评价指标,其中,准确率 Acc 的计算公式为

$$A_{cc} = \left(\frac{1}{n} \sum_{l=1}^n \frac{N_{Correct}}{N_{All}} \right) \quad (23)$$

式中 $N_{Correct}$ 和 N_{All} 分别表示为集成后正确预测样本实例和所有样本实例的数量。

2.3 参数敏感性分析和选择

本文的方法需要 2 个正则化参数 α 和 β , 以及 2 个 ADMN 方法的参数 ρ 和 μ 。 α 控制矩阵 \mathbf{W} 的稀疏性。理论上, α 越大,强调 \mathbf{W} 的稀疏性越大,而 β 控制冗余分类器数量的减少程度,较大的 β 强调更多冗余分类器的消除。

在选择 α 和 β 的最优参数时,根据文献[10]分别将 ρ 和 μ 的值临时设置为 0.001 和 1.2。首先,本文通过网格搜索策略从 $\{10^{-6}, 10^{-5}, \dots, 10^5, 10^6\}$ 中选择参数 α 和 β 。RSLDS 方法在不同 α 和 β 下的准确率见图 3,对于 $\alpha\beta$ 不同数据集,分别选择最高准确率对应的 α 和 β 为最终参数。可以发现,相比于 α ,本文方法对 β 更加敏感,这说明考虑分类器之间的冗余是正确的。在 α 和 β 的值确定之后,再使

用网格搜索的方法确定 ρ 和 μ 得最优参数, ρ 的值范围为 $\{10^{-5}, 10^{-4}, \dots, 1\}$, μ 的值范围为 $\{0.2, 0.4, \dots, 1.8, 2\}$ 。RSLDS 方法在不同 ρ 和 μ 下的准

确率见图 4, 同样选择最高准确率对应的参数为最终参数。

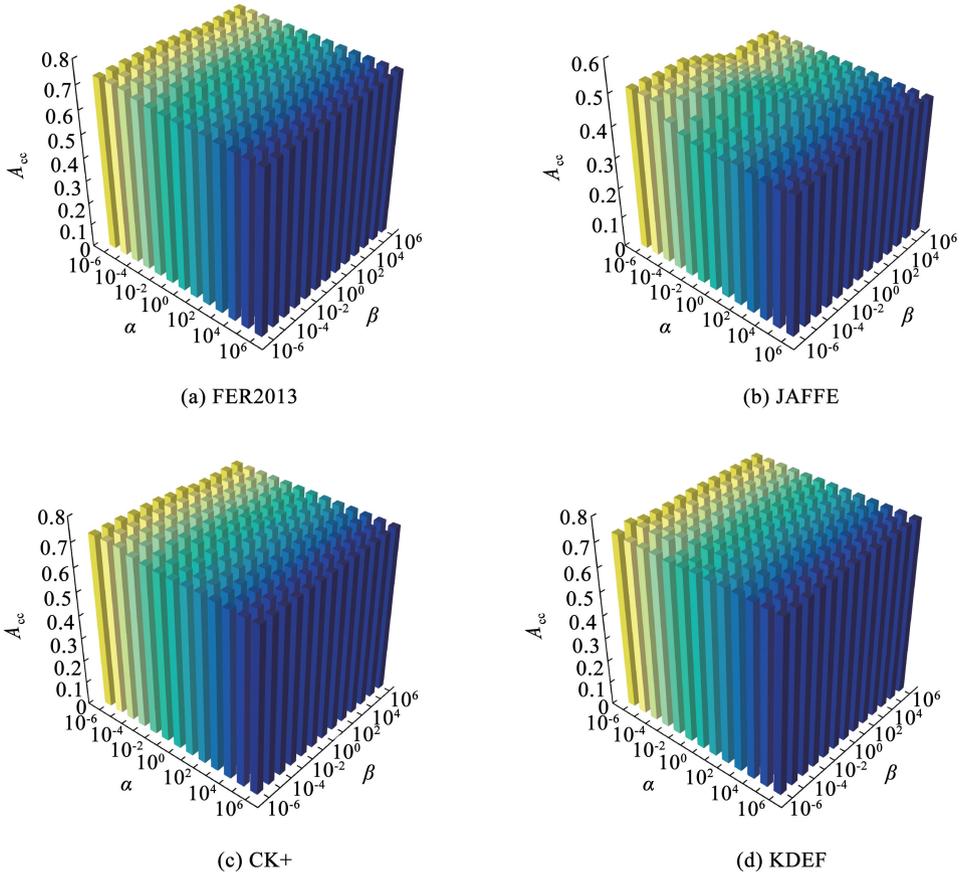


图 3 本文算法对参数 α 和 β 的敏感性分析

Fig. 3 Sensitivity analysis of parameters α and β in proposed algorithm

参数确定之后,在上述最优参数下获得本文后续的所有实验结果。集成剪枝前后,基分类器子集的准确率分布情况见图 5。图形越宽表示在当前区间分别的基分类器数量越多,图形越长表示基分类器分布范围越广,结果表明,剪枝后,图形都变得短而小,这表明大多数冗余和弱分类器在集成修剪后被成功去除。

2.4 消融实验结果与分析

为了分析不同改进点对集成剪枝性能的影响,以式(3)为基础方法,在此基础上分别考虑各个改进点的作用,式(3)考虑鲁棒性的表达式为

$$\min_w \|W^T P - Y\|_{2,1} + \alpha \|W\|_{2,1} \quad (25)$$

式(3)考虑内积正则化项的表达式为

$$\min_w \|W^T P - Y\|_F^2 + \alpha (\|WW^T\|_1 - \|W\|_F^2) \quad (26)$$

式(3)考虑依赖分数的表达式为

$$\min_w \|W^T P - Y\|_F^2 + \alpha \|W\|_{2,1} + \beta \text{Tr}(W^T D Q W) \quad (27)$$

式(3)考虑鲁棒性和式(3)的内积正则化项的

表达式为

$$\min_w \|W^T P - Y\|_{2,1} + \alpha (\|WW^T\|_1 - \|W\|_F^2) \quad (28)$$

式(3)考虑鲁棒性和式(3)考虑依赖分数的表达式为

$$\min_w \|W^T P - Y\|_{2,1} + \alpha \|W\|_{2,1} + \beta \text{Tr}(W^T D Q W) \quad (29)$$

式(3)的内积正则化项和式(3)考虑依赖分数的表达式为

$$\min_w \|W^T P - Y\|_F^2 + \alpha (\|WW^T\|_1 - \|W\|_F^2) + \beta \text{Tr}(W^T D Q W) \quad (30)$$

考虑了各个改进点的效果见表 3, 打勾表示考虑了对应的改进点,鲁棒性的选择降低了错误预测值对基分类器选择的影响。因此,考虑鲁棒性的改进点取得的效果优于考虑内积正则化项和依赖分数。在鲁棒性选择的基础上,将内积正则化项和依赖分数也进行考虑,即保证了鲁棒性的进行基分类器选择,又迫使基分类器子集数量小且低冗余,从而保证选择优质且低冗余的基分类器。

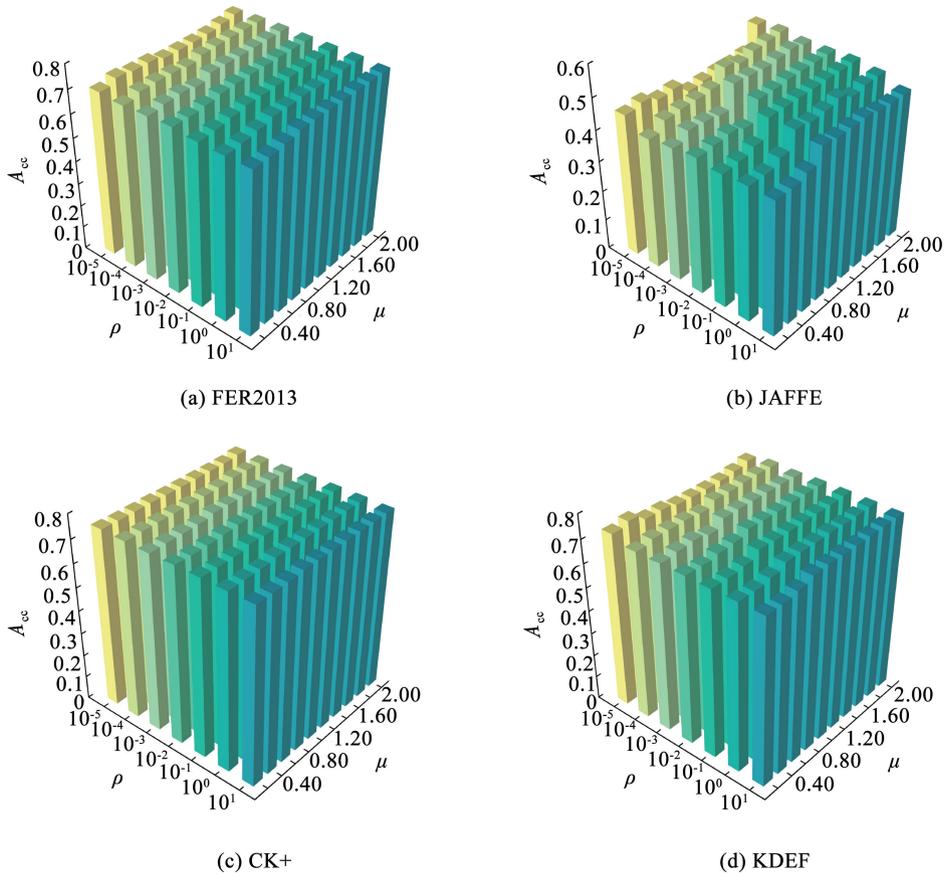


图 4 本文算法对参数 ρ 和 μ 的敏感性分析

Fig. 4 Sensitivity analysis of paramaters ρ and μ in proposed algorithm

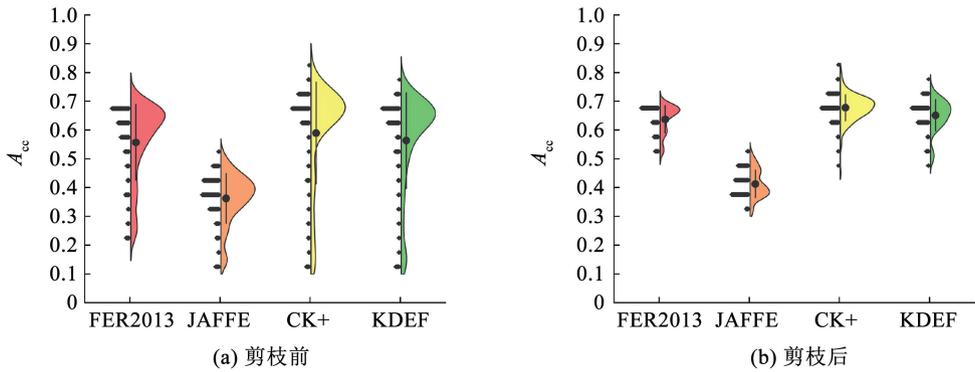


图 5 集成剪枝前后的基分类器分布

Fig. 5 Base classifier distribution before and after ensemble pruning

表 3 考虑不同模块的准确率对比

Tab. 3 Comparison of accuracy considering different modules

鲁棒性	内积正则化项	依赖分数	FER2013	JAFFE	CK +	KDEF
√	√	√	0.729 5	0.493 0	0.767 1	0.753 1
			0.734 0	0.511 1	0.775 0	0.763 3
√	√	√	0.729 4	0.507 7	0.762 5	0.753 9
			0.731 2	0.512 3	0.769 8	0.759 0
√	√	√	0.733 5	0.515 6	0.776 9	0.765 0
√	√	√	0.735 0	0.519 8	0.775 1	0.762 4
√	√	√	0.731 7	0.508 6	0.766 5	0.757 1
			0.736 4	0.526 4	0.775 4	0.765 2

2.5 RSLDS 与其他集成剪枝方法的对比

将 RSLDS 在 4 个数据集上与常见的集成剪枝方法进行了比较,结果见表 4。表中的 Baseline 表示将所有分类器的预测结果直接用多数投票方法融合得出的结果。获得最佳性能的用粗体表示。从表中可以看出,RSLDS 在 4 个数据集上均取得最高的分类

准确率,同时,RSLDS 选择的基分类器子集数量也更少。这表明,考虑分类器依赖、内积正则化项和鲁棒性的集成剪枝方法不仅仅进一步降低了基分类器池所需的存储容量,而且还提升了模型的分类性能,取得了更好的效果。

表 4 将 RSLDS 与 8 种集成剪枝方法的准确率比较

Tab. 4 Comparison of accuracy between RSLDS and eight ensemble pruning methods

数据集	FER2013		JAFPE		CK +		KDEF	
	准确率	数量	准确率	数量	准确率	数量	准确率	数量
Baseline	0.703 5	231	0.422 5	231	0.757 8	231	0.716 3	231
DREP ^[18]	0.732 0	29	0.469 5	116	0.764 5	114	0.755 1	105
ComEP ^[19]	0.710 8	209	0.507 0	3	0.759 2	224	0.720 4	209
Kappa ^[7]	0.713 0	196	0.469 5	192	0.758 6	230	0.716 3	231
QS ^[3]	0.709 4	206	0.446 0	174	0.758 6	230	0.716 3	231
RGSS&B-EP ^[20]	0.727 5	118	0.507 0	137	0.769 1	143	0.753 1	103
OO ^[21]	0.733 4	49	0.511 7	134	0.770 4	142	0.759 2	53
SDAcc ^[8]	0.732 5	18	0.511 7	129	0.769 7	142	0.757 1	57
CSLSEP ^[4]	0.735 9	29	0.511 7	133	0.771 7	143	0.761 2	50
RSLDS	0.736 4	28	0.526 4	79	0.775 4	112	0.765 2	42

2.6 统计分析

本文使用临界距离 (critical distance, CD) 来分析 RSLDS 和其他方法之间的相对性能,CD 的表达式为

$$q_{\alpha} \sqrt{\frac{K(K+1)}{6N}} \quad (31)$$

集成剪枝算法的 CD 见图 6,横轴代表各种方法在 4 个数据集上的平均排名,粗红线表明其他对比方法与提出的 RSLDS 方法在性能上没有明显差异,否则认为其他对比方法与 RSLDS 有明显差异。观察图 6 可以发现,本文提出的 RSLDS 在所有集成剪枝算法中,取得了最高的平均排序,这说明该方法在集成剪枝领域具有优越性。

一种具有依赖分数的鲁棒性稀疏低冗余集成剪枝方法。首先,该方法使用 $l_{2,1}$ 范数作为损失函数提高分类器子集的鲁棒性;其次,使用互信息来衡量不同基分类器之间的依赖关系并使用熵来评估他们的优先级;然后,考虑了内积正则化项确保选择的基分类器子集更稀疏;最后,将基分类器之间的优先级依赖、 $l_{2,1}$ 范数和内积正则化项整合到一个基于回归的目标方程中,同时考虑多样性和准确率之间的平衡,实现冗余分类器和弱分类器的修剪。为了验证提出的 RSLDS 方法的有效性和鲁棒性,在 4 个公共人脸情感数据集上进行了实验,结果表明,提出的方法不仅有效的提高人脸情感识别的准确率,而且进一步降低了集成模型的运行时间和存储空间。

参考文献

[1] GUO Yifei, HUANG Jian, XIONG Mingfu, et al. Facial expressions recognition with multi-region divided attention networks for smart education cloud applications [J]. Neurocomputing, 2022, 493: 119. DOI: 10.1016/j.neucom.2022.04.052

[2] LI Yunkai, MENG Qinghao, WANG Yaxin, et al. MMFN: emotion recognition by fusing touch gesture and facial expression information [J]. Expert Systems with Applications, 2023, 228: 120469. DOI: 10.1016/j.eswa.2023.120469

[3] LI Danyang, WEN Guihua, LI Xu, et al. Graph-based dynamic ensemble pruning for facial expression recognition [J]. Applied Intelligence, 2019, 49(9): 3188. DOI: 10.1007/s10489-019-01435-2

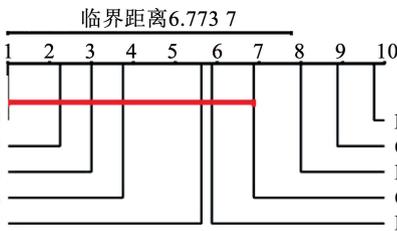


图 6 集成剪枝算法的临界距离

Fig. 6 Critical distance of ensemble pruning algorithm

3 结论

为了选择识别效果更好和冗余度更低的基分类器子集,提高集成剪枝的泛化能力和准确率,提出了

- [4] HUANG Shisong, LI Danyang, ZHANG Zhuhong, et al. CSLSEP: an ensemble pruning algorithm based on clustering soft label and sorting for facial expression recognition [J]. *Multimedia Systems*, 2023, 29: 1463. DOI: 10.1007/s00530-023-01062-5
- [5] 金志刚, 韩玥, 朱琦. 一种结合深度学习和集成学习的情感分析模型[J]. *哈尔滨工业大学学报*, 2018, 50(11): 32
JIN Zhigang, HAN Yue, ZHU Qi. A sentiment analysis model with the combination of deep learning and ensemble learning[J]. *Journal of Harbin Institute of Technology*, 2018, 50(11): 32. DOI: 10.11918/j.issn.0367-6234.201709078
- [6] ZHANG Chunxia, ZHANG Jianshe, YIN Qingyan. A ranking-based strategy to prune variable selection ensembles[J]. *Knowledge-Based Systems*, 2017, 125: 13. DOI: 10.1016/j.knsys.2017.03.031
- [7] WANG Zhenlei, ZHAO Suyun, LI Zheng, et al. Ensemble selection with joint spectral clustering and structural sparsity [J]. *Pattern Recognition*, 2021, 119(11): 108061. DOI: 10.1016/j.patcog.2021.108061
- [8] DAI Qun, YE Rui, LIU Zhuan. Considering diversity and accuracy simultaneously for ensemble pruning[J]. *Applied Soft Computing*, 2017, 58(9): 75. DOI: 10.1016/j.asoc.2017.04.058
- [9] MING D, DING C. Robust flexible feature selection via exclusive L21 regularization[C]//*Proceedings of the 28th International Joint Conference on Artificial Intelligence*. Macao: AAAI Press, 2019: 3158. DOI: 10.24963/ijcai.2019/438
- [10] DU Shiqiang, MA Yide, LI Shouliang, et al. Robust unsupervised feature selection via matrix factorization [J]. *Neurocomputing*, 2017, 241: 115. DOI: 10.1016/j.neucom.2017.02.034
- [11] LIM H, KIM D W. Pairwise dependence-based unsupervised feature selection [J]. *Pattern Recognition*, 2021, 111: 107663. DOI: 10.1016/j.patcog.2020.107663
- [12] LI Yonghao, HU Liang, GAO Wanfu. Robust sparse and low-redundancy multi-label feature selection with dynamic local and global structure preservation[J]. *Pattern Recognition*, 2023, 134: 109120. DOI: 10.1016/j.patcog.2022.109120
- [13] SUN Zhengya, HAN Jiuqi, HAO Hongwei. Selecting feature subset with sparsity and low redundancy for unsupervised learning [J]. *Knowledge-Based Systems*, 2015, 86(9): 210. DOI: 10.1016/j.knsys.2015.06.008
- [14] AFONSO M V, BIOUCAS-DIAS J M, FIGUEIREDO M A T. An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems[J]. *IEEE Transactions on Image Processing*, 2011, 20(3): 681. DOI: 10.1109/TIP.2010.2076294
- [15] GOODFELLOW L J, DUMITRU E, CARRIER P L, et al. Challenges in representation learning: a report on three machine learning contests[J]. *Neural Networks*, 2015, 64: 59. DOI: 10.1016/j.neunet.2014.09.005
- [16] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression [C]//*2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. San Francisco, IEEE, 2010: 94. DOI: 10.1109/CVPRW.2010.5543262
- [17] GOELEVELN E, DE R R, LEYMAN L, et al. The Karolinska directed emotional faces: a validation study [J]. *Cognition and Emotion*, 2008, 22(6): 1094. DOI: 10.1080/02699930701626582
- [18] LI Nan, YU Yang, ZHOU Zhihua. Diversity regularized ensemble pruning [C]//*Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 330. DOI: 10.1007/978-3-642-33460-3_27
- [19] LI Danyang, WEN Guihua, HOU Zhi, et al. RTCRelief-F: an effective clustering and ordering-based ensemble pruning algorithm for facial expression recognition [J]. *Knowledge and Information Systems*, 2019, 59(1): 219. DOI: 10.1007/s10115-018-1176-z
- [20] DAI Q. A novel ensemble pruning algorithm based on randomized greedy selective strategy and ballot [J]. *Neurocomputing*, 2013, 122(122): 258. DOI: 10.1016/j.neucom.2013.06.026
- [21] MARTÍNEZ-MUÑOZ G, SUÁREZ A. Pruning in ordered bagging ensembles [C]//*Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh: Association for Computing Machinery, 2006: 609. DOI: 10.1145/1143844.1143921

(编辑 苗秀芝)