

基于相对概率变化比的 CNN 超参数优化方法

李 慧¹, 周溪召¹, 施柏州²

(1. 上海理工大学 管理学院, 上海 200093; 2. 朝阳科技大学 理工学院, 台中 413310)

摘要: 卷积神经网络 (convolutional neural network, CNN) 已被广泛应用于图像识别领域, 其自身的超参数对图像分类问题中分类错误率的大小有较大的影响。为进一步优化 CNN 超参数, 提出了基于 Softmax 回归的相对概率变化比。应用相对概率变化比寻找对图像分类影响较大的超参数, 并根据超参数的重要性大小依次对其进行调整。为验证相对概率比的有效性, 在网络架构 1 和网络架构 2 下进行了调参实验。实验结果表明, Softmax 回归的相对概率变化比在不同的网络架构下均可反映 CNN 超参数对分类错误率的影响, 且有助于找到一个更优的超参数组合, 降低分类错误率。在 MNIST 和 CIFAR-10 数据集上的对比实验表明, 研究结果在不同数据集下都适用。

关键词: 卷积神经网络; 超参数组合; Softmax 回归; 相对概率变化比

中图分类号: TP 301-6

文献标志码: A

CNN hyper-parameters optimization method based on the change ratio of relative probability

LI Hui¹, ZHOU Xizhao¹, SHI Baizhou²

(1. Business School, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. College of Science and Engineering, Chaoyang University of Technology, Taichung 413310, China)

Abstract: Convolutional neural network (CNN) has been widely used in the field of image classification. Its hyper-parameters have a great impact on the misclassification rate. In order to further optimize CNN hyper-parameters, the change ratio of relative probability based on Softmax regression was introduced. The change ratio of relative probability was used to find the hyper-parameters which have great influence on the image classification problem, and the hyper-parameters were adjusted according to their importance. In order to verify the effectiveness of the change ratio of relative probability, experiments were carried out in Architecture 1 and Architecture 2. The experiment results show that the concept of the change ratio of relative probability introduced in Softmax regression can reflect the effect of CNN hyper-parameters on misclassification rate in both architectures, and it is more helpful to find a better combination of CNN hyper-parameters and reduce the misclassification rate. The comparative experiments on MNIST and CIFAR-10 datasets show that the above conclusions are applicable to different datasets.

Keywords: convolutional neural network; combination of hyper-parameters; Softmax regression;

收稿日期: 2020-10-23

基金项目: 国家自然科学基金资助项目 (61273042)

第一作者: 李 慧 (1996-), 女, 硕士研究生. 研究方向: 自动机器学习. E-mail: lihui112665@163.com

change ratio of relative probability

近年来,随着计算机硬件的发展以及大数据的出现,卷积神经网络(convolutional neural network, CNN)在图像识别和目标检测领域取得了较大的进展^[1-4]。其中, AlexNet 在 2012 年的 ImageNet 竞赛中以 18.9% 的错误率获得了冠军,而 VGG 网络在 2014 年的这一比赛中以 7.3% 的错误率获得了亚军。比较这两个网络可发现: VGG 网络采用 3×3 的卷积核大小,池化层采用 2×2 的窗口; AlexNet 采用 11×11、5×5 以及 3×3 的卷积核大小,池化层主要采用 3×3 的窗口。两者之间的错误率差别表明 CNN 超参数会影响模型分类正确率。Shih 等^[5]和 Bibaeva^[6]也在文章中指出, CNN 在图像识别上的分类正确率很大程度上受到自身超参数的影响,比如卷积核的高、卷积核的宽、卷积核的步长以及池化窗口的步长。

为了找到更优的 CNN 超参数组合,提高分类正确率,已有较多的方法应用于 CNN 的超参数优化。CNN 超参数优化方法可分为两类:一类为可视化优化方法,比如 Zeiler 等^[7]利用反卷积可视化 AlexNet 的第一层和第二层,将 AlexNet 的第一层卷积核大小从 11×11 减小为 7×7,第二层步长改为 2,改善后的模型在 ImageNet 2012 上的分类正确率高于原来的 AlexNet;另一类为自动超参数优化方法,在自动超参数优化中,网格搜索(grid search)通过初步确定的超参数范围,基于穷举法来确定最优的 CNN 超参数组合。Bergstra 等^[8]证明,相比于网格搜索,在同一域内进行随机搜索(random search)能够在更短的计算时间内找到相同甚至更好的模型。此外, Snoek 等^[9]提出采用贝叶斯优化(Bayesian optimization)来寻找更优的超参数。具体而言, Snoek 等^[9]为进一步提高卷积神经网络在 CIFAR-10 上的分类正确率,在 Krizhevsky 等^[10]提出的模型基础上微调了 9 个超参数,最终测试集上的分类正确率相比于文献[10]提高了 3%。

值得注意的是,上述几种方法均只能返回最终搜索到的超参数,难以得到单个超参数对 CNN 模型性能的影响。而研究单个超参数对 CNN 性能的影响不仅有助于专家对超参数进行调整,也有助于专家更好地理解算法的原理,开发出性能更优的 CNN 模型。为此,本文基于 Softmax 回归提出了相对概率变化比的概念,对 CNN 超参数进行优化。此概念不仅可以反映各超参数对模型性能

的影响程度,还能得到各超参数的最佳调整方向,有助于缩短专家的调参时间。

1 发生比变化率

19 世纪 60 年代,逻辑回归(logistic regression, LR)第一次被 Cornfield 等^[11]提出。LR 具有两大功能:一是发现对因变量影响较大的“危险”因素,比如 Domínguez-Almendros 等^[12]根据 LR 的分析结果得出,患哮喘的母亲有更大的可能性生出一个患哮喘的新生儿,换言之,患哮喘的母亲对于新生儿患哮喘来说,是一个“危险”因素;二是根据 LR 模型,判断新的样本属于哪一类别,比如 Phillips 等^[13]采用 LR 对采矿业使用的卡车油样进行分析,以此来判断卡车的健康状况。

LR 主要用于解决二分类问题,它采用 Logit 函数将线性回归预测值转化为 (0,1) 之间的概率值,用来表示事件发生的概率。Logit 函数形式为

$$g(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

LR 的因变量一般只有两种取值,即某个事件发生或某个事件不发生。假设事件发生的概率为 P ,则不发生的概率为 $(1 - P)$,概率 P 的计算为

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad (2)$$

式中, β_i 为回归系数, $i = 1, 2, \dots, n$ 。

$P/(1 - P)$ 为几率或发生比(odds),代表某个事件发生与不发生的概率比值,具体表达式为

$$\frac{P}{1 - P} = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)} \quad (3)$$

从式(3)可知,当因变量 x_k ($k \in [1, n]$) 变化时,发生比也会发生改变,本文将 x_k 增加一个单位而引起的发生比的变化称为 x_k 的发生比变化率,用 P_{CR} 表示。当 $\beta_k > 0$ ($k \in [1, n]$) 时, x_k 增加一个单位时,发生比将会增加 e^{β_k} 倍 ($P_{CR} = e^{\beta_k}$),此时, $e^{\beta_k} > 1$,这说明 x_k 的增加会导致事件发生的概率增加,即 x_k 对事件发生这一类别而言为“危险”因素。当 $\beta_k < 0$ 时,若 x_k 增加一个单位时,发生比将会减少 e^{β_k} 倍,此时, $e^{\beta_k} < 1$,这说明 x_k 的增加会导致事件发生的概率减小,即 x_k 对事件发生这一类别而言为“保健”因素。当 $\beta_k = 0$ 时, x_k 无论如何变化,都不会引起发生比的变化, x_k 这一变量对事件的发

生不造成任何影响。此外, x_k 的 $|P_{CR} - 1|$, 即 $|e^{\beta_k} - 1|$ 的值越大, 代表第 k 个特征对发生比的影响程度越大。假设有一事件为是否有恶性肿瘤, 为避免恶性肿瘤的产生, 则人体应多增加“保健”因素的作用, 而适当减少“危险”因素的作用。

2 相对概率变化比

2.1 相对概率变化比的概念及其意义

Softmax 回归是一种广义的逻辑回归模型, 在 Softmax 回归中, 参数 $\theta_j = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jn})$ 具有与 LR 模型中参数类似的意义。为更好地解释 Softmax 回归中参数 θ_j 的意义, 相对概率变化比这一概念被引入到本文中。举例来说, 对于给定的样本 x 被分类为第 d ($d \in [1, 2, \dots, m]$) 类的概率计算如式 (4) 所示, 同样, 样本 x 被分类为第 t ($t \in [1, 2, \dots, m], t \neq d$) 类的概率计算如式 (5) 所示。这两者之间的概率比被认为是类别 d 相对于类别 t 的发生比 P_{OR}^{dt} , 具体计算过程如式 (6) 所示。从式 (6) 可知, 第 k 个自变量 x_k 的变化, 会导致类别 d 相对于类别 t 的发生比发生变化, 这种由于自变量 x_k 的变化而造成的 P_{OR}^{dt} 变化, 在本文中被命名为相对概率变化比。类别 d 与类别 t 之间的相对概率变化比表示为 P_{CR}^{dt} 。当第 k 个自变量 x_k 增加一个单位时, P_{CR}^{dt} 的计算如式 (7) 所示。

$$P(y = d|x; \theta_d, b_d) = \frac{e^{\beta_{d1}x_1 + \beta_{d2}x_2 + \dots + \beta_{dn}x_n + b_d}}{\sum_{j=1}^m e^{\theta_j^T x + b_j}} \quad (4)$$

$$P(y = t|x; \theta_t, b_t) = \frac{e^{\beta_{t1}x_1 + \beta_{t2}x_2 + \dots + \beta_{tm}x_n + b_t}}{\sum_{j=1}^m e^{\theta_j^T x + b_j}} \quad (5)$$

$$P_{OR}^{dt} = \frac{e^{\beta_{d1}x_1 + \beta_{d2}x_2 + \dots + \beta_{dn}x_n + b_d}}{e^{\beta_{t1}x_1 + \beta_{t2}x_2 + \dots + \beta_{tm}x_n + b_t}} \quad (6)$$

$$P_{CR}^{dt} = \frac{e^{\beta_{dk}}}{e^{\beta_{tk}}} \quad (7)$$

若 $P_{CR}^{dt} = 1$, 表示 x_k 的变化并不会引起类别 d 相对于类别 t 的发生比 (P_{OR}^{dt}) 变化, 可解释为 x_k 这一特征不是类别 d 与类别 t 的主要区别因素。若 $P_{CR}^{dt} > 1$, 表示 x_k 的增加将会引起类别 d 的发生概率相对于类别 t 的发生概率有所增加。相对于类别 t , x_k 这一特征对类别 d 来说是“危险”因素; 而相对于类别 d , x_k 这一特征对类别 t 来说是“保健”因素。反之, 若 $P_{CR}^{dt} < 1$, 表示 x_k 的增加, 将会引起类别 d 的发生概率相对于类别 t 的发生概率有所减少。相对

于类别 t , x_k 这一特征对类别 d 来说是“保健”因素; 而相对于类别 d , x_k 这一特征对类别 t 来说是“危险”因素。假设当前的样本 x 属于类别 t , 若需要调整样本 x 中的特征, 使调整后的样本 x 相比于类别 t 有更大的可能被划分为类别 d , 那么此时可增加类别 d 中的“危险”因素, 降低类别 t 中的“保健”因素。

本文采用 $|P_{CR}^{dt} - 1|$ 来衡量变量 x_k 对分类结果的影响程度。举例来说, 若增加 x_p ($p \in [1, 2, \dots, m]$) 这一变量, 在类别 d 和类别 t 中, 引起其相对概率变化比为 P_{CR}^{dt1} , 而增加 x_q ($q \in [1, 2, \dots, m], q \neq p$) 这一变量, 在类别 d 和类别 t 中, 引起其相对概率变化比为 P_{CR}^{dt2} 。若 $|P_{CR}^{dt1} - 1| > |P_{CR}^{dt2} - 1|$, 意味着调整 x_p 相比于调整 x_q , 更有助于样本 x 在类别 t 和类别 d 中进行相互转换。

2.2 相对概率变化比在 CNN 超参数优化中的应用

CNN 超参数组合所对应的测试集分类错误率是连续数值, 根据连续数值的范围, 设定每个类别的测试集错误率区间。比如: 将测试集分类错误率处于 $[0, 0.25]$ 这一区间的超参数组合划分为低错误率类别 d ; 处于 $(0.25, 0.5]$ 这一区间的超参数组合则为中错误率类别 t ; 处于 $(0.5, 1]$ 的超参数组合为高错误率类别 u 。 x_k ($k = 1, 2, \dots, n$) 表示其中的一个超参数, n 为超参数的个数总和。根据 Softmax 回归结果, 计算每个超参数的 P_{CR}^{dt} , P_{CR}^{tu} , $|P_{CR}^{dt} - 1|$ 和 $|P_{CR}^{tu} - 1|$ 。

超参数 x_k 所对应的 P_{CR}^{dt} 和 P_{CR}^{tu} 存在 4 种情况。第一种情况为 $P_{CR}^{dt} > 1$, $P_{CR}^{tu} > 1$, 如图 1(a) 所示。在这种情况下, 若增加 x_k , 低错误率类别相比于中错误率类别, 其发生概率会增加, 中错误率类别相比于高错误率类别, 其发生概率也会增加。增加 x_k , 将有助于降低分类错误率, 从高错误率类别往低错误率类别移动。即当超参数组合处于高错误率类别和中错误率类别时, 要降低分类错误率, 应增加 x_k 。此外, 若当前类别为低错误率类别, 减小 x_k 将更有助于分类为中错误率类别, 分类错误率升高。因此, 处于低错误率类别时, 按照图 1(a) 中的趋势, 为进一步降低分类错误率, 应增加 x_k 。

超参数 x_k 的第二种情况为 $P_{CR}^{dt} < 1$, $P_{CR}^{tu} > 1$, 如图 1(b) 所示。在这种情况下, 当分类错误率处于高错误率类别时, 若增加 x_k , 中错误率类别相比于高错误率类别的发生概率要增加, 中错误率类别相比于低错误率类别的发生概率要增加, 即增加 x_k , 将有助于超参数组合从高错误率类别往中错误率类别移动。当超参数组合处于中错误率类别时, 若继续增加 x_k , 其超参数组合仍最大可能被分

类为中错误率类别，不会朝着更低的分类错误率类别(低错误率类别)移动。因此，若要进一步降低分类错误率，应尝试减小 x_k 。而当超参数组合处于低错误率类别时，若增加 x_k ，有更大的可能性提高分类错误率，被分类为中错误率类别。因此，在低错误率类别时，若要进一步降低分类错误率，应该选择减小 x_k 。

超参数 x_k 的第三种情况为 $P_{CR}^{dl} < 1, P_{CR}^{tu} < 1$ ，如图1(c)所示。在这种情况下，若减小 x_k ，低错误率类别相比于中错误率类别的发生概率会增加，中错误率类别相比于高错误率类别的发生概率会增加。减小 x_k ，将有助于超参数组合从高错误率类别往低错误率类别移动。当超参数组合处于高错误率类别和中错误率类别时，为降低分类错误率，应减小 x_k 。若当前类别已为低错误率类别，增加 x_k ，将有更大的可能性被分类为中错误率类别，分类错误率升高，因此若已处于低错误率类别时，为进一步降低分类错误率，应尝试减小 x_k 。

超参数 x_k 的第四种情况为 $P_{CR}^{dl} > 1, P_{CR}^{tu} < 1$ ，如图1(d)所示。在这种情况下，当超参数组合处于高错误率类别时，减小 x_k ，中错误率类别的发生概率相比于高错误率类别的发生概率会增加，中错误率类别的发生概率相比于低错误率类别的发生概率也会增加。因此，当处于高错误率类别时，应减小 x_k ，使超参数组合更符合中错误率类别的特征，初步降低分类错误率。当处于中错误率类别时，继续减小 x_k ，其超参数组合最大可能被分类为中错误率类别，不会朝着更低的分类错误率类别(低

错误率类别)移动。因此，若要进一步降低分类错误率，应尝试增加 x_k 。而当超参数组合处于低错误率类别时，若减小 x_k ，有更大可能提高分类错误率，被分类为中错误率类别。因此，在低错误率类别时，若要进一步降低分类错误率，应该选择增加 x_k 。

综上4种情况分析可发现，若当前超参数组合为高错误率类别，若想通过调整超参数 x_k 进一步降低分类错误率， x_k 的调整方向可先依据中错误率类别和高错误率类别的相对概率变化比 P_{CR}^{tu} 来决定。若 $P_{CR}^{tu} < 1$ ，则减小 x_k ，若 $P_{CR}^{tu} > 1$ ，则增加 x_k 。当超参数组合为中错误率类别时，若要通过调整 x_k 来进一步降低分类错误率，调整方向可依据 P_{CR}^{dl} 来决定。若 $P_{CR}^{dl} > 1$ ，则应增加 x_k ，若 $P_{CR}^{dl} < 1$ ，则应减小 x_k 。同样，若已经处于低错误率类别，要进一步通过调整 x_k 来降低分类错误率，其调整方向仍由 P_{CR}^{dl} 来决定。当对超参数进行调整时，超参数的调整顺序由 $|P_{CR}^{dl} - 1|$ 和 $|P_{CR}^{tu} - 1|$ 的大小决定，当需调整的超参数组合为高错误率类别时，依据各超参数的 $|P_{CR}^{tu} - 1|$ 值对其进行排序，依据排序顺序对其进行调整。这是因为超参数 x_k 所对应的 $|P_{CR}^{tu} - 1|$ 值越大，对划分中错误率类别和高错误率类别的影响程度越大。同样，当需调整的超参数组合为中错误率类别和低错误率类别时，其超参数的调整顺序应依据 $|P_{CR}^{dl} - 1|$ 的大小来决定。

3 实验设计与实验结果

为探究 CNN 的超参数对图像分类错误率的影响

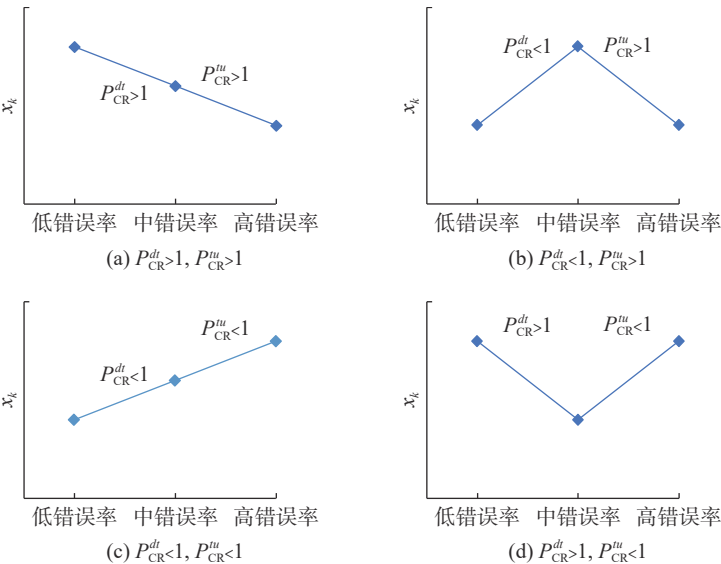


图1 x_k 对分类错误率的影响
Fig.1 Effect of x_k on misclassification rate

响, 本文以 MNIST 数据集随机选取的 600 个数据为研究对象, 其中训练集样本数为 400, 验证集样本数为 100, 测试集样本数为 100。本文基于架构 1 对 CNN 超参数进行了探究, 架构 1 的示意图如图 2 所示, 主要有一层输入层, 一层卷积层, 一层池化层, 两层全连接层, 最后是输出层。

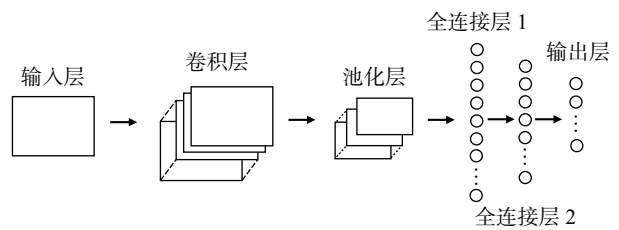


图 2 架构 1 的示意图
Fig.2 Diagram of Architecture 1

在架构 1 下, 本文主要探究了 9 个超参数: 卷积核的高、卷积核的宽、池化层的高、池化层的宽、池化层的步长、卷积核的个数、卷积核的步长、第一层全连接层的神经元个数和第二层全连接层的神经元个数^[3]。前 4 个超参数的取值均为 [2,14] 之间的整数, 卷积核的个数取值为 [4,8] 之间的整数, 卷积核的步长限定为 [1,5] 之间的整数, 池化窗口的步长限定为 [1,7] 之间的整数, 第一层全连接层和第二层全连接层的神经元个数均限定为 [60,140] 之间的整数。

在架构 1 中, 基于各超参数的取值范围, 随机产生了 500 组超参数的组合, 去除不合理架构, 得到合理架构数为 393。每组超参数组合在 MNIST 数据集上的分类错误率易受到初始权重的影响。为得到较为准确的测试集分类错误率, 事先随机产生了 70 个不同的随机种子。每组超参数进行了 70 次实验, 对 70 次实验结果进行排序。排序后对位于中间的 30 次实验结果分别取训练集、验证集, 以及测试集上的平均值, 并将其作为这组超参数下的训练集分类错误率、验证集分类错误率和测试集分类错误率。

3.1 Logistic 回归实验(发生比变化率)

测试集分类错误率是衡量模型训练效果的重要指标, 本文依据每组超参数所产生的测试集分类错误率来对数据进行离散化, 总共划分为两类。当测试集分类错误率小于或者等于 0.25 时, 将被划分为低错误率类别; 当大于 0.25 时, 将被划分为高错误率类别。低错误率类别的样本数为 174, 高错误率类别的样本数为 219。训练集数据和测试集数据将按照 0.75 : 0.25 的比例进行划分。

输入数据均依据式(8)进行标准化。

z = (x - μ) / σ (8)

式中: x 为需要被标准化的原始数据; μ 和 σ 分别为训练数据的均值和方差。

实验通过 python 语言中的 sklearn 包来实现。LR 分析结果如表 1 所示, 其在测试集的分类正确率为 91.91%。

表 1 架构 1 下 LR 参数(MNIST)
Tab.1 Parameters of LR in Architecture 1 (MNIST)

变量	β	P_{CR}	$ P_{CR} - 1 $
x_1	-0.4813	0.6179	0.3821
x_2	0.0502	1.0515	0.0515
x_3	-0.1036	0.9015	0.0985
x_4	3.9565	52.2749	51.2749
x_5	1.1926	3.2957	2.2957
x_6	0.2127	1.2371	0.2371
x_7	1.6700	5.3126	4.3126
x_8	-0.1551	0.8562	0.1438
x_9	-0.1364	0.8724	0.1276

为验证 LR 中的 P_{CR} 是否有助于优化 CNN 超参数, 基于 LR 的 P_{CR} 对 393 组中最差的超参数组合进行调参, 测试集分类错误率从 77.93% 降至 51.50%, 下降幅度为 26.43%。同时, 为找到一个更优的 CNN 超参数组合, 对 393 组超参数实验数据中最优的超参数组合进行调参。但根据 LR 的 P_{CR} 对其进行调整后, 测试集分类错误率没有降低, 初步表明 LR 的发生比变化率未能较好地反映 CNN 超参数对分类错误率的影响。为进一步探究更优的超参数优化方法, 本文后续基于 Softmax 回归得到了相对概率变化比, 利用相对概率变化比来优化超参数。

3.2 Softmax 回归实验(相对概率变化比)

在 Softmax 回归中, 分类错误率被划分为 3 个等级: 当测试集分类错误率处于 [0,0.25] 之间时, 被分类为低错误率类别 d ; 当处于 (0.25,0.5] 时, 被分类为中错误率类别 t ; 当大于 0.5 时, 被分类为高错误率类别 u 。在架构 1 中被分类为低错误率类别的超参数组合数为 174, 中错误率类别的为 147, 高错误率类别的为 72。

为提高 Softmax 回归的分类正确率, 本文采用了单层感知机(SLP)来对数据进行分类, 主要通过调用 python 语言中的 keras 包来编程实现。在本实

验中, SLP 包含输入层和输出层, 输出层的激活函数为“Softmax”, 损失函数为交叉熵损失函数, 优化器为 Adam。由于迭代次数对模型的训练有较大的影响, 当设置较大时, SLP 会出现过拟合的现象, 而设置较小会导致 SLP 的训练不足。为找到较优的迭代次数, 在迭代次数设置为 500, 600 和 700 的情况下进行实验。在迭代次数为 500 时, 其测试集上的分类正确率为 80.39%; 迭代次数增加到 600 时, 测试集的分类正确率为 87.45%; 而继续增加到 700 时, 测试集的分类正确率下降到 82.35%。因此, 在 SLP 中, 较为合适的迭代次数为 600。又因为批处理大小 BatchSize 会对模型的泛化性能造成一定的影响, 为找到较为合适的 BatchSize, 探讨了 BatchSize 设置为 16, 32 和 64 时, 测试集分类正确率的变化。当 BatchSize 为 16 时, 测试集分类正确率为 81.56%; 增加到 32 时, 测试集分类正确率提高到 87.45%; 而继续增加到 64 时, 测试集分类正确率并未进一步提高, 下降到了 76.86%。因此, SLP 中较为合适的 BatchSize 为 32。此外, 通过 SLP 得到的架构 1 中各超参数的相对概率变化比如表 2 所示。

表 2 架构 1 下各超参数的相对概率变化比 (MNIST)

Tab.2 Change ratios of relative probability of each hyper-parameter in Architecture 1 (MNIST)

变量	P_{CR}^{dt}	P_{CR}^{tu}	$ P_{CR}^{dt}-1 $	$ P_{CR}^{tu}-1 $
x_1	1.625 5	1.515 1	0.625 5	0.515 1
x_2	0.952 1	2.342 8	0.047 9	1.342 8
x_3	1.187 4	1.368 2	0.187 4	0.368 2
x_4	0.032 5	0.113 4	0.967 5	0.886 6
x_5	0.508 1	0.295 7	0.491 9	0.704 3
x_6	0.742 4	0.998 9	0.257 6	0.001 1
x_7	0.298 8	0.438 6	0.701 2	0.561 4
x_8	0.952 5	1.327 8	0.047 5	0.327 8
x_9	1.420 7	0.976 8	0.420 7	0.023 2

为探究 Softmax 回归中的相对概率变化比是否有助于 CNN 的超参数优化, 基于相对概率变化比对 393 组中最差的超参数组合进行了调整, 测试集分类错误率由 77.93% 下降到 13.53%, 下降幅度高达 64.40%, 这一下降幅度大于基于 LR 的发生比变化率进行调整的下降幅度 (26.43%)。为进一步找到更优的超参数组合, 降低分类错误率, 对 393 组实验数据中最优的超参数组合基于相对概率变化比进行调整, 找到了更优的 CNN 超

参数组合, 其测试集分类错误率降至 8.73%, 相比最初的 9.90%, 降低了 1.17%, 而基于 LR 的发生比变化率对超参数进行调整后, 未降低分类错误率。这些数据表明, Softmax 回归的相对概率变化比相比于 LR 中的发生比变化率更有助于找到一个更优的超参数组合, 降低模型的分类错误率。

3.3 实验结果分析

3.3.1 架构对实验结果的影响

在上述实验中, CNN 的超参数优化均基于架构 1 进行, 为排除网络架构对实验结果的影响, 本文在架构 2 下基于 MNIST 数据集也进行了实验。架构 2 中有两层卷积层, 两层池化层, 主要有 16 个超参数, 各超参数在本文中的命名及其取值范围如表 3 所示。

表 3 架构 2 中的超参数及其取值范围

Tab.3 Hyper-parameters in Architecture 2 and their value ranges

超参数	取值范围	超参数	取值范围
卷积核1的个数	[4, 8]	卷积核2的高	[2, 4]
卷积核1的高	[2, 5]	卷积核2的宽	[2, 4]
卷积核1的宽	[2, 5]	卷积核2的步长	[1, 2]
卷积核1的步长	[1, 2]	池化窗口2的高	[2, 4]
池化窗口1的高	[2, 5]	池化窗口2的宽	[2, 4]
池化窗口1的宽	[2, 5]	池化窗口2的步长	[1, 2]
池化窗口1的步长	[1, 2]	第一层全连接层的神经元个数	[60, 140]
卷积核2的个数	[12, 20]	第二层全连接层的神经元个数	[60, 140]

基于架构 2, 随机产生了 500 组超参数组合进行实验, 去除掉不合理的超参数组合后为 280 组超参数组合。对这 280 组超参数进行 Softmax 回归分析和 LR 分析, 各参数的相对概率变化比如表 4 所示, 发生比变化率如表 5 所示。

根据相对概率变化比对 280 组超参数组合中的最差超参数组合进行调参, 可将测试集分类错误率由原来的 49.80% 下降到 9.9%, 下降幅度为 39.90%。而根据发生比变化率进行调参, 仅可将测试集分类错误率由原来的 49.80% 下降到 48.36%, 其下降幅度为 1.44%, 远小于基于相对概率变化比的下降幅度 39.90%。此外, 依据相对概率变化比对 280 组超参数组合中的最优超参数组合进行调参, 可将测试集分类错误率由原来的 8.03% 下降到 6.53%, 而依据发生比变化率对其进行调参, 仅下降到 6.70%。上述数据表明, 相比于 LR 的发生比变化率, 基于 Softmax 回归的相对概率变化比在不

同架构下对 CNN 的超参数进行优化, 所得结果更优。

表 4 架构 2 下各超参数的相对概率变化比 (MNIST)

Tab.4 Change ratios of relative probability of each hyper-parameter in Architecture 2 (MNIST)

变量	P_{CR}^{dt}	P_{CR}^{tu}	$ P_{CR}^{dt} - 1 $	$ P_{CR}^{tu} - 1 $
x_1	1.7297	1.1120	0.7297	0.1120
x_2	0.6857	2.1394	0.3143	1.1394
x_3	1.1448	0.6394	0.1448	0.3606
x_4	0.0537	0.1610	0.9463	0.8390
x_5	0.6949	0.4837	0.3051	0.5163
x_6	1.1060	0.7672	0.1060	0.2328
x_7	0.1731	0.5413	0.8269	0.4587
x_8	1.3939	1.0194	0.3939	0.0194
x_9	0.7434	2.0661	0.2566	1.0661
x_{10}	0.9614	1.3716	0.0386	0.3716
x_{11}	0.2499	0.4786	0.7501	0.5214
x_{12}	0.8238	0.6592	0.1762	0.3408
x_{13}	1.0816	0.5677	0.0816	0.4323
x_{14}	0.9203	0.4795	0.0797	0.5205
x_{15}	0.8047	1.2206	0.1953	0.2206
x_{16}	1.4637	0.9372	0.4637	0.0628

表 5 架构 2 下 LR 参数 (MNIST)

Tab.5 Parameters of LR in Architecture 2 (MNIST)

变量	β	e^β	$ e^\beta - 1 $
x_1	-0.2689	0.7641	0.2359
x_2	-0.0102	0.9898	0.0102
x_3	-0.2131	0.8080	0.1920
x_4	3.7225	41.3708	40.3708
x_5	0.3304	1.3915	0.3915
x_6	-0.0991	0.9056	0.0944
x_7	2.2307	9.3065	8.3065
x_8	-0.1603	0.8518	0.1482
x_9	-0.0020	0.9979	0.0021
x_{10}	-0.2848	0.7521	0.2479
x_{11}	1.9949	7.3516	6.3516
x_{12}	1.3146	3.7233	2.7233
x_{13}	0.4616	1.5867	0.5867
x_{14}	0.1240	1.1320	0.1320
x_{15}	0.2183	1.2439	0.2439
x_{16}	-0.0564	0.9451	0.0549

3.3.2 数据集对实验结果的影响

上述实验均在 MNIST 数据集下进行, 为进一步排除数据集对实验结果的干扰, 本文也在 CIFAR-10 数据集上进行了实验。基于架构 1, 随机产生了 500 组超参数组合, 去除掉不合理的组合后为 393 组超参数组合。对上述合理的 393 组超参数进行 Softmax 回归分析和 LR 分析, 各参数的相对概率变化比如表 6 所示, 发生比变化率情况如表 7 所示。根据表 6 的结果对 393 组超参数组合中的最差超参数组合进行调参, 可将其测试集分类错误率由原来的 80.36% 下降到 61.43%。根据表 7 中的发生比变化率进行调参, 其测试集错误率仅下降到 74.9%。为找到一个更优的 CNN 超参数组合, 根据表 6 的相对概率变化比对 393 组超参数组合中的最优超参数组合进行调参, 可将其测试集分类错误率由原来的 61.53% 下降到 58.16%。

表 6 架构 1 下各超参数的相对概率变化比 (CIFAR-10)

Tab.6 Change ratios of relative probability of each hyper-parameter in Architecture 1 (CIFAR-10)

变量	P_{CR}^{dt}	P_{CR}^{tu}	$ P_{CR}^{dt} - 1 $	$ P_{CR}^{tu} - 1 $
x_1	2.3348	1.6788	1.3348	0.6788
x_2	0.7164	0.7058	0.2836	0.2942
x_3	0.7465	1.5468	0.2535	0.5468
x_4	0.1893	0.0535	0.8107	0.9465
x_5	0.7774	0.2573	0.2226	0.7427
x_6	1.5895	1.5599	0.5895	0.5599
x_7	0.4425	0.5391	0.5575	0.4609
x_8	0.8971	1.4369	0.1029	0.4369
x_9	0.9658	0.6677	0.0342	0.3323

表 7 架构 1 下 LR 参数 (CIFAR-10)

Tab.7 Parameters of LR in Architecture 1 (CIFAR-10)

变量	β	e^β	$ e^\beta - 1 $
x_1	-0.7336	0.4801	0.5199
x_2	0.3895	1.4763	0.4763
x_3	0.4056	1.5002	0.5002
x_4	1.6891	5.4148	4.4148
x_5	0.4135	1.5122	0.5122
x_6	-0.3754	0.6869	0.3131
x_7	0.7146	2.0434	1.0434
x_8	-0.1386	0.8705	0.1295
x_9	0.1039	1.1095	0.1095

而根据表 7 的发生比变化率对其进行调参,并不能降低测试集的分类错误率。这些数据表明,基于 Softmax 回归的相对概率变化比在不同数据集下对 CNN 的超参数进行调参,其结果也要优于 LR 中的发生比变化率的调参结果。

4 结 论

卷积神经网络在图像识别领域已经取得了较大的成就,超参数组合对识别错误率有较大的影响,但要找到合适的 CNN 超参数组合却是一个时间成本较高的过程。为节约时间,优化 CNN 超参数组合,本文提出了基于 Softmax 回归的相对概率变化比来明确 CNN 各超参数对分类错误率的影响程度以及各超参数的调整方向。为验证该方法的有效性,在架构 1 下,运用 Softmax 回归和逻辑回归(LR)对 MNIST 数据集的 393 组 CNN 超参数组合进行了分析,并依据 Softmax 回归的相对概率变化比和 LR 中的发生比变化率分别对 393 组超参数组合中的最差和最优的超参数组合进行了调整。结果表明,Softmax 回归的相对概率变化比相比于 LR 中的发生比变化率更能反映 CNN 超参数对分类错误率的影响,且更有助于找到一个更优的超参数组合,降低分类错误率。为排除最初的架构对这一结论的影响,在架构 2 下,针对 MNIST 数据集也进行了实验。实验结果表明,基于 Softmax 回归的相对概率变化比在不同架构下对 CNN 的超参数进行调整,其调整结果仍要优于 LR 的调整结果。为进一步排除数据集对这一结论的影响,在 CIFAR-10 上进行了实验,实验结果表明,在 CIFAR-10 上,基于 Softmax 回归的相对概率变化比对超参数进行调整的结果仍要优于 LR 的发生比变化率的调整结果。

参考文献:

- [1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 1–14.
- [2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84–90.
- [3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580–587.
- [4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779–788.
- [5] SHIH P C, LI H, LIN X, et al. Application of metaheuristic algorithms for determining the structure of a convolutional neural network with a small dataset[C]//Proceedings of the 3rd International Conference on Computer Science and Application Engineering. Sanya: ACM, 2019: 87.
- [6] BIBAIEVA V. Using metaheuristics for hyper-parameter optimization of convolutional neural networks[C]//Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing. Aalborg: IEEE, 2018: 1–6.
- [7] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 818–833.
- [8] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization[J]. *The Journal of Machine Learning Research*, 2012, 13(1): 281–305.
- [9] SNOEK J, LAROCHELLE H, ADAMSANESSA R P. Practical Bayesian optimization of machine learning algorithms[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2012: 2951–2959.
- [10] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. *Handbook of Systemic Autoimmune Diseases*, 2009, 1(4): 1–58.
- [11] CORNFIELD J, GORDON T, SMITH W W. Quantal response curves for experimentally uncontrolled variables[J]. *Bulletin of the International Statistical Institute*, 1961, 38(3): 97–115.
- [12] DOMÍNGUEZ-ALMENDROS S, BENÍTEZ-PAREJO N, GONZALEZ-RAMIREZ A R. Logistic regression models[J]. *Allergologia et Immunopathologia*, 2011, 39(5): 295–305.
- [13] PHILLIPS J, CRIPPS E, LAU J W, et al. Classifying machinery condition using oil samples and binary logistic regression[J]. *Mechanical Systems and Signal Processing*, 2015, (60/61): 316–325.

(编辑: 丁红艺)