DOI:10.19651/j.cnki.emt.2106497

基于增强稀疏自编码器与 Softmax 回归的医学诊断*

孟祥莲 蒋 巍 李晓芳
(常州工学院 计算机信息工程学院 常州 213032)

摘 要：为了提升医学诊断的预测精度,设计了增强稀疏自编码器和 Softmax 回归的特征学习和分类阶段组合方法。在稀疏自编码器(SAE)网络的特征学习中,通过惩罚网络的权重实现稀疏性,结合反向传播学习方法将变化向后传递并迭代优化成本函数。在 Softmax 回归分类阶段中,利用带动量的小批量梯度下降法来优化 Softmax 分类器的交叉熵,结合小批数据计算模型误差更新模型参数并实现收敛性。将所提出方法用于心脏病、宫颈癌和慢性肾病(CKD)数据集实验,其预测精度分别为 91%、97%和 98%,并且表现出较高的特征学习和鲁棒的分类性能。

关键词：稀疏自动编码器;无监督学习;Softmax 回归;医学诊断
中图分类号：TP391.41 **文献标识码**：A **国家标准学科分类代码**：510.4099

Medical diagnosis based on enhanced sparse autocoder and Softmax regression

Meng Xianglian Jiang Wei Li Xiaofang
(School of Computer Science and Information Engineering, Changzhou Institute of Technology, Changzhou 213032, China)

Abstract: In order to improve the prediction accuracy of medical diagnosis, a feature learning and classification stage combination method based on enhanced sparse self encoder and Softmax regression is designed. In the feature learning of the sparse self encoder (SAE) network, the sparsity is realized by punishing the weight of the network, and the change is transmitted backward and the cost function is optimized iteratively. In the stage of Softmax regression classification, the cross entropy of Softmax classifier is optimized by small batch gradient descent method, and the model error is calculated with small batch data, and the model parameters are updated to achieve convergence. The prediction accuracy of the proposed method is 91%, 97% and 98% respectively for heart disease, cervical cancer and chronic kidney disease (CKD) data sets, and shows high feature learning and robust classification performance.

Keywords: sparse automatic encoder;unsupervised learning;Softmax regression;medical diagnosis

0 引 言

医学诊断是由临床医生通过分析病人的病历、实验检查和身体检查等推断影响个人疾病的过程。准确地诊断可以及时的发现个人的潜在病理,而误诊可能直接危及生命。与临床医生短缺和昂贵的手动诊断相比,基于机器学习(ML)的诊断可以显着改善医疗保健系统,并减少由临床医生压力、疲劳和经验不足等导致的误诊[1]。因此,ML 正逐步应用于医学诊断并开发出许多诊断模型[2]。然而,医学数据的不平衡性和实验数据的高成本等因素阻碍了 ML 在医学领域的发展。无监督特征学习方法因其不完全依赖于实验数据而受到广泛关注,并且适合于数据不平衡时的训练模型。实现特征学习的方法主要分为两种:监督学习和无监督学习。其中,监督学习包括字典学习[3]和多层感知器(MLP)[4]等;无监督学习包括独立成分分析、矩阵分解、聚类和自动编码器。通常,利用 L1 正则化[5]或 Kullback-Leibler (KL)散度[6]这两种方法来实现稀疏自编码器(SAE)网络的稀疏性惩罚。SAE 并没有对网络的权重进行正则化,而是将正则化强加在激活上。因此,使用这种结构可以获得次优性能,其中稀疏性使得网络难以实现接近零的成本函数[7]。

本文将改进的稀疏自动编码器(SAE)和 Softmax 分类器结合起来应用于医学诊断。SAE 对权重进行正则化,而不是像常规 SAE 那样进行激活,并使用 Softmax 分类器来执行分

收稿日期:2021-04-24
* 基金项目:国家自然科学基金(61901063)、教育部人文社会科学研究青年基金(19YJCZH120)、江苏高校“青蓝工程”基金(2020)、常州市科技计划基金(CE20205042)项目资助

类任务。为了证明该方法的有效性,使用的慢性肾病(CKD)数据集、宫颈癌风险因素数据集和 Framingham 心脏研究数据集对所提出方法进行验证。同时,使用不同的性能评估指标来评估所提出的方法的性能并与其他方法进行比较。

1　增强稀疏自编码器

1.1　自动编码器(AE)

自动编码器是用于无监督特征学习的神经网络,由输入层、隐藏层和输出层组成[8]。3 层自动编码器(AE)的基本结构,如图 1 所示。

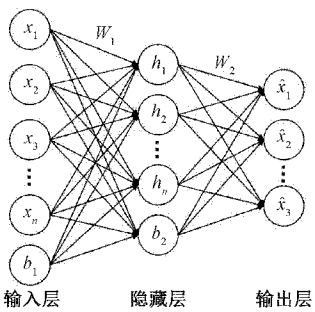

图 1　自动编码器的结构

当给定输入数据时,AE 有助于自动发现导致最佳分类的特征。自动编码器可分为变分自动编码器和正则化自动编码器。其中,正则化自编码器主要用于解决后续分类需要最优特征学习的问题。正则化自动编码器的示例包括降噪、压缩和稀疏自动编码器。本文旨在实现稀疏自动编码器(SAE)来更有效地从原始数据中学习表示,从而简化分类过程,并最终提高分类器的预测性能。

1.2　稀疏自动编码器(SAE)

SAE 是一种无监督学习方法,用于从未标记的数据中自动学习特征[9]。在这种类型的自动编码器中,训练准则涉及稀疏性惩罚。通常,SAE 的成本函数是通过惩罚隐藏层中的激活来构造的。对于任何特定的样本,网络通过激活少量节点来学习编码。通过在网络上引入稀疏约束,例如限制隐藏单元的数量,该算法可以从数据中学习更好的关系[10]。自动编码器由两个函数组成:编码器和解码器。编码器映射 d 维输入数据以获得隐藏的表示。相反,解码器将隐藏的表示映射回尽可能接近编码器输入的 d 维向量[11]。假设 m 为输入特征,n 为隐藏层神经元,则编码和解码过程可以分别表示:

$$\boldsymbol{a}^1=\begin{bmatrix}a_1^1\\ \vdots\\ a_n^1\end{bmatrix}=\begin{bmatrix}w_{1,1}^1 & w_{1,2}^1 & \cdots & w_{1,m}^1\\ \vdots & \vdots & \ddots & \vdots\\ w_{n,1}^1 & w_{n,2}^1 & \cdots & w_{n,m}^1\end{bmatrix}\begin{bmatrix}x_1\\ \vdots\\ x_m\end{bmatrix}+\begin{bmatrix}b_1^1\\ \vdots\\ b_n^1\end{bmatrix} \tag{1}$$

$$\boldsymbol{a}^2=\begin{bmatrix}a_1^2\\ \vdots\\ a_n^2\end{bmatrix}=\begin{bmatrix}w_{1,1}^2 & w_{1,2}^2 & \cdots & w_{1,m}^2\\ \vdots & \vdots & \ddots & \vdots\\ w_{n,1}^2 & w_{n,2}^2 & \cdots & w_{n,m}^2\end{bmatrix}\begin{bmatrix}a_1^1\\ \vdots\\ a_n^1\end{bmatrix}+\begin{bmatrix}b_1^2\\ \vdots\\ b_m^2\end{bmatrix} \tag{2}$$

其中,$\boldsymbol{w}^1\in R^{n,m}$ 和 $\boldsymbol{w}^2\in R^{n,m}$ 分别为隐藏层和输出层的权重矩阵;$\boldsymbol{b}^1\in R^{n,1}$ 和 $\boldsymbol{b}^2\in R^{m,1}$ 分别为隐藏层和输出层的偏差矩阵;向量 $\boldsymbol{a}^1\in R^{n,1}$ 为输出层的输入;向量 $\boldsymbol{a}^2\in R^{m,1}$ 为稀疏自编码器的输出;向量 $\boldsymbol{x}\in R^{m,1}$ 为输入层的原始输入。

1.3　增强稀疏自编码器

本文采用均方误差函数 E_{MSE} 作为输入 $\boldsymbol{x}$ 和重构输入 $\boldsymbol{a}^2$ 之间的重构误差函数。此外,在误差函数中引入正则化函数 Ω_s 的稀疏性,并结合惩罚权重 $\boldsymbol{w}^1\in R^{n,m}$ 和 $\boldsymbol{w}^2\in R^{n,m}$ 来实现稀疏性。因此,稀疏自动编码器的成本函数 E_{SAE} 可以表示为:

$$E_{SAE}=E_{MSE}+\Omega_s \tag{3}$$

均方误差函数和正则化函数可以分别表示为:

$$E_{MSE}=\frac{1}{m}\sum_{i=1}^{m}(x_i-a_i^2)^2 \tag{4}$$

$$\Omega_s=\frac{1}{m}\sum_{i-1}^{m}\left[(x_i+10)\log\frac{x_i+10}{a_i^2+10}+(10-x_i)\log\frac{10-x_i}{10-a_i^2}\right] \tag{5}$$

当数据从稀疏自动编码器的输入传输到输出,下一阶段包括评估成本函数和微调模型参数来获得最佳性能。同时,成本函数 E_{SAE} 并没有明确地将网络的权重和偏差关联起来。因此,还需定义敏感度来感知 E_{SAE} 中的变化,并通过反向传播学习方法将变化向后传递并迭代优化成本函数,本文采用随机梯度下降法,更新输出层的偏差和权重的随机梯度下降可以分别表示为:

$$\boldsymbol{b}'^2=\boldsymbol{b}^2-\eta^2\frac{\partial E_{SAE}}{\partial \boldsymbol{b}^2} \tag{6}$$

$$\boldsymbol{w}'^2=\boldsymbol{w}^2-\eta^2\frac{\partial E_{SAE}}{\partial \boldsymbol{w}^2} \tag{7}$$

其中,η^2 表示相对于输出层的学习速率。成本函数 E_{SAE} 的导数测量函数值相对于其输入值变化的敏感性。此外,梯度表示输入参数需要改变的程度以最小化成本函数。同时,利用链式法则计算梯度。因此,式(6)和(7)可以改写为:

$$\boldsymbol{b}'^2=\boldsymbol{b}^2-\eta^2\frac{\partial E_{SAE}}{\partial \boldsymbol{a}^2}\cdot\frac{\partial \boldsymbol{a}^2}{\partial \boldsymbol{b}^2} \tag{8}$$

$$\boldsymbol{w}'^2=\boldsymbol{w}^2-\eta^2\frac{\partial E_{SAE}}{\partial \boldsymbol{a}^2}\cdot\frac{\partial \boldsymbol{a}^2}{\partial \boldsymbol{w}^2} \tag{9}$$

SAE 输出层的灵敏度表示为:

$$\boldsymbol{S}^2=\frac{\partial E_{SAE}}{\partial \boldsymbol{a}^2} \tag{10}$$

因此,式(8)和(9)可以改写为:

$$\boldsymbol{b}'^2=\boldsymbol{b}^2-\eta^2\boldsymbol{S}^2 \tag{11}$$

$$\boldsymbol{w}'^2=\boldsymbol{w}^2-\eta^2\boldsymbol{S}^2(\boldsymbol{a}^1)^{\mathrm{T}} \tag{12}$$

其中,

$$\boldsymbol{S}^2=\begin{bmatrix}s_1^2\\ \vdots\\ s_m^2\end{bmatrix}=\begin{bmatrix}\dfrac{-(x_1+10)}{\lg(a_1^2+10)}+\dfrac{10-x_1}{\lg(10-a_1^2)}-(x_1-a_1^2)\\ \vdots\\ \dfrac{-(x_m+10)}{\lg(a_m^2+10)}+\dfrac{10-x_m}{\lg(10-a_m^2)}-(x_m-a_m^2)\end{bmatrix} \tag{13}$$

使用相同的方法计算 $\boldsymbol{S}^1$，可以将灵敏度传输回隐藏层：

$$\boldsymbol{b}'^1=\boldsymbol{b}^1-\eta^1\boldsymbol{S}^1 \tag{14}$$

$$\boldsymbol{w}'^1=\boldsymbol{w}^1-\eta^1\boldsymbol{S}^1(\boldsymbol{x})^{\mathrm{T}} \tag{15}$$

其中，η^1 表示相对于隐藏层的学习速率，而定义为：

$$\boldsymbol{S}^1=\begin{bmatrix}s_1^1\\ \vdots\\ s_m^1\end{bmatrix}=\begin{bmatrix}s_1^2w_{1,1}^2+s_2^2w_{2,1}^2+\cdots+s_m^2w_{m,1}^2\\ \vdots\\ s_1^2w_{1,n}^2+s_2^2w_{2,n}^2+\cdots+s_m^2w_{m,n}^2\end{bmatrix} \tag{16}$$

2　Softmax 回归

利用从 E_{SAE} 中学习到的特征对分类器进行训练。Softmax 回归又称多项式 logistic 回归(MLR)，是 logistic 回归的推广，可用于多类分类。在文献[12]中，Softmax 分类器可应用于多个二元分类任务并取得了良好的性能。Softmax 函数将输出解释为概率的方法，其表示为：

$$f(x_i)=\frac{\mathrm{e}^{x_i}}{\sum_{j=1}^{k}\mathrm{e}^{x_j}},i=1,2,\cdots,N \tag{17}$$

其中，$x_1,x_2,\cdots,x_N$ 表示输入值，输出 $f(x_i)$ 表示样本属于第 i 个标签的概率。对于 N 个输入样本，使用交叉熵成本函数测量 Softmax 层的误差：

$$L(\boldsymbol{w})=\frac{1}{N}\sum_{n=1}^{N}H(p_n,q_n)=$$

$$\frac{1}{N}\sum_{n=1}^{N}[y_n\log\hat{y}_n+(1-y_n)\log(1-\hat{y}_n)] \tag{18}$$

其中，真实概率 p_n 为实际标签，q_n 为预测值，$H(p_n,q_n)$ 是 p_n 和 q_n 之间差异的度量。此外，神经网络可能会陷入局部极小值，因此算法假设已经达到全局最小值，从而导致非最优性能。为了防止局部极小问题并进一步提高分类器的性能，本文采用带动量的小批量梯度下降法来优化分类器的交叉熵成本。该优化算法将训练数据分成小批量，然后用小批数据计算模型误差并更新模型参数，并且动量具有较好的收敛性。

本文所提出的方法可视化的流程，如图 2 所示。对初始数据集进行预处理，然后将其分为训练集和测试集。训练集用于以无监督的方式训练分析自动编码器，将测试集转换并输入到训练后的模型中以获得低维表示数据集。低维训练集用于训练 Softmax 分类器，并用低维测试集对 Softmax 分类器的性能进行测试。因此，分类器仅看到低维训练集，因此不存在数据泄漏。

3　实验分析

3.1　数据集

本文选取 3 种疾病数据集用于所提出方法的验证，以显示其在不同医疗诊断情况下的表现。数据集包括：1)Framingham 心脏研究数据集，该数据集从 Kaggle 网站获得，包含 4 238 个样本和 16 个特征；2)宫颈癌危险因素数据集取自加利福尼亚大学欧文分校(UCI)机器学习数据库，包含 858 个实例和 36 个属性；3)慢性肾病(CKD)数据集也是从 UCI 的机器学习数据库中获得，包含 400 个样本和 25 个特征。使用均值插补来处理数据集中缺失的变量。

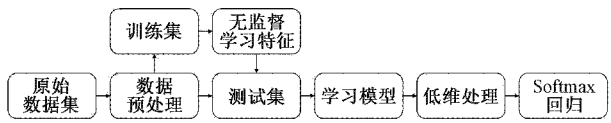

图 2　方法流程

3.2　参数选取

E_{SAE} 的训练参数包括：$\eta^1=0.01$，$\eta^2=0.1$，$n=25$，周期为 200。Softmax 分类器的超参数为：学习率为 0.01，小批量样本数为 32，动量值为 0.9，周期数为 200。根据文献[13]可知，这些超参数在不同的神经网络应用中都具有最佳性能。

3.3　评价指标

使用如下性能指标来评估所提出方法的有效性：准确性、精确性、召回率和 $F1$ 分数。

1)准确度是正确分类的实例与测试集中实例总数的比率：

$$AC=\frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

2)精度是正确预测的实例在预测患有疾病(即阳性)的实例中所占的比率：

$$P=\frac{TP}{TP+FP} \tag{20}$$

3)召回率是正确预测患者的比例：

$$R=\frac{TP}{TP+FN} \tag{21}$$

4)$F1$ 分数是衡量精确性和召回率之间平衡的指标：

$$F1=\frac{2P\cdot R}{P+R}=\frac{2TP}{2TP+FP+FN} \tag{22}$$

其中，真阳性(TP)表示病人被正确地预测为患病；假阳性(FP)表示健康人被错误地预测为患病；真阴性(TN)表示正确预测健康的人为健康；假阴性(FN)表示错误地预测病人为健康。

3.4　性能分析

为了验证所提出方法的有效性，将其与 LR、CART、SVM、KNN、LDA 和传统的 Softmax 回归进行了基准测试。为了体现所提方法的改进性能，对这些算法不进行任何参数调整。因此，使用了 Python 编程语言中软件机器学习库 Sklearn 中的默认参数值。采用 K-Fold 交叉验证技术对所有模型进行评价。表 1～3 分别给出了在 Framingham 心脏研究、宫颈癌危险因素和 CKD 数据集上测试的实验结

果。同时,图 3~5 分别给出传统 Softmax 分类器和所提方法在各种疾病预测模型中的性能的受试者工作特性(ROC)曲线。通过绘制真阳性率(TPR)和假阳性率(FPR)曲线得到 ROC 曲线,以此说明二元分类器的诊断能力。

表 1　本文方法和其他分类器在 Framingham 数据集上的性能　%

模型	准确度	精度	召回率	*F*1
本文方法	91	93	90	92
LR	83	84	86	84
CART	75	74	75	74
SVM	82	78	82	80
KNN	81	75	81	77
LDA	83	81	83	82
Softmax	86	84	88	86

表 2　本文方法和其他分类器在宫颈癌数据集上的性能　%

模型	准确度	精度	召回率	*F*1
本文方法	97	98	95	97
LR	94	96	91	93
CART	90	93	96	94
SVM	94	90	93	91
KNN	93	98	95	96
LDA	95	93	91	92
Softmax	94	97	91	94

表 3　本文方法和其他分类器在 CKD 数据集上的性能　%

模型	准确度	精度	召回率	*F*1
本文方法	98	97	97	97
LR	98	93	97	95
CART	95	97	95	96
SVM	96	94	96	95
KNN	94	93	89	91
LDA	96	97	93	95
Softmax	96	95	97	96

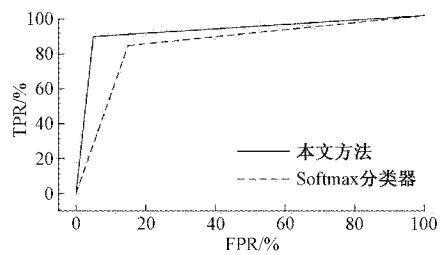

图 3　心脏病模型的受试者操作特性(ROC)曲线

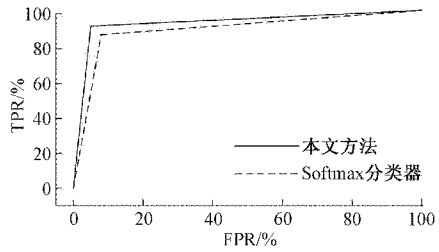

图 4　宫颈癌模型的 ROC 曲线

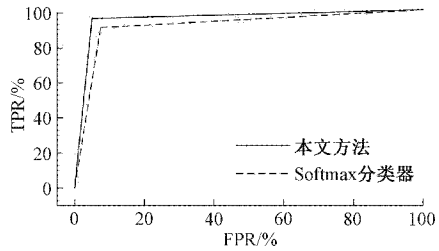

图 5　CKD 模型的 ROC 曲线

从实验结果可以看出,稀疏自编码器提高了 Softmax 分类器的性能,并通过各种模型的 ROC 曲线进一步验证。本文比其他机器学习算法的性能表现更好。此外,还考虑了模型在各种疾病预测中的错误分类。对于心脏病的预测,本文方法的 FPR 为 7%,假阴性率(FNR)为 10%。模型特异性(即真实阴性率(TNR))为 93%,TPR 为 90%。对于宫颈癌数据集得到的预测结果为:FPR = 3%,FNR=5%,TNR=97%,TPR=95%。CKD 的预测结果为:FPR=0,FNR=3%,TNR=100%,TPR=97%。

3.5　方法对比

为了进一步验证本文方法的性能,将本文方法与现有文献方法对心脏病数据集预测模型进行了比较,其他方法包括:基于 PSO 和 Softmax 回归特征选择方法[14],基于特征选择的 PSO 双层集成方法[15],基于 NB、Bayes 网络(BN)、RF 和 MLP 的集成分类器[16],基于 NB 和 LR 混合方法[17],基于混合 RF 与线性模型(HRFLM)[18],基于 LR 和 LASSO 回归[19],基于 NB 和高级加密标准(AES)的检测方法[20],基于 ANN 和模糊层次分析法(Fuzzy AHP)的混合方法[21],基于稀疏自编码器的 ANN 分类器[22]。本文方法与其他方法在心脏病数据集上的比较,如表 4 所示。

表 4　本文方法与其他方法在心脏病数据集上的比较　%

模型	精度	模型	精度
本文方法	91.00	HRFLM[18]	88.40
PSO+Softmax[14]	88.40	LASSO+LR[19]	89.00
PSO 双层集成[15]	85.71	NB+AES[20]	89.77
NB+BN+RF+MLP[16]	85.48	ANN+FAHP[21]	91.00
NB+LR[17]	87.40	SAE[22]	90.00

将本文方法与现有文献方法对宫颈癌数据集预测模型进行了比较，其他方法包括：基于主成分分析(PCA)的支持向量机[23]，C5.0 决策树[24]，基于隔离林(iForest)、合成少数过采样技术(SMOTE)和 RF 的多级分类过程[25]，基于稀疏自编码器的 ANN 分类器[26]，基于 C5.0 和 RF 的混合方法[27]。本文方法与其他方法在宫颈癌数据集上的比较，如表 5 所示。

表 5　本文方法与其他方法在宫颈癌数据集上的比较

%

模型	精度	模型	精度
本文方法	97.00	iForest+SMOTE+RF[25]	96.93
SVM+PCA[23]	94.03	SAE+ANN[26]	96.05
C5.0[24]	96.00	C5.0+RF[27]	96.90

将本文方法与现有文献方法对 CKD 数据集预测模型进行了比较，其他方法包括：优化的 XGBoost 方法[28]，自适应 Boosting 方法(AdaBoost)[29]，NB 和决策树的混合分类器(NBTree)[30]，XGBoost[31]，MLP 神经网络[32]。本文方法与其他方法在 CKD 数据集上的比较，如表 6 所示。

表 6　本文方法与其他方法在 CKD 数据集上的比较

%

模型	精度	模型	精度
本文方法	98.00	NBTree[30]	97.75
优化 XGBoost[28]	96.70	XGBoost[31]	97.29
AdaBoost[29]	88.66	MLP[32]	96.10

由表 4～6 可见，所提出的带有 Softmax 回归的稀疏自动编码器在各种疾病预测中的性能最佳。实验结果表明，稀疏自编码器可以有效地表达特征从而提高性能。这进一步证明了利用相关数据训练分类器的重要性，这是由于稀疏自编码器会显著影响预测模型的性能。同时，不仅通过执行算法的超参数调整，而且采用适当的特征学习技术可以获得良好的分类性能。

4　结　　论

本文设计了基于增强的稀疏自动编码器和 Softmax 回归的疾病预测方法。自动编码器通过惩罚隐藏层中的激活来实现稀疏性，从而使得网络的成本函数接近于 0。本文方法在心脏病、宫颈癌和 CKD 三种疾病上进行了实验，其准确率分别为 91%、97%和 98%，优于传统的 Softmax 回归和其他算法，与其他预测模型进行了比较研究，本文方法可以有效地提高预测精度。因此，本文方法可以作为一种有效的疾病医学诊断检测方法。

参考文献

[1] TIAN SH N, LU J H, GU B Y, et al. Medical diagnosis system based on fast-weights scheme[J]. Instrumentation, 2020, 7(1): 51-57.

[2] 郑光远，刘峡壁，韩光辉. 医学影像计算机辅助检测与诊断系统综述[J]. 软件学报，2018，29(5)：1471-1514.

[3] 王军浩，闫德勤，刘德山，等. 融合极端学习机的判别性分析字典学习算法[J]. 计算机科学，2020，47(5)：137-143.

[4] 蔚文婧，王寻，张鹏远，等. 一种基于多层感知器的房颤心电图检测方法[J]. 中国医学物理学杂志，2020，37(3)：332-336.

[5] 王尧，段锦，叶得前，等. 基于 RTV 模型图像分解的去雾算法[J]. 长春理工大学学报(自然科学版)，2020，43(4)：99-106.

[6] 王立可，崔小莉，张力戈. 多重属性过滤深度特征合成算法[J]. 计算机工程与应用，2020，56(12)：169-174.

[7] 张志敏，柴变芳，李文斌. 基于稀疏自编码器的属性网络嵌入算法[J]. 计算机工程，2020，46(7)：98-103，109.

[8] 尚青霞，周磊，冯亮. 基于降噪自动编码器的多任务优化算法[J]. 大连理工大学学报，2019，59(4)：417-426.

[9] 李晴晴，侯瑞春，丁香乾. 基于改进堆叠自编码器的滚动轴承故障诊断[J]. 计算机工程与设计，2019，40(7)：2064-2070.

[10] 张劲波，曾德生，骆金维. 块匹配约束下波束快速稀疏分解算法研究[J]. 微型电脑应用，2020，36(9)：39-41.

[11] 傅红笋，张艳. 源独立邻近梯度法求解频率域全波形稀疏约束反演问题[J]. 黑龙江大学自然科学学报，2020，37(4)：395-400.

[12] 李胜辉，白雪，董鹤楠，等. 基于小波与栈式稀疏自编码器的电力电缆早期故障定位方法研究[J]. 国外电子测量技术，2019，38(5)：146-151.

[13] 冉鹏，王灵，李昕，等. 改进 Softmax 分类器的深度卷积神经网络及其在人脸识别中的应用[J]. 上海大学学报(自然科学版)，2018，24(3)：352-366.

[14] WISAM E, AKHAN A, ABDUL H Z. Evolving deep learning architectures for network intrusion detection using a double PSO metaheuristic[J]. Computer Networks, 2020, 168(26): 107-119.

[15] 马学森，谈杰，陈树友，等. 云计算多目标任务调度的优化粒子群算法研究[J]. 电子测量与仪器学报，2020，34(8)：133-143.

[16] BEULAH C C, CAROLIN J S. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques[J]. Informatics in Medicine Unlocked, 2019, 16(1): 100-115.

[17] 轩华，李冰. 基于异步次梯度法的 LR 算法及其在多阶段 HFSP 的应用[J]. 运筹与管理，2015，24(6)：121-127.

[18] DURGADEVI V, KARTHIKEYAN R. Ensemble of

heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset[J]. Computer Methods and Programs in Biomedicine, 2020, 19(8):105-120.

[19] 谭勇,谢林柏,冯宏伟,等. 基于 LASSO 回归的红外火焰探测器的设计与实现[J]. 激光与红外,2019,49(6):720-724.

[20] 郑天琪,方献更. 一种抵抗侧信道攻击的 AES 算法协处理器架构设计[J]. 电子测试,2020(14):36-39.

[21] 陈树婷,谭大鹏. 基于 SA-ANN 的认知机制建模与识别优化算法[J]. 电子学报,2018,46(8):2011-2019.

[22] 李晓彬,牛玉广,葛维春,等. 基于改进堆叠自编码网络的电站辅机故障预警[J]. 仪器仪表学报,2019,40(6):39-47.

[23] 徐浩. 基于 PCA 算法和 SVM 的人脸识别系统[J]. 信息技术与信息化,2019(11):96-98.

[24] 石志凯,朱国胜,雷龙飞,等. 基于 C5.0 决策树的 NAT 设备检测方法[J]. 计算机科学, 2018,45(S1):323-327.

[25] IJAZ M F, ATTIQUE M, SON Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods [J]. Sensors, 2020, 20(10):119-131.

[26] 陈超强,蒋磊,王恒. 基于 SAE 和 LSTM 的下肢外骨骼步态预测方法[J]. 计算机工程与应用,2019,55(12):110-116,154.

[27] 刘丹,杨风暴,卫红,等. 基于多分类器的 C5.0 决策树植被分类方法[J]. 图学学报,2017,38(5):722-728.

[28] 周盛山,汤占军,王金轩,等. EEMD 和 CNN-XGBoost 在风电功率短期预测的应用研究[J]. 电子测量技术,2020,43(22): 55-61.

[29] 欧阳潇琴,王秋华. 基于改进权值更新和选择性集成的 AdaBoost 算法[J]. 软件导刊,2020,19(4): 257-262.

[30] 杨小军,钱鲁锋,别致. 基于 WEKA 平台的决策树算法比较研究[J]. 舰船电子工程,2018,38(10): 34-36,97.

[31] 罗春芳,张国华,刘德华,等. 基于 Kmeans 聚类的 XGBoost 集成算法研究[J]. 计算机时代, 2020(10):12-14.

[32] 李超,杨艳. 基于改进网中网神经网络的交通标志识别[J]. 信息技术,2019,43(9):137-140.

作者简介

孟祥莲,副教授,主要研究方向为人工智能。

E-mail:hesi5783882959@163.com

蒋巍,讲师,主要研究方向为机器学习。

李晓芳,教授,主要研究方向为信息获取与处理。