



文章编号:1001-9081(2019)10-2809-06

DOI:10.11772/j.issn.1001-9081.2019040624

改进的弹性网模型在深度神经网络中的应用

冯明皓¹, 张天伦¹, 王林辉¹, 陈 荣^{1*}, 连少静²

(1. 大连海事大学 信息科学技术学院, 辽宁 大连 116026; 2. 河北大学 数学与信息科学学院, 河北 保定 071002)

(*通信作者电子邮箱 rchen.cs@gmail.com)

摘要:由于具有较高的模型复杂度,深层神经网络容易产生过拟合问题,为了减少该问题对网络性能的不利影响,提出一种基于改进的弹性网模型的深度学习优化方法。首先,考虑到变量之间的相关性,对弹性网模型中的L1范数的不同变量进行自适应加权,从而得到L2范数与自适应加权的L1范数的线性组合。其次,将改进的弹性网络模型与深度学习的优化模型相结合,给出在这种新正则项约束下求解神经网络参数的过程。然后,推导出改进的弹性网模型在神经网络优化中具有群组选择能力和Oracle性质,进而从理论上保证该模型是一种更加鲁棒的正则化方法。最后,在多个回归问题和分类问题的实验中,相对于L1、L2和弹性网正则项,该方法的回归测试误差可分别平均降低87.09、88.54和47.02,分类测试准确度可分别平均提高3.98、2.92和3.58个百分点。由此,在理论和实验两方面验证了改进的弹性网模型可以有效地增强深层神经网络的泛化能力,提升优化算法的性能,解决深度学习的过拟合问题。

关键词:神经网络模型;深度学习;正则化方法;弹性网模型;过拟合

中图分类号:TP183 **文献标志码:**A

Improved elastic network model for deep neural network

FENG Minghao¹, ZHANG Tianlun¹, WANG Linhui¹, CHEN Rong^{1*}, LIAN Shaojing²

(1. College of Information Science and Technology, Dalian Maritime University, Dalian Liaoning 116026, China;

2. College of Mathematics and Information Science, Hebei University, Baoding Hebei 071002, China)

Abstract: Deep neural networks tend to suffer from overfitting problem because of the high complexity of the model. To reduce the adverse effects of the problem on the network performance, an improved elastic network model based deep learning optimization method was proposed. Firstly, considering the strong correlation between the variables, the adaptive weights were assigned to different variables of L1-norm in elastic network model, so that the linear combination of the L2-norm and the adaptively weighted L1-norm was obtained. Then, the solving process of neural network parameters under this new regularization term was given by combining improved elastic network model with the deep learning optimization model. Moreover, the robustness of this proposed model was theoretically demonstrated by showing the grouping selection ability and Oracle property of the improved elastic network model in the optimization of neural network. At last, in regression and classification experiments, the proposed model was compared with L1-norm, L2-norm and elastic network regularization term, and had the regression error decreased by 87.09, 88.54 and 47.02 and the classification accuracy improved by 3.98, 2.92 and 3.58 percentage points respectively. Thus, theory and experimental results prove that the improved elastic network model can effectively improve the generalization ability of deep neural network model and the performance of optimization algorithm, and solve the overfitting problem of deep learning.

Key words: neural network model; deep learning; regularization method; elastic network model; overfitting

0 引言

近年来,深度学习^[1]技术受到广泛关注,并在众多应用领域有着较好表现。理论上,深层神经网络能够拟合任意分布的数据,但是在实际中,有限的数据资源和较高的模型复杂度使得神经网络很难具有理想的泛化能力,过拟合现象由此

产生。该现象是机器学习中一种常见的病态问题,解决该问题的方法通常分为三种:第一种是扩充训练数据规模,例如,Szegedy等^[2]通过对原有图像样例进行翻转、裁剪等变化来增加训练数据,从而提升了神经网络在图像分类中的性能;与之不同,陈文兵等^[3]通过生成小样本数据来增加训练样本的数量,这种合成数据的方法同样可以降低神经网络的过拟合程

收稿日期:2019-04-15;修回日期:2019-07-03;录用日期:2019-07-09。

基金项目:国家自然科学基金资助项目(61672122, 61402070, 61602077);辽宁省自然科学基金资助项目(20170540097, 2015020023);辽宁省科学事业公益研究基金资助项目(GY-2017-0005);中央高校基本科研业务费资助项目(3132019207, 3132019355)。

作者简介:冯明皓(1995—),男,天津人,硕士研究生,主要研究方向:深度学习、最优化算法; 张天伦(1991—),男,河北保定人,博士研究生,主要研究方向:机器学习、计算视觉; 王林辉(1995—),男,山东烟台人,硕士研究生,主要研究方向:深度学习、文本分类; 陈荣(1969—),男,辽宁大连人,教授,博士,CCF会员,主要研究方向:人工智能、软件工程; 连少静(1990—),女,河北邯郸人,硕士,主要研究方向:机器学习。



度。第二种是更改训练方式,最具代表性的工作是 Dropout^[4],该方法通过在每次优化迭代中随机删除一些神经元来降低模型的复杂程度;除此之外,Zhang 等^[5]通过监控损失值的变化来决定网络训练的中止或重启,从而避免过拟合问题。第三种是在损失函数中引入正则项因子,对网络模型中的参数进行约束优化求解。与第一种方法相比,正则化方法没有增加额外的训练数据,从而不会加剧计算负担,同时,也不会受到冗余数据和噪声数据的影响;与第二种方法相比,正则化方法具有较好的收敛性,不会导致较长的训练周期。此外,正则化方法具有可靠的理论基础和严格的理论推导,因此,在解决过拟合问题的工作中,对正则化方法的研究与应用最为普遍。

正则化方法最早由 Tikhonov 等^[6]在 1963 年提出,并于 20 世纪 90 年代以后,成为一种主流的解决机器学习中解决病态问题的有效方法。本质上,该方法是将正则项作为含有解的先验知识引进经验风险中,从而约束解空间的范围,进而获得理想中的稳定解。在理论上,Antoniadis 等^[7]提出一个理想的正则化方法应该具备以下四个性质:1)连续性。所求参数的估计值在范围内应当连续,以获得一个更稳定的解。2)无偏性。所求参数的估计值应当是近似无偏的,以获得一个偏差较小的模型。3)稀疏性。所求模型能够将较小参数直接压缩为 0,降低模型的复杂程度。4)Oracle 性质。正则项能够正确识别模型的能力,可用渐进正态性和变量选择一致性来解释。在该领域里,这四个性质被广泛用作评价正则化方法的标准,因此在设计正则化方法时,应尽量保证所提出的模型具备以上性质。

在实际中,被广泛使用的正则化方法主要有:最小绝对值收敛和选择算子(Least Absolute Shrinkage and Selection Operator, LASSO)模型、岭回归模型和弹性网模型。其中:LASSO 模型又被称为 L1 正则化方法,该方法可以保证求解结果具有稀疏性^[8],降低原始数据维度,并可以过滤出重要的特征,因而该模型通常被用于高维数据的建模问题,例如,Cui 等^[9]利用 LASSO 模型构造特征池来对高维数据进行特征提取;Tang 等^[10]提出一种基于 L1 正则化方法的稀疏自动编码器模型。但是 LASSO 模型忽略了变量之间的相关性,不满足无偏性。岭回归模型又被称为 L2 正则化方法,这种方法在参数估计中限制较大的解,从而启发式地得到趋近于零的解,并且该方法可以保留变量之间的相关性,在样本维度高于样本规模时,可以获得光滑的稳定解。因此为了避免神经网络模型过于复杂,L2 正则化方法常被用于惩罚神经网络中较大的参数^[11];同时,Jin 等^[12]通过在神经网络优化问题中引入 L2 正则项来提高神经网络对离群点数据的鲁棒性。但是 L2 正则化方法往往得到较稠密的解,因而不具备稀疏性。弹性网模型是前两种方法的线性组合,该模型既具备 L1 正则化的特征选择能力和稀疏性的特点,又具备 L2 正则化保留变量之间相关性的特点,因而该模型在防止过拟合的问题中有更加广泛的应用^[13-14];但是在理论上,弹性网很难满足无偏性和 Oracle 性质。

为了更好地解决深度学习里的过拟合问题,本文在神经网络的损失函数中引入一种改进的弹性网模型,该模型可以被看作是 L2 正则化与加权的 L1 正则化的线性组合。在该模型里,L1 正则化中的不同变量被自适应地赋予不同的权重因

子,在优化神经网络参数时,这种加权的方式可以有选择性地保留重要的权重分量,使得到的网络参数具有稀疏性,从而降低过拟合的风险。同时,通过理论推导可以证明,在 L2 正则化的协同作用下,该模型具有合理的群组选择能力,因而相关性强的权重分量得以同时去除或者保留。除此之外,该模型还具备 Oracle 性质,可以保证对系数不为零的参数进行无偏估计。因而该模型是一种功能性更强的正则化模型。为了进一步验证该模型的实际效果,本文做了充分的对比实验,通过实验结果验证了该方法可以有效地防止深度学习中的过拟合问题,而且正则化的效果要明显优于现有的主流模型。

1 基于改进的弹性网的神经网络优化模型

深层神经网络(如图 1 所示)的学习目标是在给定的样本集上最小化期望损失,可以定义为:

$$E(J(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*)) = \int_{\mathbf{x}} J(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*) p(\mathbf{x}) d(\mathbf{x})$$

其中: \mathbf{w} 和 \mathbf{b} 分别是网络的权重和偏置,也是深度学习的优化对象; \mathbf{x} 是神经网络的输入,即样本属性; \mathbf{y}^* 是样本的期望输出; $p(\mathbf{x})$ 为样本的概率密度函数; $J(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*)$ 是关于权重 \mathbf{w} 和偏置 \mathbf{b} 的函数,该函数用来度量实际输出 \mathbf{y} 和期望输出 \mathbf{y}^* 之间的距离(可以是欧氏距离,也可以是 Kullback-Leibler(K-L) 散度)。

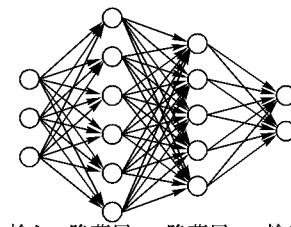


图 1 深层神经网络示意图

Fig. 1 Schematic diagram of deep neural network

为了提升神经网络的性能,一种最有效的办法是通过加入正则化项对损失函数进行约束求解,有如下定义:

$$J(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*) + R(\mathbf{w}) \quad (1)$$

其中: $R(\mathbf{w})$ 为正则化项。常用正则化项有: $R(\mathbf{w}) = \lambda_1 \sum_{i=1}^n |w_i|$,即 L1 正则化; $R(\mathbf{w}) = \lambda_2 \sum_{i=1}^n w_i^2$,即 L2 正则化; $R(\mathbf{w}) = \lambda_1 \sum_{i=1}^n |w_i| + \lambda_2 \sum_{i=1}^n w_i^2$,即弹性网。 λ 为正则项参数,表示正则项在整体模型中占的比重。这些模型的性质已在引言中进行讨论,为了克服这些现有模型的不足,一种改进的弹性网正则化模型被引入到神经网络的优化过程中。

首先,引入这种改进的弹性网模型:

$$R(\mathbf{w}) = \lambda_1^* \sum_{i=1}^n w_i' |w_i| + \lambda_2 \sum_{i=1}^n w_i^2 \quad (2)$$

其中:

$$w_i' = (|w_i^*| + \varepsilon)^{-\gamma}$$

ε 为一个非常小的数,用来防止分母为 0; w_i^* 为弹性网模型的参数优化结果; λ_1 必须和弹性网的参数相同; λ_1^* 可以和 λ_1 相同也可以不同;参数 γ 是一个正数。

最后,将改进的弹性网模型代入式(1),则深度学习优化



模型的定义可变形为:

$$J(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*) + \lambda_1^* \sum_{i=1}^n |w_i| + \lambda_2 \sum_{i=1}^n w_i^2 \quad (3)$$

在分类实验里,损失函数为交叉熵(Cross Entropy, CE)函数,定义如下:

$$J_{\text{CE}}(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*) = \frac{1}{n} \left(- \sum_{j=1}^n \sum_{i=1}^c y_{ji}^* \log y_{ji} \right)$$

其中: c 为输出层神经元的个数; n 为样本数。在回归实验里,损失函数为均方误差(Mean Squared Error, MSE),定义如下:

$$J_{\text{MSE}}(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*) = \frac{1}{n} \left(\sum_{i=1}^n (y_i^* - y_i)^2 \right)$$

具体地,这里选取目前最流行的深度学习优化算法为研究对象,即自适应矩估计(Adaptive moment estimation, Adam)^[15]。其中,在 Adam 模型中梯度的计算如下:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \frac{\partial J(\mathbf{w}, \mathbf{b})}{\partial \mathbf{w}} + \lambda_1 \sum_{i=1}^n w_i' \operatorname{sgn}(w_i) + \lambda_2 \sum_{i=1}^n 2w_i$$

其中: $\operatorname{sgn}(\cdot)$ 为符号函数, $\frac{\partial J(\mathbf{w}, \mathbf{b})}{\partial \mathbf{w}}$ 可以通过神经网络的BP

算法得到。基于改进的弹性网的神经网络优化算法的流程如算法 1 所示。

算法 1 基于改进的弹性网的神经网络优化算法。

输入 数据集 (\mathbf{X}, \mathbf{Y}) , Adam 算法超参数, 正则项系数 λ 。
输出 训练后的模型参数。

1) 输入数据集,构建弹性网模型进行迭代:

$$J(\mathbf{w}, \mathbf{b}; \mathbf{x}, \mathbf{y}^*) + \lambda_1 \sum_{i=1}^n |w_i| + \lambda_2 \sum_{i=1}^n w_i^2$$

2) 得到弹性网模型的参数后,按式(2) 构建改进的弹性网。

3) 按式(3) 构建新的优化模型,重新进行迭代。

2 改进的弹性网的相关性质

2.1 Oracle 性质

为了防止反向传播中的梯度消失现象,通常采用的激活函数为修正线性单元(Rectified Linear Unit, ReLU),此函数的形式如下:

$$f(x) = \max(0, x)$$

不失一般性,考虑神经元输出不为零的情况,此时一个神经元的前向计算可以被视为一个线性回归过程。为了表述方便,下面的证明中偏置暂时不被考虑。设第 l 层神经元的输入为 $n \times c^{l-1}$ 维度的矩阵 \mathbf{X} ,设期望输出为 $n \times c^l$ 维度的矩阵 \mathbf{Y} ,其中 c^{l-1} 和 c^l 分别表示第 l 层和第 $l-1$ 层神经元的个数。在衡量实际输出与期望输出的距离时,使用欧氏距离作为标准,此时优化目标可以表示如下:

$$\bar{\mathbf{w}}^* = \arg \min_{\mathbf{w}} \left(|\mathbf{Y} - \mathbf{X}\mathbf{w}|^2 + \lambda_1^* \sum_{j=1}^{c^{l-1} \times c^l} \hat{w}_j |w_j| + \lambda_2 \sum_{j=1}^{c^{l-1} \times c^l} w_j^2 \right)$$

Marquardt^[16]提出:对输入进行一些改造后,岭回归可以被视为最小二乘估计。基于此观点,对输入 \mathbf{X} 进行改造后,优化目标可以转换为如下问题:

$$\hat{\mathbf{w}}^* = \arg \min_{\mathbf{w}} \left(|\mathbf{Y}^* - \mathbf{X}^* \mathbf{w}|^2 + \lambda_1^* \sum_{j=1}^{c^{l-1} \times c^l} \hat{w}_j |w_j| \right)$$

其中: $\mathbf{X}^* = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} \mathbf{I} \end{pmatrix}$,维度为 $(n + c^l) \times c^{l-1}$; $\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix}$,维度

为 $(n + c^l) \times c^l$,根据改造后的数据集 $(\mathbf{X}^*, \mathbf{Y}^*)$,有:

1) $\mathbf{Y}^* = \mathbf{X}^* \tilde{\mathbf{w}} + \boldsymbol{\varepsilon}^*$,对于固定的 $\lambda_2, \boldsymbol{\varepsilon}_i^*$ 独立同分布服从 $N(0, \sigma^2)$, $\tilde{\mathbf{w}}$ 为模型的真实参数;

$$2) \mathbf{D}_n^* = \frac{1}{n} \mathbf{X}^{*\top} \mathbf{X}^* = \mathbf{D}_n + \frac{\lambda_2}{n} \mathbf{I} \rightarrow \mathbf{D} = \frac{1}{n} \mathbf{X}^T \mathbf{X}.$$

不妨设 $A = \{j \mid \tilde{w}_j \neq 0\}$,总共有 p_0 个不为0的真实参数,

$A^* = \{i \mid \tilde{w}_i^* \neq 0\}$,矩阵 \mathbf{D} 可以分块为 $\begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 \\ \mathbf{D}_3 & \mathbf{D}_4 \end{bmatrix}$,其中 \mathbf{D}_1 的维度为 $p_0 \times p_0$,定义为不等于0的参数对应的样本属性。

定理 1 Oracle 性质。假设 $(\lambda_1^*/n) \rightarrow 0, \lambda_1^* n^{(r-1)/2} \rightarrow \infty$ 并且 $(\lambda_2/n) \rightarrow 0$,则 Oracle 性质满足:

1) 演进正态性, $\sqrt{n}(\tilde{\mathbf{w}}_A^* - \tilde{\mathbf{w}}_A) \xrightarrow{d} N(0, \sigma^2 \mathbf{D}_1^{-1})$ 。

2) 变量选择一致性, $\lim_{n \rightarrow \infty} P(A^* = A) = 1$ 。

证明

对性质 1 的证明:

令 $\mathbf{w} = \tilde{\mathbf{w}} + \frac{\mathbf{u}}{\sqrt{n}}$,其中 \mathbf{u} 为任意 $c^{l-1} \times c^l$ 维矩阵,定义为:

$$\psi(\mathbf{u}) = \left(\mathbf{Y}^* - \mathbf{X}^* \left(\tilde{\mathbf{w}} + \frac{\mathbf{u}}{\sqrt{n}} \right) \right)^2 + \lambda_1^* \sum_{j=1}^{c^{l-1} \times c^l} \hat{w}_j \left| \tilde{w}_j + \frac{u_j}{\sqrt{n}} \right|$$

令 $\hat{\mathbf{u}} = \arg \min \psi(\mathbf{u})$,则 $\hat{\mathbf{u}} = \sqrt{n}(\tilde{\mathbf{w}}^* - \tilde{\mathbf{w}})$ 。其中: $\tilde{\mathbf{w}}^*$ 为模型参数估计值, $\tilde{\mathbf{w}}$ 为模型参数真实值,并且

$$\psi(\mathbf{0}) = (\mathbf{Y}^* - \mathbf{X}^* \tilde{\mathbf{w}})^2 + \lambda_1^* \sum_{j=1}^{c^{l-1} \times c^l} \hat{w}_j |\tilde{w}_j|$$

那么经过化简,可以得到:

$$\psi(\mathbf{u}) - \psi(\mathbf{0}) = V_1 + V_2 + V_3 + V_4$$

其中:

$$V_1 = \mathbf{u}^T \frac{\mathbf{X}^{*\top} \mathbf{X}^*}{n} \mathbf{u}$$

$$V_2 = -\boldsymbol{\varepsilon}^{*\top} \mathbf{X}^* \frac{\mathbf{u}}{\sqrt{n}}$$

$$V_3 = -\frac{\mathbf{u}^T \mathbf{X}^{*\top} \boldsymbol{\varepsilon}^*}{\sqrt{n}}$$

$$V_4 = \frac{\lambda_1^*}{\sqrt{n}} \sum_{j=1}^{c^{l-1} \times c^l} \hat{w}_j \sqrt{n} \left(\left| \tilde{w}_j + \frac{u_j}{\sqrt{n}} \right| - |\tilde{w}_j| \right)$$

容易得知:

$$\frac{\boldsymbol{\varepsilon}^{*\top} \mathbf{X}^*}{\sqrt{n}}, \frac{\mathbf{X}^{*\top} \boldsymbol{\varepsilon}^*}{\sqrt{n}} \xrightarrow{d} Q = N(0, \sigma^2 \mathbf{D}_n^*)$$

作出如下假设:

(A) 存在序列 $\{a_n\}$,使得 $a_n \rightarrow \infty$,并且有

$$\lim_{n \rightarrow \infty} a_n (\mathbf{w}^* - \tilde{\mathbf{w}}) = 0$$

$$\lim_{n \rightarrow \infty} \frac{\lambda_1^*}{\sqrt{n}} = 0$$

$$\lim_{n \rightarrow \infty} a_n^{\gamma} \frac{\lambda_1^*}{\sqrt{n}} \rightarrow \infty$$

对于 V_4 部分,可分情况来看讨论:

1) 当 $u_j = 0$ 时,易得 $V_4 \xrightarrow{P} 0$ 。

2) 当 $u_j \neq 0, \tilde{w}_j \neq 0$ 时,有 $\hat{w}_j \xrightarrow{P} |\tilde{w}_j|^{-\gamma}$,



$\sqrt{n} \left(\left| \tilde{w}_j + \frac{u_j}{\sqrt{n}} \right| - |\tilde{w}_j| \right) \rightarrow u_j \operatorname{sgn}(\tilde{w}_j)$, 根据 Slutsky 定理的得:

$$V_4 \xrightarrow{P} 0.$$

3) 当 $u_j \neq 0, \tilde{w}_j = 0$ 时, $\sqrt{n} \left(\left| \tilde{w}_j + \frac{u_j}{\sqrt{n}} \right| - |\tilde{w}_j| \right) = |u_j|$,

$$\frac{\lambda_1^* \hat{w}_j}{\sqrt{n}} = \frac{\lambda_1^*}{\sqrt{n}} (\sqrt{n})^\gamma \frac{1}{|\sqrt{n} w_j^*|^\gamma},$$

根据假设(A), 令序列 a_n 为 \sqrt{n} , 可以得到 $\frac{\lambda_1^*}{\sqrt{n}} (\sqrt{n})^\gamma \rightarrow \infty$; 由于 $\gamma > 0$, 则 $\frac{1}{|\sqrt{n} w_j^*|^\gamma} =$

$$\frac{1}{|\sqrt{n}(w_j^* - 0)|^\gamma} \rightarrow \infty,$$

所以 $V_4 \xrightarrow{P} \infty$ 。

综上, 令 $V_5 = \psi(\mathbf{u}) - \psi(\mathbf{0})$, 则有:

$$V_5 \xrightarrow{d} \begin{cases} \mathbf{u}_A^T \left(\mathbf{D}_1 + \frac{\lambda_2}{n} \mathbf{I} \right) \mathbf{u}_A - \mathbf{u}_A^T \mathbf{Q}_A - Q_A \mathbf{u}_A, \\ u_j = 0 \text{ or } j \in A \\ \infty, \text{ 其他} \end{cases}$$

此时 V_4 为凸函数并且唯一极小值点为:

$$\left(\left(\mathbf{D}_1 + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} Q_A, 0 \right)$$

其中 $Q_A = N(0, \sigma^2 \mathbf{D}_1)$, 则有:

$$\hat{\mathbf{u}}_A = \sqrt{n} (\tilde{\mathbf{w}}_A^* - \hat{\mathbf{w}}_A) \xrightarrow{d} N(0, \sigma^2 \mathbf{D}_1^{-1})$$

对性质 2) 的证明:

命题可以等价于 $\forall j' \notin A, P(j' \in A^*) \rightarrow 0$, 由 KKT(Karush-Kuhn-Tucker) 条件可得:

$$2\mathbf{X}_{j'}^{*\top} (\mathbf{Y}^* - \mathbf{X}^* \tilde{\mathbf{w}}^*) = \lambda_1^* \hat{w}_{j'}, \quad (4)$$

其中: $\mathbf{X}_{j'}^{*\top}$ 代表所有样本的某个和权重 $\tilde{w}_{j'}^*$ 对应的属性向量。

由对性质 1) 的证明可得 $\frac{\lambda_1^*}{\sqrt{n}} \hat{w}_{j'} = \frac{\lambda_1^*}{\sqrt{n}} (\sqrt{n})^\gamma \frac{1}{|\sqrt{n} w_{j'}^*|^\gamma} \rightarrow \infty$, 则

式(4) 等号左侧可以写成:

$$\begin{aligned} & \frac{2}{\sqrt{n}} \mathbf{X}_{j'}^{*\top} (\boldsymbol{\varepsilon}^* + \mathbf{X}^* \tilde{\mathbf{w}} - \mathbf{X}^* \tilde{\mathbf{w}}^*) = \\ & 2 \left(\frac{\mathbf{X}_{j'}^{*\top} \mathbf{X}^* \sqrt{n} (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^*)}{n} + \frac{\mathbf{X}_{j'}^{*\top} \boldsymbol{\varepsilon}^*}{\sqrt{n}} \right) \end{aligned}$$

由对性质 1) 的证明和 Slutsky 定理可知:

$\frac{\mathbf{X}_{j'}^{*\top} \mathbf{X}^* \sqrt{n} (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^*)}{n}$ 和 $\frac{\mathbf{X}_{j'}^{*\top} \boldsymbol{\varepsilon}^*}{\sqrt{n}}$ 分别收敛于某两个正态分布,

则 $P(j' \notin A) \leq P(2\mathbf{X}_{j'}^{*\top} (\mathbf{Y}^* - \mathbf{X}^* \tilde{\mathbf{w}}^*) = \lambda_1^* \hat{w}_{j'}) \rightarrow 0$ 。

综上, 通过证明改进的弹性网模型满足性质 1) 与性质 2), 说明了该模型具有 Oracle 性质。该性质表明: 该模型的估计值可以以 1 的概率正确估计非零的参数, 并且估计值的非零部分服从渐近正态分布。

2.2 群组选择能力

定理 2 群组选择能力。给定神经元的输入和期望输出, 假设估计值 $\tilde{w}_i^*, \tilde{w}_j^*$ 均大于 0, 并且 $\tilde{w}_i^*, \tilde{w}_j^*$ 对应的为不同的样本属性, 定义 $B(i, j) = \frac{1}{|\mathbf{Y}|} |\tilde{w}_i^* - \tilde{w}_j^*|$, 所以

$$B(i, j) \leq \frac{1}{\lambda_2} \left(|\mathbf{X}_i - \mathbf{X}_j| + \frac{\gamma \lambda_1^*}{2 |\mathbf{Y}|} |\tilde{w}_i - \tilde{w}_j| \right)$$

证明 由假设可以得到 $\operatorname{sgn}(\tilde{w}_i^*) = \operatorname{sgn}(\tilde{w}_j^*)$, 定义:

$$\tilde{\mathbf{w}}^* = \arg \min_{\mathbf{w}} L(\lambda_1^*, \lambda_2, \mathbf{w})$$

让 L 分别对 \tilde{w}_i^* 和 \tilde{w}_j^* 求偏导为 0, 可得

$$-2\mathbf{X}_i^T (\mathbf{Y} - \mathbf{X} \tilde{\mathbf{w}}^*) + \lambda_1^* \hat{w}_i \operatorname{sgn}(\tilde{w}_i^*) + 2\lambda_2 \tilde{w}_i^* = 0 \quad (5)$$

$$-2\mathbf{X}_j^T (\mathbf{Y} - \mathbf{X} \tilde{\mathbf{w}}^*) + \lambda_1^* \hat{w}_j \operatorname{sgn}(\tilde{w}_j^*) + 2\lambda_2 \tilde{w}_j^* = 0 \quad (6)$$

将式(5) 减去式(6), 在等号两边加上绝对值, 化简得

$$\begin{aligned} \lambda_2 |\tilde{w}_i^* - \tilde{w}_j^*| &= \\ &\left| (\mathbf{X}_i^T - \mathbf{X}_j^T) \mathbf{r} - \frac{\lambda_1^*}{2} (\hat{w}_i - \hat{w}_j) \operatorname{sgn}(\tilde{w}_i^*) \right| \end{aligned}$$

其中: $\mathbf{r} = \mathbf{Y} - \mathbf{X} \tilde{\mathbf{w}}^*$ 。

由 $|A| - |B| \leq |A \pm B| \leq |A| + |B|$, 可得:

$$\lambda_2 |\tilde{w}_i^* - \tilde{w}_j^*| \leq |(\mathbf{X}_i^T - \mathbf{X}_j^T) \mathbf{r}| + \left| \frac{\lambda_1^*}{2} (\hat{w}_i - \hat{w}_j) \right|$$

易知: $L(\lambda_1^*, \lambda_2, \tilde{w}_i^*) \leq L(\lambda_1^*, \lambda_2, \tilde{w}^* = 0)$, 即 $|\mathbf{r}| \leq |\mathbf{Y}|$ 。则

$$\lambda_2 |\tilde{w}_i^* - \tilde{w}_j^*| \leq |(\mathbf{X}_i^T - \mathbf{X}_j^T) \mathbf{Y}| + \left| \frac{\lambda_1^*}{2} (\hat{w}_i - \hat{w}_j) \right|$$

不等式两端同时除以 $|\mathbf{Y}|$ 和 λ_2 , 可得:

$$B(i, j) \leq \frac{1}{\lambda_2} \left(|\mathbf{X}_i - \mathbf{X}_j| + \frac{\lambda_1^*}{2 |\mathbf{Y}|} |\hat{w}_i - \hat{w}_j| \right)$$

假设 \hat{w}_j 的相容估计为 $|\tilde{w}_j|^{-\gamma}$, 定义一个函数 $f(x) = x^{-\gamma}$,

由中值定理得 $|f(x) - f(y)| \leq \gamma |x - y|$ 。

同理:

$$|\hat{w}_i - \hat{w}_j| = ||\tilde{w}_i|^{-\gamma} - |\tilde{w}_j|^{-\gamma}| \leq \gamma ||\tilde{w}_i| - |\tilde{w}_j||$$

由绝对值不等式 $||a| - |b|| \leq |a \pm b| \leq |a| + |b|$, 可得

$$\gamma ||\tilde{w}_i| - |\tilde{w}_j|| \leq \gamma |\tilde{w}_i - \tilde{w}_j|$$

则

$$B(i, j) \leq \frac{1}{\lambda_2} \left(|\mathbf{X}_i - \mathbf{X}_j| + \frac{\gamma \lambda_1^*}{2 |\mathbf{Y}|} |\tilde{w}_i - \tilde{w}_j| \right)$$

不失一般性, 可假设: 输入 \mathbf{X} 服从标准正态分布。

令 $\rho = \mathbf{X}_i^T \mathbf{X}_j$, 则

$$|\mathbf{X}_i - \mathbf{X}_j| = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^T (\mathbf{X}_i - \mathbf{X}_j)} = \sqrt{2(1 - \rho)}$$

根据标准正态分布的性质, 可得 $\rho \approx 1$, $|\mathbf{X}_i - \mathbf{X}_j| \approx 0$, 当两个变量强相关时, 则这两个变量对于决策变量的影响是近似相同的, 模型分配给它们的真实参数也近似相同, 即 $|\tilde{w}_i - \tilde{w}_j|$ 趋近于 0, 那么 $B(i, j)$ 也会趋近于 0, 即 $|\tilde{w}_i^* - \tilde{w}_j^*|$ 趋近于 0。通过以上证明可以看出, 群组选择能力具体表现为: 若两个参数同时为 0, 则这两个参数对应的变量会被模型同时去掉; 若两个参数近似相同且不为 0, 则这两个参数对应的变量会被模型同时保留。

3 实验与分析

3.1 实验数据集

本文实验采用的数据集包括分类实验数据集和回归实验数据集两个部分, 这些数据集的信息分别汇总在表 1~2 中。这些数据集来自 UCI 机器学习数据库网站和 KEEL 数据集网站, 并且在回归问题中都是对单一属性进行回归预测。为了更好地进行实验, 数据集在标准化处理后被划分为训练集、验证集和测试集。



表1 实验中所使用的分类数据集

Tab. 1 Datasets used in the classification experiment

数据集	属性数	分类数	样本数		
			训练集	验证集	测试集
Human	561	6	500	200	200
Image	19	7	500	200	200
Car	6	4	500	200	200
Balance	4	3	400	100	100
Pima	8	2	400	150	150
Sports	59	2	500	200	200
Contraceptive	9	3	500	200	200
Vehicle	18	4	500	150	150

表2 实验中所使用的回归数据集

Tab. 2 Datasets used in the regression experiment

数据集	属性数	样本数		
		训练集	验证集	测试集
Airfoil	5	300	100	100
Concrete	8	300	100	100
Estate	6	250	80	80
Power	4	300	100	100
Wine	11	500	200	200
Mortgage	15	500	200	200
Stock	9	500	200	200
Weather	9	500	200	200

3.2 实验设置

所有实验的实验环境为 Windows 10(64位)操作系统, Python 3.6 以及 Tensorflow 1.5 GPU 版本, 显卡为 GTX-1060 6 GB显存。实验中用到的神经网络为全连接的多隐层前向传播网络模型, 并且对于同一个数据集采用的网络结构相同。超参数 λ_1 和 λ_2 的取值范围是 $\{10^7, 10^6, \dots, 10^{-7}\}$, 在每个数据集上固定网络结构, 经过多次实验, 使得 L1、L2 和弹性网(Elastic Net, EN) 模型取得最好结果的参数作为该数据集上实验对比所用的参数, 改进的弹性网(Advanced EN, AEN)模型的 λ_1^* 和 λ_2 设置与 EN 相同, 后续实验可以验证, 在 EN 达到最好效果时, 改进的弹性网仍能继续改善网络的性能。训练集用来进行模型参数的学习, 验证集用来评估不同阶段的模型的表现; 测试集用来评估训练结束后的模型的性能。分类实验采用交叉熵(CE)作为验证集上的评价指标, 同时采用准确度(Accuracy, ACC)作为测试集上的评价指标; 回归实验采用均方误差(MSE)作为验证集和测试集上的评价指标。

3.3 实验分析

本文主要对比的方法有: 在 L1 正则项约束下的优化模型(L1), 在 L2 正则项约束下的优化模型(L2), 在弹性网约束下的优化模型(EN), 以及在本文方法约束下的优化模型(AEN)。图 2 展示了在其中四个数据集的验证集上的对数损失值随迭代次数的变化趋势。图 3~4 为在不同数据集上训练结果和测试结果之间的差值, 这些差值反映了模型的过拟合程度, 差值越大, 过拟合情况越严重。具体表现为: 在分类问题上, 测试准确率越低于训练准确率; 在回归问题上, 测试误差越高于训练误差。表 3~4 则给出了以上方法在不同测试集上的最终测试结果, 以及 AEN 相对于 L1、L2 和 EN 的平均准确率提升数值和均方误差下降数值。

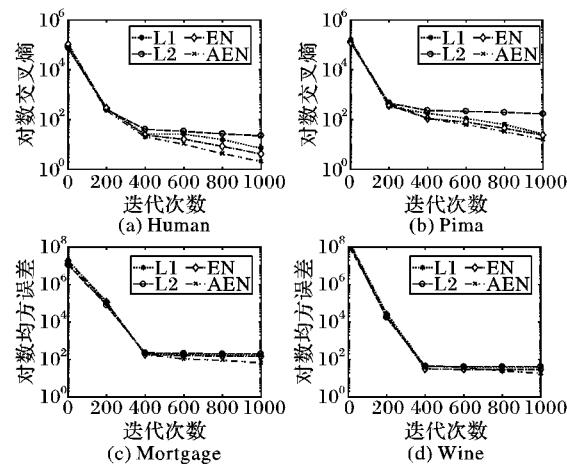


图2 四个数据集的对数损失变化曲线

Fig. 2 Logarithm loss curve of four datasets

表3 四种方法在分类数据测试集的准确率

Tab. 3 Accuracy of four methods on classification test sets

数据集	准确率(ACC)/%			
	L1	L2	EN	AEN
Human	80.00	83.99	81.49	87.50
Image	82.99	83.49	82.99	87.99
Car	92.50	93.99	92.50	94.99
Balance	92.00	92.00	94.99	97.00
Pima	75.99	75.99	72.67	77.33
Sports	82.99	84.50	81.00	86.00
Contraceptive	46.50	45.50	48.50	50.00
Vehicle	78.66	80.66	80.66	82.67
提升平均值	3.98	2.92	3.58	—

表4 四种方法在回归数据测试集的最终结果

Tab. 4 Final results of four methods on regression test sets

数据集	均方误差(MSE)			
	L1	L2	EN	AEN
Airfoil	23.8702	28.0700	22.6269	14.8174
Concrete	72.3202	58.3473	47.7001	45.2417
Estate	220.7445	270.2336	157.6485	152.9333
Power	124.3598	128.1608	126.8739	74.1758
Wine	43.4986	49.4433	37.9569	10.9873
Mortgage	327.3597	357.4577	190.8707	94.3630
Stock	80.5898	70.6077	48.6104	39.8147
Weather	448.6036	390.6130	388.5385	212.2920
下降平均值	87.09	88.54	47.02	—

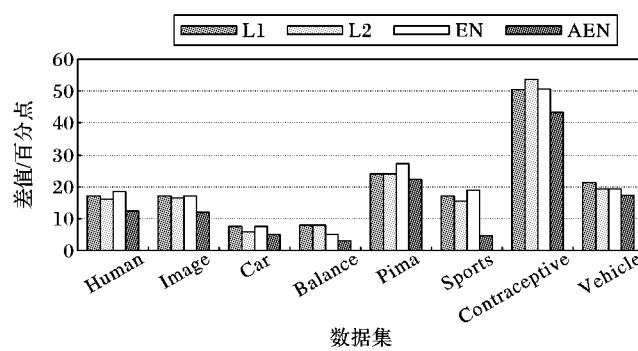


图3 分类数据集的训练集和测试集准确率之差

Fig. 3 Accuracy difference between training set and test set of classification datasets

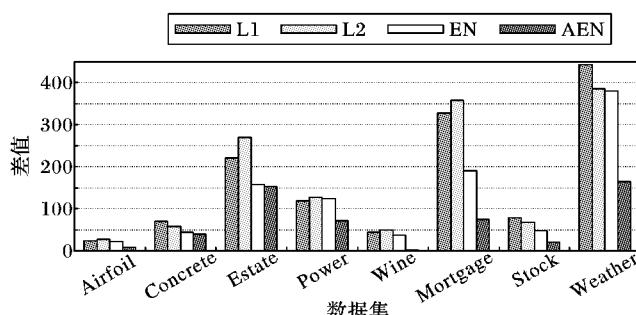


图4 回归数据集的训练集和测试集损失之差
Fig.4 Loss difference between training set and test set of regression datasets

从图3~4可看出,AEN在防止过拟合的问题中取得了较为突出的表现。除此之外,从表3~4可看出:AEN方法在大多数情况下可以得到更低的回归损失值和较高的分类准确率,且通过图2中在验证集上的损失变化曲线可以看出:L1方法由于其稀疏性更容易得到一个较低的损失;L2方法由于其平滑的特性,曲线一般更加光滑稳定,但一般不会得到一个较低的损失;EN方法由于将两者结合,经常会得到一个比前两者更好的结果;AEN方法在不同的数据集上可以得到比其他方法更低的预测损失。结合图3~4和表3~4可以得出:AEN方法不仅可以在训练阶段使得深层神经网络模型更鲁棒地拟合数据分布,而且所得到的模型在未知数据上也有较好的泛化能力。产生这些结果的原因可以被归结为以下两点:1)在弹性网的框架下,能够兼顾稀疏性和平滑性,在能够减轻模型复杂度的同时又容易找到最优解;2)改进的弹性网模型是在弹性网基础上对L1部分再进行加权,对不同的参数,配给不同的权重变量,对于较大的参数将会配给一个较小的权重,对于较小的参数将会配给一个较大的权重,这样在迭代更新时,较大的、重要的参数将会更容易被保留下,较小的、不重要的参数也会更容易接近0。

4 结语

本文对深度学习以及正则化方法进行了研究与分析。首先讨论了L1正则化,L2正则化和弹性网正则化,并基于这些正则化模型的优势与缺点,提出一种基于弹性网的改进模型。然后,将这个改进的弹性网正则化方法与深度学习算法相结合,提出一种对神经网络参数进行约束求解的新方法。在理论上,证明了这个改进的弹性网模型具有群组选择能力和Oracle性质,这些性质的证明可以保证改进的方法在一定程度上避免深度学习中的过拟合问题,从而提高深层神经网络的泛化能力。在实验上,通过对不同方法在多个验证集和测试集的表现,可以看到改进的弹性网模型不仅可以在训练阶段较鲁棒地收敛至较低的损失值,在对未知样本的预测阶段也体现出了较好的泛化能力,通过对实验结果进行分析,得出改进的弹性网模型取得好的效果的原因,日后主要的研究目标是把本文方法应用在更多的实际问题中。

参考文献(References)

- [1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.
 - [2] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1–9.
 - [3] 陈文兵, 管正雄, 陈允杰. 基于条件生成式对抗网络的数据增强方法[J]. 计算机应用, 2018, 38(11): 3305–3311. (CHEN W B, GUAN Z X, CHEN Y J. Data augmentation method based on conditional generative adversarial net model[J]. Journal of Computer Applications, 2018, 38(11): 3305–3311.)
 - [4] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.
 - [5] ZHANG Z, LUO P, LOY C C, et al. Facial landmark detection by deep multi-task learning[C]// Proceedings of the 2014 European Conference on Computer Vision, LNCS 8694. Cham: Springer, 2014: 94–108.
 - [6] TIKHONOV A N. Solution of incorrectly formulated problems and the regularization method[J]. Doklady Akademii Nauk SSSR, 1963, 151: 501–504.
 - [7] ANTONIADIS A, FAN J. Regularization of wavelet approximations[J]. Journal of the American Statistical Association, 2001, 96(455): 939–967.
 - [8] LIAN L, LIU A, LAU V K N. Weighted LASSO for sparse recovery with statistical prior support information[J]. IEEE Transactions on Signal Processing, 2018, 66(6): 1607–1618.
 - [9] CUI C, WANG D. High dimensional data regression using Lasso model and neural networks with random weights[J]. Information Sciences, 2016, 372: 505–517.
 - [10] TANG J, DENG C, HUANG G. Extreme learning machine for multilayer perceptron[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(4): 809–821.
 - [11] PHAISANGITISAGUL E. An analysis of the regularization between L2 and dropout in single hidden layer neural network[C]// Proceedings of the 7th International Conference on Intelligent System, Modelling and Simulation. Piscataway: IEEE, 2016: 174–179.
 - [12] JIN J, CHEN C L P. Regularized robust broad learning system for uncertain data modeling[J]. Neurocomputing, 2018, 322: 58–69.
 - [13] 李光早, 王士同. 基于稀疏表示和弹性网络的人脸识别[J]. 计算机应用, 2017, 37(3): 901–905. (LI G Z, WANG S T. Face recognition based on sparse representation and elastic network [J]. Journal of Computer Applications, 2017, 37(3): 901–905.)
 - [14] LI Q, SUN Y, WANG C, et al. Elastic net hypergraph learning for image clustering and semi-supervised classification[J]. IEEE Transactions on Image Processing, 2017, 26(1): 452–463.
 - [15] KINGMA D P, BA J L. Adam: a method for stochastic optimization[EB/OL]. [2019-01-10]. <https://arxiv.org/pdf/1412.6980.pdf>.
 - [16] MARQUARDT D W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation[J]. Technometrics, 1970, 12(3): 591–612.
- This work is partially supported by the National Natural Science Foundation of China (61672122, 61402070, 61602077), the Natural Science Foundation of Liaoning Province (20170540097, 2015020023), the Public Research Funds for Scientific Ventures of Liaoning Province (GY-2017-0005), the Fundamental Research Funds for the Central Universities (3132019207, 3132019355).
- FENG Minghao, born in 1995, M. S. candidate. His research interests include deep learning, optimization algorithm.
- ZHANG Tianlun, born in 1991, Ph. D. candidate. His research interests include machine learning, computer vision.
- WANG Linhui, born in 1995, M. S. candidate. His research interests include deep learning, text categorization.
- CHEN Rong, born in 1969, Ph. D., professor. His research interests include artificial intelligence, software engineering.
- LIAN Shaojing, born in 1990, M. S. Her research interests include machine learning.