

特征增量极限学习机

赵中堂^{1,2} 郑小东¹

(郑州航空工业管理学院智能工程学院 郑州 450046)¹

(北京理工大学智能机器人与系统高精尖创新中心 北京 100081)²

摘 要 在机器学习的不同应用领域,出现了很多优秀的极限学习机分类模型。研究者往往愿意公开这些模型的结构以及参数,但不愿公开原始训练数据。针对如何仅利用现有的模型和少量具有新特征的样本得到一个更高效的识别模型的问题,提出一种特征增量极限学习机算法。该算法能从具有新特征的样本中学习知识,提高现有模型的识别精度。在真实世界图像和三轴加速度传感器数据集上的测试结果表明,该算法能有效地工作,在不需要以往训练样本参与的情况下,能一定程度上提高已有模型的识别精度,得到新的识别模型。

关键词 增量学习,普适计算,迁移学习,机器学习

中图法分类号 TP393 **文献标识码** A

Feature Incremental Extreme Learning Machine

ZHAO Zhong-tang^{1,2} ZHENG Xiao-dong¹

(School of Intelligent Engineering, Zhengzhou University of Aeronautics, Zhengzhou 450046, China)¹

(Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing 100081, China)²

Abstract In different application fields of machine learning, many excellent classification models of extreme learning machine were produced. Researchers are often willing to share the structure and parameters of these models, but are reluctant to share the original training data. To solve the problem of how to use a small number of samples with new features and the existing classifier to generate a more efficient classifier, this paper proposed a feature incremental extreme learning machine, which can learn knowledge from samples with new features and improve the recognition accuracy of existing models. Test results on real world datasets show that the proposed algorithm can work effectively and improve the recognition accuracy of existing models, without the participation of previous training samples.

Keywords Incremental learning, Pervasive computing, Transfer learning, Machine learning

1 背景

在计算机视觉、自然语言处理、生物信息学、基于多模传感数据的行为识别等领域,目前已有许多好的极限学习机(Extreme Learning Machine, ELM)模型。这些模型往往使用大量训练数据,为得到构成 ELM 模型的超级参数,即使在最先进的计算设备上,也需要数周甚至更长的时间来完成训练。有些研究人员愿意将模型结构及参数分享给其他人员,但是却不愿意将宝贵的数据公开出来。

然而,这些模型本身由于条件所限,在某些任务上的识别性能不高。比如,基于图像的模型难以识别复杂背景图像中的物体,基于加速度传感器数据的行为识别模型对摔倒行为的识别精度较低。发生这种情况的原因在于,仅利用原始数据及现有的特征提取技术难以找到区分度高的特征。

随着科技的发展,当新的传感设备数据可以辅助原有数据,或者新的特征被发现时,这种问题就会得到解决。

基于这样的认识,我们提出了本文的研究,研究的前提

是:1)已知某个 ELM 模型的结构和参数,其能识别的类别集为 A ,但对其中某些类别(构成集合 B)的识别性能不高;2)训练该模型的原始数据未知;3)已有少量标定好的样本特征(包含原有 ELM 模型中的特征和新发现的特征),但不足以训练得到一个好的模型。研究任务是:利用 1)和 3)得到一个效率更高的 ELM 模型,使得对 B 集合中的类别的识别能力增强,且并不降低对集合 $C=A-B$ 中类别的识别能力。我们将研究得到的算法称为特征增量极限学习机(Feature Incremental ELM)。

2 相关工作

过去数年间,增量学习^[1]的研究已经取得了一些进展。

在神经网络方面,增量径向基函数网络^[2](Radial Basis Function networks, RBF networks)是一种新型的 RBF 网络,它能够根据错分类别的累计错误信息动态地插入新的径向基函数中心,并根据 Neural-Gas 算法选择新中心的半径。AR-TMAP 是一种基于自适应共振理论的增量学习网络,模糊

本文受国家自然科学基金项目(U1504609),北京理工大学智能机器人与系统高精尖创新中心开放基金项目(2018IRS09),河南省 2017 科技发展规划项目(172102210525),河南省高等学校青年骨干教师培养计划项目(2017GGJS111),河南省高等教育教学改革研究与实践一般项目(2017SJGLX400)资助。

赵中堂(1978—),男,博士,副教授,CCF 会员,主要研究方向为机器学习,E-mail:112864041@qq.com(通信作者)。

ARTMAP^[3](FAM)是它的一个变种。FAM 吸纳了模糊集理论,用以决定 ARTMAP 网络的动态变化。概率 FAM^[4]混合了 FAM 和概率神经网络(Probabilistic Neural Networks, PNNs)以达到在线学习和预测的目的,它通过引入 PNN 提高了 FAM 的分类精度。上述方法的主要缺点在于学习速度慢,易受噪音数据的影响;容易导致过拟合,概化能力不强;仅能对模型中已有类别更新,不能添加对新类数据的识别能力。文献[5]基于 PNN 和可调节的模糊聚类算法(Adjustable Fuzzy Clustering algorithm, AFC)实现了一种增量学习方法。该方法可以轻易地从新的训练样本中获取知识以提高模型的识别能力;还可以自由地增加对新类别的识别能力,删除对不感兴趣类别的识别能力;无需保存以往使用的训练数据,以节省存储空间。然而,观察其实验结果,对于大约 1000 个样本,采用 5 折交叉验证的方法,其增量学习的过程需要 264.3 s,对于智能终端之类的资源受限设备,其计算时间是无法忍受的。新加坡南洋理工大学的 Huang 教授等人提出了极限学习机算法^[6],ELM 是一个单隐层前馈神经网络,与传统的神经网络求解算法不同,它不需要通过反复的迭代过程调整神经元的输入权重和偏置值,而是通过求解线性系统的 Moore-Penrose 伪逆和范数最小二乘解的方法计算输出权重。因此,ELM 具有计算速度快、鲁棒性强的特点,但是其只能批量学习。在此基础上,文献[7]提出的 OSELM(Online Sequential Extreme Learning Machine)算法能够以增量的方式处理数据,只有那些新出现的单个或者成块数据需要参与学习,已经参与过学习的数据无需再次参与。然而,OSELM 只能处理模型中已有类别的样本,不能处理新出现类别的样本。

在支持向量机研究方面,张浩然分析了回归最小二乘支持向量机的性质^[8],并在此基础上根据分块矩阵计算公式和核函数矩阵本身的特点设计了支持向量机的增量式学习算法和在线学习算法。该算法能充分利用历史训练结果,减少了存储空间和计算时间。该文献对回归问题进行了充分论述,但没有讨论在分类问题上的求解以及如何对新类别样本进行增量学习。

在随机森林研究方面,王爱平提出了一种增量式极端随机森林分类器^[9](Incremental Extremely Random Forest, IERF),用于处理数据流特别是小样本数据流的在线学习问题。IERF 算法中新到达的样本将被存储到相应的叶节点,并通过 Gini 系数来确定是否对当前叶节点进行分裂扩展。然而,该方法仍需要存储所有样本。

综上所述,已有几类机器学习算法具有了对已知类别进行增量学习的功能,然而它们普遍具有时间复杂度高的缺点,不适用于穿戴式设备资源受限的环境。目前鲜有对特征增量的研究。

3 特征增量机器学习方法

3.1 ELM 的网络结构

ELM 是近年来出现的一种新的机器学习方法,它计算速度快、泛化能力强^[6,10-14],吸引了众多科研人员的注意。ELM

既可以处理回归问题,又可以处理分类问题。针对所研究问题的特点,我们在此关注的是其分类能力。

如图 1 所示,ELM 算法总体上是一个前向神经网络结构,其仅有一个隐结点层(Hidden Layer),因此具有结构简单特点。输入结点层(Input Layer)包含所有输入神经元结点,每个输入神经元结点对应样本的一个特征;隐结点层(Hidden Layer)包含所有隐层神经元结点;输出结点层(Output Layer)包含所有输出神经元结点,每个输出神经元结点对应多分类问题的一个类别。对于训练集合 $S = \{(\mathbf{x}_i, \mathbf{t}_i) | \mathbf{x}_i \in R^n, \mathbf{t}_i \in R^m, i = 1, 2, \dots, N\}$,它共含有 N 个样本,其中每个样本 $\mathbf{x}_i = \langle f_1^i, f_2^i, \dots, f_n^i \rangle$ 是一个行向量,其长度 n 对应样本特征的个数; $\mathbf{t}_i$ 也是一个行向量,其长度 m 对应类别的个数。当 $\mathbf{x}_i$ 的类标号为 j 时,向量 $\mathbf{t}_i$ 的第 j 位为 1,其他各位全部为 0。

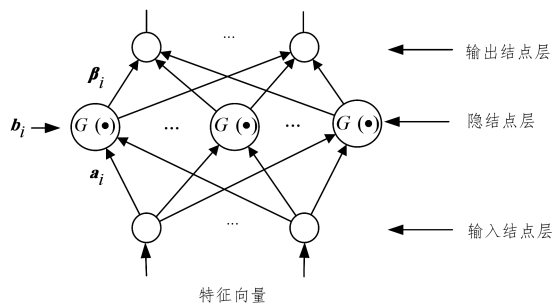


图 1 ELM 算法的网络结构

3.2 ELM 网络的训练求解

用单隐层前向神经网络对该问题建模,选定输入结点个数为 n ,隐层结点个数为 $\tilde{N}$,输出结点个数为 m ,激励函数为 $G(\mathbf{a}, \mathbf{b}, \mathbf{x})$ 。对每一个样本 $\mathbf{x}_i$,有:

$$\begin{aligned} & [G(\mathbf{a}_1, \mathbf{b}_1, \mathbf{x}_1), \dots, G(\mathbf{a}_{\tilde{N}}, \mathbf{b}_{\tilde{N}}, \mathbf{x}_1)] \begin{bmatrix} \beta_{11} & \beta_{21} & \dots & \beta_{m1} \\ \beta_{12} & \beta_{22} & \dots & \beta_{m2} \\ \vdots & \vdots & & \vdots \\ \beta_{1\tilde{N}} & \beta_{2\tilde{N}} & \dots & \beta_{m\tilde{N}} \end{bmatrix} \\ & = [o_{11}, o_{12}, \dots, o_{1m}] \end{aligned} \quad (1)$$

其中, $\mathbf{a}_i = [a_{i1}, a_{i2}, \dots, a_{in}]$ 表示第 i 个隐结点到所有输入结点之间权值组成的向量; $\mathbf{b}_i$ 表示第 i 个隐结点的阈值; $\mathbf{a}_i$ 和 $\mathbf{b}_i$

的值随机生成,其值位于区间 $[-1, 1]$ 之间。 $\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{i\tilde{N}} \end{bmatrix}$ 表示

第 i 个输出结点到每一个隐结点之间的权值组成的向量; $\mathbf{o}_i = [o_{i1}, o_{i2}, \dots, o_{im}]$ 表示 m 个输出结点的值组成的向量。

训练集合 $S = \{(\mathbf{x}_i, \mathbf{t}_i) | \mathbf{x}_i \in R^n, \mathbf{t}_i \in R^m, i = 1, 2, \dots, N\}$ 中的所有 N 个样本均可以写成式(1)的形式,进而把所有 N 个等式写成如式(2)的矩阵形式:

$$\begin{bmatrix} G(\mathbf{a}_1, \mathbf{b}_1, \mathbf{x}_1) & \dots & G(\mathbf{a}_{\tilde{N}}, \mathbf{b}_{\tilde{N}}, \mathbf{x}_1) \\ G(\mathbf{a}_1, \mathbf{b}_1, \mathbf{x}_2) & \dots & G(\mathbf{a}_{\tilde{N}}, \mathbf{b}_{\tilde{N}}, \mathbf{x}_2) \\ \vdots & & \vdots \\ G(\mathbf{a}_1, \mathbf{b}_1, \mathbf{x}_N) & \dots & G(\mathbf{a}_{\tilde{N}}, \mathbf{b}_{\tilde{N}}, \mathbf{x}_N) \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} & \dots & \beta_{m1} \\ \beta_{12} & \beta_{22} & \dots & \beta_{m2} \\ \vdots & \vdots & & \vdots \\ \beta_{1\tilde{N}} & \beta_{2\tilde{N}} & \dots & \beta_{m\tilde{N}} \end{bmatrix} = \begin{bmatrix} o_{11} & o_{12} & \dots & o_{1m} \\ o_{21} & o_{22} & \dots & o_{2m} \\ \vdots & \vdots & & \vdots \\ o_{N1} & o_{N2} & \dots & o_{Nm} \end{bmatrix} \quad (2)$$

训练集合 S 对应的单隐层前向网络的优化目标是使预测

误差趋于 0, 即 $\sum_{i=1}^N \|o_i - t_i\| = 0$, 亦即存在 β_i, a_i 和 b_i , 使得:

$$\begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_1) \\ G(a_1, b_1, x_2) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_2) \\ \vdots & & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_N) \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} & \cdots & \beta_{m1} \\ \beta_{12} & \beta_{22} & \cdots & \beta_{m2} \\ \vdots & \vdots & & \vdots \\ \beta_{1\tilde{N}} & \beta_{2\tilde{N}} & \cdots & \beta_{m\tilde{N}} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{Nm} \end{bmatrix}$$

上式可以采用矩阵的形式, 写得更加紧凑一些:

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$$

其中:

$$\mathbf{H} = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_1) \\ \vdots & & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_N) \end{bmatrix}_{N \times \tilde{N}}$$

$$\boldsymbol{\beta} = [\beta_1 \quad \cdots \quad \beta_m]_{\tilde{N} \times m} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{m,1} \\ \vdots & & \vdots \\ \beta_{1,\tilde{N}} & \cdots & \beta_{m,\tilde{N}} \end{bmatrix}_{\tilde{N} \times m}$$

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{Nm} \end{bmatrix}_{N \times m}$$

按照传统的方法, 训练一个单隐层前向网络, 我们需要找到

特定的 β_i, a_i 和 $b_i (i=1, 2, 3, \dots, \tilde{N})$, 使得:

$$\|H(\hat{a}_1, \dots, \hat{a}_{\tilde{N}}, \hat{b}_1, \dots, \hat{b}_{\tilde{N}})\boldsymbol{\beta} - \mathbf{T}\| = \min_{\beta, a_i, b_i} \|H(a_1, \dots, a_{\tilde{N}}, b_1, \dots, b_{\tilde{N}})\boldsymbol{\beta} - \mathbf{T}\|$$

Huang^[6]指出, a_i 和 b_i 的值无需调整, 只需在算法开始

时随机生成即可, 那么问题就转换为:

$$\|H(a_1, \dots, a_{\tilde{N}}, b_1, \dots, b_{\tilde{N}})\hat{\boldsymbol{\beta}} - \mathbf{T}\| = \min_{\hat{\boldsymbol{\beta}}} \|H(a_1, \dots, a_{\tilde{N}}, b_1, \dots, b_{\tilde{N}})\boldsymbol{\beta} - \mathbf{T}\|$$

则该问题的最小范数二乘解为:

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^\dagger \mathbf{T} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{T}$$

其中, $\mathbf{H}^\dagger$ 是矩阵 $\mathbf{H}$ 的 Moor-Penrose 伪逆。

3.3 特征增量学习机

特征增量学习机 (Feature Online Incremental Extreme Learning Machine, FOIELM) 的核心是如何增加 ELM 模型的输入结点数并调整输入结点与隐结点之间的权值。

令已有标注数据 $S = \{(x_s^i, t_s^i) \mid x_s^i \in R^n, t_s^i \in R^m, i=1, 2, \dots, N\}$, 一共有 N 个样本, 每个样本有 n 个特征。令增量数据为 $D = \{(x_d^i, t_d^i) \mid x_d^i \in R^{n'}, t_d^i \in R^m, i=1, 2, \dots, N'\}$, 一共有 N' 个样本, 每个样本有 n' 个特征。

若将 S 和 D 放入同一个集合同时训练, 则可以采用矩阵的形式来描述该问题, 即:

$$\begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix} \boldsymbol{\beta}_{s+d} = \begin{bmatrix} \mathbf{T}_s \\ \mathbf{T}_d \end{bmatrix}$$

按照上述 ELM 理论, 可以求解:

$$\boldsymbol{\beta}_{s+d} = \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix}^\dagger \begin{bmatrix} \mathbf{T}_s \\ \mathbf{T}_d \end{bmatrix}$$

$$= \left(\begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix}^\top \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix}^\top \begin{bmatrix} \mathbf{T}_s \\ \mathbf{T}_d \end{bmatrix}$$

进一步:

$$\boldsymbol{\beta}_{s+d} = \left(\begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix}^\top \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix} \right)^{-1} (\mathbf{H}_s^\top \mathbf{T}_s + \mathbf{H}_d^\top \mathbf{T}_d)$$

令:

$$\mathbf{K}_{s+d} = \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix}^\top \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix} = \mathbf{H}_s^\top \mathbf{H}_s + \mathbf{H}_d^\top \mathbf{H}_d = \mathbf{K}_s + \mathbf{K}_d$$

其中:

$$\mathbf{K}_s = \mathbf{H}_s^\top \mathbf{H}_s$$

$$\mathbf{K}_d = \mathbf{H}_d^\top \mathbf{H}_d$$

则:

$$\begin{aligned} \boldsymbol{\beta}_{s+d} &= (\mathbf{K}_{s+d})^{-1} (\mathbf{K}_s \mathbf{K}_s^{-1} \mathbf{H}_s^\top \mathbf{T}_s + \mathbf{H}_d^\top \mathbf{T}_d) \\ &= (\mathbf{K}_{s+d})^{-1} (\mathbf{K}_s \boldsymbol{\beta}_s + \mathbf{H}_d^\top \mathbf{T}_d) \\ &= (\mathbf{K}_{s+d})^{-1} ((\mathbf{K}_{s+d} - \mathbf{K}_d) \boldsymbol{\beta}_s + \mathbf{H}_d^\top \mathbf{T}_d) \\ &= (\mathbf{K}_{s+d})^{-1} ((\mathbf{K}_{s+d} - \mathbf{H}_d^\top \mathbf{H}_d) \boldsymbol{\beta}_s + \mathbf{H}_d^\top \mathbf{T}_d) \\ &= \boldsymbol{\beta}_s + (\mathbf{K}_{s+d})^{-1} (\mathbf{H}_d \mathbf{T}_d - \mathbf{H}_d \mathbf{H}_d \boldsymbol{\beta}_s) \\ &= \boldsymbol{\beta}_s + (\mathbf{K}_{s+d})^{-1} \mathbf{H}_d^\top (\mathbf{T}_d - \mathbf{H}_d \boldsymbol{\beta}_s) \end{aligned} \quad (3)$$

从式(3)可以看出, $\boldsymbol{\beta}_{s+d}$ 是 $\boldsymbol{\beta}_s, \mathbf{H}_s, \mathbf{H}_d$ 和 $\mathbf{T}_d$ 的函数, 而 $\mathbf{H}_s$ 是计算 $\boldsymbol{\beta}_s$ 时的一个中间结果。这意味着, 在已有新一批数据 D 到来时, 我们只需要计算 $\mathbf{H}_d$ 和 $\mathbf{T}_d$, 利用旧数据计算的模型参数 $\mathbf{H}_s$ 和 $\boldsymbol{\beta}_s$, 就可以得到新的模型。因此, 该模型是一个增量学习模型。

3.4 不同长度特征的处理方法

在 ELM 网络中, 当从输入结点向隐层结点计算矩阵 $\mathbf{H}$ 时:

$$\mathbf{H} = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_1) \\ \vdots & & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_{\tilde{N}}, b_{\tilde{N}}, x_N) \end{bmatrix}_{N \times \tilde{N}}$$

由于数据集 S 和数据集 D 中样本的特征长度不同, 其对应的输入结点个数不一样, D 中特征个数要比 S 中多, 需要为新增的特征增加输入权重 a 和对应的偏执权重 b 。

$$\begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_d \end{bmatrix} \boldsymbol{\beta}_{s+d} = \begin{bmatrix} \mathbf{T}_s \\ \mathbf{T}_d \end{bmatrix}$$

为应对特征数目增多的情况, 我们重新设置 ELM 网络结构, 为其预留多个虚拟的输入结点, 比如预留 L 个输入结点位置。同理, 输入权重向量也做同样的设置。

假设当前输入的一个特征向量为 $\mathbf{x}_i = \langle f_i^1, f_i^2, \dots, f_i^n \rangle$, 对应的 ELM 输入结点向量为:

$$\mathbf{y}_i = \langle f_i^1, f_i^2, \dots, f_i^n, 0_i^{n+1}, \dots, 0_i^L \rangle$$

即 $\mathbf{y}_i = \mathbf{x}_i * \mathbf{M}$ 。其中:

$$\mathbf{M} = \left[\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & 1 & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{n \times n} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & 0 & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{n \times (L-n)} \right]_{n \times L}$$

对于增加特征后的向量 $\mathbf{x}_i' = \langle f_i^1, f_i^2, \dots, f_i^n, f_i^{n+1}, \dots, f_i^{n'} \rangle$, 其特征数为 n' , 则调整 $\mathbf{M}$ 矩阵为:

$$\mathbf{M} = \left[\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{n' \times n'} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n' \times (L-n')} \right]_{n' \times L}$$

这样就能把原始数据和增量数据的样本长度统一, 其均为 L 。

同理, 对于输入权重 $\mathbf{a}_i = [a_{i1}, a_{i2}, \dots, a_{iL}]$, 初始时是一个

随机向量,大小为 L 。根据输入特征向量的大小,也对其进行类似处理,调整 $\mathbf{M}$ 中对角矩阵和零矩阵的大小。

4 实验

4.1 数据准备

我们在两个公共的图像数据集和一个自己采集的行为数据集上做验证实验。

本实验所采用的两个图像数据集均来自 UCI 数据库¹⁾,一个称为 image segments,另一个称为 satellite images。

image segments 问题所采用数据库中的图像是从 7 幅室外图像中抽取出来的,共有 2310 个子图像,每个图像的大小为 3 像素×3 像素,可以从每幅子图像中提取 19 个特征,用以将其分类到 7 幅原始图像中的一个。

satellite images 问题中的图像数据库全部来自地球卫星多光谱扫描仪(landsat multispectral scanner)扫描到的一个场景(scene),该扫描仪具有 4 个不同的谱段,可以在一帧图像中同时扫描到 4 幅图像。satellite images 问题选取 1 帧图片,从中选取感兴趣的子区域(82 像素×100 像素),然后从该子区域中抽取更小的图片碎片(3 像素×3 像素)。该问题的目标是使用这些图片碎片的 36 个特征,并将其中心像素分类到 6 个分类(red soil,cotton crop,gray soil,damp gray soil,soil with vegetation stubble,and very damp gray soil)中的一个。

image segments 数据库和 satellite images 数据库的相关信息如表 1 所列。

表 1 UCI 数据集的相关信息

数据集名	特征个数	类别个数	样本个数
image segment	19	7	2310
satellite images	36	6	6435

对于行为识别,我们用加速度传感器数据区分静止、散步、跑步、上楼梯和下楼梯这 5 种行为,并采用滑动窗口的方法来提取特征。从加速度传感器的信号波形观察,在加速度传感器各个轴上的信号波形呈周期变化。对于走路动作来说,大概每秒钟对应一个动作周期,其他动作的周期会更短。由于手机的采样频率为 32 Hz,即每秒钟采集 32 个数据,我们设窗口的大小为 64 个数据,这样可以包含两个周期的信号。滑动的步长设为窗口的一半,即为 32 个数据。这种数据处理方法的优点在文献[15-17]中有所述述。

对每个窗口中的数据,分别取 x,y,z 轴数据的均值、方差、离散傅里叶变换后的能量,再提取各个轴之间的相关系数。这样,每一个样本可以得到 12 个特征。

均值能够反映某类动作的剧烈程度^[18-20],例如某人的散步和慢跑在垂直方向的加速度数据波形图明显不同,它们的波形中心相距甚远。方差能够反映数据偏离中心的幅度,例如散步和下楼梯的加速度大小在均值上相似,而下楼梯的加速度值的变化范围要明显大于散步。离散傅里叶变换将数据从时域转换到频域,因此能量特征能够区分运动周期有区别的动作,如散步和慢跑。相关系数能将多个方向都有运动的动作和仅在某一个方向有运动的动作区分开来,如散步和上楼梯^[15-16]。处理后的行为样本如表 2 所列。

表 2 日常行为的样本信息

行为	标签	样本数量
静止	1	4520
下楼梯	2	4293
散步	3	4327
跑步	4	4245
上楼梯	5	4369

4.2 算法性能评估

本节将基于上述数据集验证本文使用的 CIELM 算法,共验证了 3 方面的内容:模型选择,识别异常图像的实验,分类器性能 VS 新样本数量的实验。实验的硬件平台是一台普通的 PC 机,CPU 主频为 2.6 GHz,软件平台是 MATLAB2009a。ELM 源代码下载自算法提出者 Huang 教授的主页²⁾。

4.2.1 模型选择

模型选择是指选择最优的网络结构和选择机器学习算法中最优的参数,它一般是模型相关的。对于 CIELM 算法对应的 ELM 网络结构而言,只有隐层结点的个数是需要优化选择的。

为了选择最优的隐层结点个数,本文采取了训练和验证的方法。整个数据集既被作为训练数据,又被作为测试数据。隐层结点的个数为 1~300 之间遍历,相应的实验结果如图 2 所示。

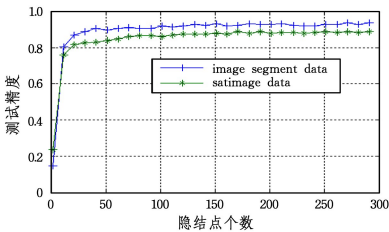


图 2 不同隐结点个数时 ELM 算法的性能

如图 2 所示,当隐层结点的个数多于 100 时,模型就能取得最好的分类性能且比较稳定。因此,在后续实验中,我们选择 100 个隐层结点作为最优的模型结构。

4.2.2 性能测试

数据集 image segment 中共有 2310 个样本,19 个特征,按照 4:1:1 的比例确定训练数据、增量数据和测试数据。其中,训练数据的特征数随机选择 18 个,增量数据的特征数为 19 个。训练两个模型,并分别进行测试,一个是用训练数据训练得到模型 ModelA,用测试数据中对应的 18 个特征进行测试;另一个是在 ModelA 的基础上用增量数据按照本文特征增量方法得到 ModelB,用测试数据中的 19 个特征进行测试。

数据集 satellite image 中共有 6435 个样本,36 个特征,按照 8:1:1 的比例确定训练数据、增量数据和测试数据。其中,训练数据的特征数随机选择 34 个,增量数据的特征数为 36 个。模型训练和测试的方法与 image segment 数据集相同。

行为识别数据集共有 21754 个样本,12 个特征,按照

1) <http://archive.ics.uci.edu/ml/>.
2) www.ntu.edu.sg/home/egbhuang

8:1:1 的比例确定训练数据、增量数据和测试数据。其中,训练数据的特征数随机选择 8 个,增量数据的特征数为 12 个。模型训练和测试的方法与 image segment 数据集相同。

实验结果如表 3 所列。

表 3 算法性能测试结果

(单位:%)

数据集	ModelA 测试精度	ModelB 测试精度
Image Segment	91.8	93.5
Satellite Images	78.3	79.3
行为识别数据	92.3	95.7

观察表 3 的实验结果可知,在 3 种数据集上,运用特征增量学习算法均可以从模型 ModelA 得到一个效果更好的新模型 ModelB。其原因在于,训练 ModelA 时的特征数少于训练 ModelB 的特征数,使得 ModelB 中蕴含的区分知识更多、更丰富。但这不能一概而论“特征多就好”,对于这 3 个数据集,其特征都是对分类有贡献的,因此是越多越好,如果特征对于分类没有贡献,会适得其反。

结束语 为了充分利用已有 ELM 模型中蕴含的知识,同时也为了降低数据标定的成本,本文提出了一种特征增量极限学习机模型,其能从具有新特征的样本中学习知识以增强模型的性能。在公开的图像数据集和我们自己采集的运动行为数据集上进行测试,测试结果表明该方法能够在一定程度上提高模型的识别精度。

参 考 文 献

[1] LOSING V,HAMMER B,WERSING H. Incremental on-line learning:A review and comparison of state of the art algorithms[J]. Neurocomputing,2018(275):1261-1274.

[2] FRITZKE B. Fast learning with incremental RBF networks[J]. Neural Process. Lett.,1994,1(1):2-5.

[3] CARPENTER G A,GROSSBERG S,ROSEN D. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system[J]. Neural Netw.,1991,4:759-771.

[4] LIM C P,HARRISON R F. Probabilistic fuzzy artmap: An autonomous neural network architecture for Bayesian probability estimation[C]// Proc. of Int. Conf. Artif. Neural Netw., 1995:148-153.

[5] WANG Z L,JIANG M,HU Y H,et al. An Incremental Learning Method Based on Probabilistic Neural Networks and Adjustable Fuzzy Clustering for Human Activity Recognition by Using Wearable Sensors[J]. IEEE Transactions on Information Technology in Biomedicine,2012,16(4):691-699.

[6] HUANG G B,WANG D H,LAN Y. Extreme learning machines:a survey[J]. International Journal of Machine Learning and Cybernetics,2011,2(2):107-122.

[7] LIANG N Y,HUANG G B,SARATCHANDRAN P,et al. A fast and accurate online sequential learning algorithm for feed-forward networks[J]. IEEE Transactions on Neural Networks, 2006,17:1411-1423.

[8] 张浩然,汪晓东. 回归最小二乘支持向量机的增量和在线式学习算法[J]. 计算机学报,2006,29(3):400-406.

[9] 王爱平,万国伟,程志全,等. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报,2011(9):2059-2074.

[10] HUANG G B. What are Extreme Learning Machines? Filling the Gap between Frank Rosenblatt’s Dream and John von Neumann’s Puzzle”[J]. Cognitive Computation,2015,7:263-278.

[11] HUANG G B, BAI Z, KASUN L L C, et al. Local Receptive Fields Based Extreme Learning Machine [J]. IEEE Computational Intelligence Magazine,2015,10(2):18-29.

[12] FENG G,HUANG G B,LIN Q,et al. Error minimized extreme learning machine with growth of hidden nodes and incremental learning[J]. IEEE Transactions on Neural Networks, 2009, 20(8):1352-1357.

[13] CHEN Y Q,ZHAO Z T,WANG S Q,et al. Extreme learning machine based device displacement free activity recognition model [J]. Soft Computing,2012,16(9):1617-1625.

[14] HUANG G B,CHEN L. Enhanced random search based incremental extreme learning machine [J]. Neurocomputing, 2008, 71(16):3460-3468.

[15] GEPPERTEH A, HAMMER B. Incremental learning algorithms and applications[C]//European Sympoisum on Artificial Neural Networks (ESANN). 2016.

[16] 徐经纬. 基于无线传感器的人体行为感知研究[D]. 南京:南京大学,2012.

[17] ZHAO Z T,CHEN Y Q,LIU J F,et al. Cross-people Mobile-phone based Activity Recognition[C]//IJCAI2011. 2011:2545-2550.

[18] FIGO D,DINIZ P C,FERREIRA D R,et al. Preprocessing techniques for context recognition from accelerometer data[J]. Personal and Ubiquitous Computing,2010,14(7):645-662.

[19] WANG X L. High accuracy distributed target detection and classification in sensor networks based on mobile agent framework[D]. University of Tennessee,2004.

[20] HAN J W, MICHELINE K, PEI J. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann,2000.