基于 $m\times2$ 正则化交叉验证的神经网络超参数调优方法

曹学飞[1],杨 帆[1],李济洪[2],王瑞波[2],牛 倩[1]
(1. 山西大学 自动化与软件学院,山西 太原 030006;
2. 山西大学 现代教育技术学院,山西 太原 030006)

摘 要:超参数调优是神经网络建模的关键问题。针对传统的超参数调优方法存在的问题,该文提出了一种基于 $m\times2$ 正则化交叉验证的超参数调优方法。目的是给出一种适用于复杂模型、大数据集背景下的计算开销较小且稳健的超参数调优方法。该方法的思想是从完整的数据集上选取少部分数据进行调优,避免模型在数据集较大时非常耗时的超参数调优难题;在 $m\times2$ 交叉验证的基础上设置正则化条件均衡训练集与验证集之间的分布差异,从而减少分布不一致带来的性能波动;使用信噪比作为调优的优化目标,从而可以综合考虑模型性能评价指标的均值和方差;并采用正交设计选择相关性较低的超参数组合以提高调优效率。以命名实体任务为例进行实验,在 CoNLL 2003 数据集上的实验结果显示,提出的调优方法能够选到和网格搜索性能上没有显著差异的超参数组合,且调优时间可显著降低约 66%。

关键词:$m\times2$ 交叉验证;正则化;神经网络;超参数调优;信噪比
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2024)04-0168-06
doi:10.20165/j.cnki.ISSN1673-629X.2024.0025

A Method for Hyper-parameter Tuning of Neural Network Based on $m\times2$ Regularized Cross-validation

CAO Xue-fei[1], YANG Fan[1], LI Ji-hong[2], WANG Rui-bo[2], NIU Qian[1]
(1. School of Automation and Software Engineering, Shanxi University, Taiyuan 030006, China;
2. School of Modern Educational Technology, Shanxi University, Taiyuan 030006, China)

Abstract: Hyper-parameter tuning is a key issue in neural network modeling. From the viewpoint of the problems of traditional hyper-parameter tuning methods, we propose a hyper-parameter tuning method based on $m\times2$ regularized cross-validation. The goal is to present a robust hyper-parameter tuning method with low computational cost suitable for complex models and large datasets. The idea of the proposed method is to select a small number of data from the complete dataset for tuning, so as to avoid the time-consuming problem of hyper-parameter tuning when the dataset is large. Then, on the basis of $m\times2$ cross-validation, regularization is adopted to balance the distribution difference between the training set and the validation set to reduce the performance fluctuation caused by the distribution inconsistency. The signal-to-noise ratio is used as the metric of hyper-parameter tuning, so that the mean and variance of the model performance can be comprehensively considered. The orthogonal design is used to select a combination of hyper-parameters with low correlation to improve the tuning efficiency. The experimental results on the CoNLL 2003 dataset show that the proposed method can obtain a combination of hyper-parameters that is not significantly different from the grid search, and the tuning time can be significantly reduced by about 66%.

Key words: $m\times2$ cross-validation; regularization; neural network; hyper-parameter tuning; signal-to-noise

0 引 言

近年来,神经网络模型广泛应用在自然语言处理(Natural Language Processing, NLP)的各项任务中,特别是在一些序列标注任务上取得了很好的效果,如语义角色标注[1]、命名实体识别[2]等。然而,神经网络模型建模时也存在一个问题:模型的性能在很大程度上依赖于超参数的配置[3-5]。即模型的输出结果对超参数的设置较为敏感,这就导致模型性能可能不稳定[6]。

收稿日期:2023-06-28 **修回日期**:2023-10-30
基金项目:国家自然科学基金(61806115,62076156)
作者简介:曹学飞(1981-),男,博士,CCF 会员(A2717M),研究方向为自然语言处理。

因此,超参数的调优就成为神经网络模型的一个重要研究问题。

目前,超参数的调优主要有网格搜索[7]、贪心搜索[8]和随机搜索[8]等方法。网格搜索通过遍历所有可能的超参数组合,可找到最优性能的组合。该方法的缺点是计算代价过高,因此,相比网格搜索,实践中更多使用贪心搜索。贪心搜索采用的是分布调优、局部最优的思想,通过逐个设置超参数的取值去尝试提升模型性能,即每次选择一个超参数调优,其它超参数固定或已经调至最优,如此下去直到所有的超参数调优完毕。贪心搜索的缺点是可能会得到局部最优而不是全局最优。随机搜索则不尝试所有可能的超参数组合,而是首先为每个超参数定义一个分布,然后从指定的分布中采样固定数量的超参数设置进行调优,随机搜索比网格搜索的效率要高,但是不能保证一定能找到比较好的超参数组合[8],而且,随机搜索也不能根据上一步的调优结果来调整其下一步的行为,这意味着选择不当的超参数,反而会阻碍模型的有效学习。

使用上述方法进行模型的超参数调优,通常的做法是将数据集切分为训练集、验证集和测试集,将每一种不同的超参数组合看作一个独立的模型,在训练集上进行训练,在测试集上进行测试,检验某一种超参数组合下的模型的性能评价指标是否提高[9]。这种做法存在两个问题:其一,随机地将数据集切分为训练集、验证集和测试集,预测标记的分布可能存在较大差异,导致实验结果波动大。而且,简单的在测试集上对性能评价指标的大小进行比较,而没有对比性能评价指标的方差,或没有进行显著性检验,这样的结果往往不可靠,不具有统计意义上的科学性[10-13]。因此,在超参数调优时,应综合考虑性能评价指标的均值和方差去判别模型优劣,使得选择得到的超参数组合较为稳健。其二,当数据集较大时,超参数调优所需的计算开销也非常大,甚至在一般的计算资源下无法计算,此时,要得到方差的估计是非常困难的,进而导致后续的显著性检验无法进行。

鉴于此,该文提出了一种基于 $m\times2$ 正则化交叉验证的神经网络模型的超参数调优方法。该方法的主要思想是:

(1)从完整的数据集上选取少部分数据进行调优,避免模型复杂或数据集较大时非常耗时的超参数调优难题。

(2)在选取的小数据集上采用 $m\times2$ 交叉验证对每组超参数组合进行性能评估,$m\times2$ 交叉验证可以避免小数据集上数据不充分和分布不平衡带来的性能波动。

(3)以实验设计中的信噪比为优化目标进行调优,综合考虑模型性能评价指标的均值和方差,提高调优结果的稳健性。

(4)考虑到对所有可能的超参数组合进行完全实验所需的计算量较大,特别是 $m\times2$ 交叉验证对每组超参数组合都需要进行 $2m$ 次实验,引入正交设计选择相关性较低的超参数组合进行调优,从而提高调优效率。在这一点上,该文的思路与文献[14]一致,都是希望通过正交设计提高调优效率,与文献[14]不同的是,该文进一步设计正则化条件来均衡数据的分布。

该文的目的是给出一种适用于大模型、大数据集背景下的计算开销较小且稳健的超参数调优方法,且该调优方法得到的最优超参数组合在完整的数据集上应与网格搜索的结果相当或没有显著差异,实验结果也验证了这一点。以命名实体识别(Named Entity Recognition,NER)[15]任务为例,在基于 LSTM[16]神经网络上的实验结果显示,该调优方法可以得到与网格搜索性能相当的 F1 值,且计算开销显著降低了约 66%。

1　$m\times2$ 正则化交叉验证的调优方法

1.1　$m\times2$ 正则化交叉验证

将数据集做 m 次随机切分,实施 m 次 2 折交叉验证,称为 $m\times2$ 交叉验证[17-18]。但是依照随机切分来构建 $m\times2$ 交叉验证,容易导致训练集和验证集分布差异过大,增大了实验结果的方差,产生不可靠的结论[19-20]。因此,该文在 $m\times2$ 交叉验证的基础上,设计正则化条件约束训练集、验证集上数据分布的差异,使得切分得到的训练集、测试集分布更加均衡,称之为 $m\times2$正则化交叉验证。具体来讲:

- $m\times2$ 交叉验证:假定数据集 D 由 n 个样本组成,即 $D=\{d_1,d_2,\cdots,d_n\}$,D 上某次对半切分记为 $\{D^{\text{train}},D^{\text{valid}}\}$,其中 $D^{\text{train}}\cup D^{\text{valid}}=D$,$D^{\text{train}}\cap D^{\text{valid}}=\varnothing$。$m\times2$ 交叉验证的切分集可记为:$P=<S_i,S_i^{\text{T}}>$,其中 $S_i=(D_i^{\text{train}},D_i^{\text{valid}})$,$S_i^{\text{T}}=(D_i^{\text{valid}},D_i^{\text{train}})$,$i=1,2,\cdots,m$。$<S_i,S_i^{\text{T}}>$为一个切分对,$S_i^{\text{T}}$ 为 S_i 的对折切分,某个切分下的 D_i^{train} 称为训练集,D_i^{valid} 为验证集。

- 正则化条件:记 $\text{MF}=\{\text{MF}^{(1)},\text{MF}^{(2)},\cdots,\text{MF}^{(K)}\}$ 为分布差异度量函数集,$\text{MF}^{(k)}=(f_i^{(k)}=g_k(D_i^{\text{train}},D_i^{\text{valid}}),i=1,2,\cdots,m)$ 为长度为 m 的差异度量向量,$g_k(D_i^{\text{train}},D_i^{\text{valid}})$ 为某次切分下训练集 D_i^{train} 和验证集 D_i^{valid} 的分布差异度量函数,$k=1,2,\cdots,K$。在不同的 NLP 任务中,g_k 可以有不同的定义,例如,对于 NER 任务,定义 $g_1(D_i^{\text{train}},D_i^{\text{valid}})$ 表示实体标签的分布在训练集和验证集上的差异,$g_2(D_i^{\text{train}},D_i^{\text{valid}})$ 可以表示训练集和验证集上句子长度分布的差异。将实体标签的分布

差异度量函数作为 $m \times 2$ 交叉验证的正则化条件。

1.2 分布差异度量函数

对于 NLP 任务,预测标签或者特征大多是离散的,因此,该文使用离散随机变量分布一致性检验的卡方统计量来度量训练集和验证集之间的分布差异。设随机变量 L 表示某种实体预测标签,其取值为离散集合 $\{l_1, l_2, \cdots, l_J\}$,则 L 在训练集和验证集上的分布差异,可用如下卡方统计量来度量:

$$\chi^2 = \sum_{j=1}^{J} \frac{n^{\text{train}} (r_j^{\text{train}} - r_j)^2 + n^{\text{valid}} (r_j^{\text{valid}} - r_j)^2}{r_j} \quad (1)$$

式中, n^{train} 和 n^{valid} 为其在训练集及验证集上出现的频次, r_j, r_j^{train} 和 r_j^{valid} 为第 j 种标签值 l_j 在数据集、训练集及验证集上的频率。实验中以单位自由度的差异度量函数 χ^2 / J 来构成 $\text{MF}^{(K)}$,具体实施时,可以经验性地选择 $|f_i^{(k)}| \leqslant 1$。因此, $m \times 2$ 正则化交叉验证的思想就是通过优化数据切分来控制 $|f_i^{(k)}|$ 的大小。

1.3 信噪比

采用 $m \times 2$ 正则化交叉验证可获得每种超参数组合下模型性能评价指标估计的均值和方差,这样就可以引入实验设计中的信噪比作为超参数调优目标来选择最优的超参数组合,从而选择得到的模型更为稳健。信噪比的定义为:

$$\eta = -10\log\left(\frac{\mu^2}{\sigma^2}\right) \quad (2)$$

其中, μ 为性能评价指标的期望, σ^2 为性能评价指标的方差。实验以 F1 值作为模型性能评价指标,由于 F1 值越大表示模型性能越好,因此采用望大特性的信噪比[21],在 $m \times 2$ 交叉验证下,其计算公式为:

$$\eta = -10\log\left(\frac{1}{m}\sum_{i=1}^{m} \text{F1}^{(i)}\right) \quad (3)$$

其中, $\text{F1}^{(i)}$ 为某种超参数组合下第 i 次实验得到的 F1 值。

2 实 验

2.1 任 务

NER 旨在识别出文本中的命名实体并标记为预定义的类别。该文使用 CoNLL 2003 (https://www.clips.uantwerpen.be/conll2003/ner/) 命名实体识别数据集作为实验数据集,实体被定义为四种类型:地名(LOC)、组织机构名(ORG)、人名(PER)和其他(MISC)。

2.2 模 型

图 1 所示的模型包含三部分,Embeddings 层将词映射为向量,经过双向 LSTM(BiLSTM)层的训练学习,由 CRF 层输出带有 B,I,O 的标签序列,最后基于输出的标签序列重构得到句子中的实体。

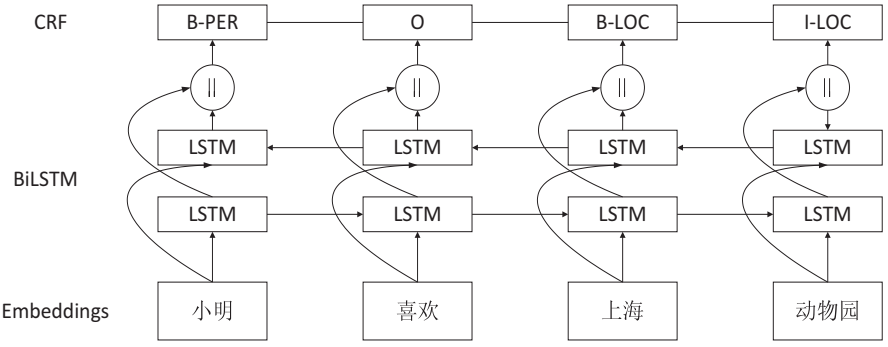

图 1 基于 LSTM 的 NER 模型架构

2.3 实验设置

CoNLL 2003 命名实体识别数据集标准的切分为训练集(14 987 句)、验证集(3 644 句)和测试集(3 486 句)。为了避免大数据集上耗时的调优难题,该文将训练集和验证集合并,从中随机抽取 4 000 句作为超参数调优所用的小数据集,先进行多次对折切分,再按照 1.1 节和 1.2 节介绍的方法选取 $|f_i^{(k)}| \leqslant 1$ 的 m 对切分(实验中 m 取 7)。

超参数调优时以信噪比为调优指标,模型的性能评价指标使用 F1 值。由于 F1 值的取值范围为[0,1],而信噪比是以性能评价指标的估计值服从正态分布为基础的,该文对 F1 值进行了 logit 变换,以使其理论上取值在 $(-\infty, +\infty)$,从而更接近正态性假设。因此,实验中涉及到的均值、方差和信噪比也都是基于 logit 变换后的值计算的。

模型的超参数设置采用了 Yang 等[3]的推荐设置,包括:Embeddings 层采用 GloVe[22]方法预训练的 100 维向量来初始化;使用 mini-batch 的随机梯度下降算法进行训练,batch-size 为 10,初始学习率为 0.015,学习率衰减系数为 0.05;dropout rate[23]设置为 0.5;此外,LSTM 层的节点数为 200,迭代次数为 100。将上述超参数配置作为实验的基线组合,从中选取了 4 个超参数(见表 1)进行调优来说明该方法,并在该超参数的基线值上下扩展取值,构成超参数的调优值

空间。表 1 给出了 4 个超参数调优的取值。

表 1　调优超参数及其取值

超参数	取值	说明
hidden	100, 200, 300	LSTM 层的节点数
lr_init	0.01, 0.015, 0.02	初始学习率
decay	0.01, 0.05, 0.1	学习率衰减系数
epoch	100, 200, 300	迭代次数

3　结果与分析

3.1　基于正交设计选择的超参数组合的调优结果

对表 1 中的 4 个超参数,如果进行网格搜索,需要评估 3^4 组不同的超参数组合,每个组合如果再进行 7×2 次交叉验证的实验,实验的总次数为 1 134($3^4\times14$)。为减少实验次数,提高调优效率,该文采用 $L_{18}(3^7)$ 正交表来选择相关性较低的超参数组合,即从 3^4 组超参数组合中选出了 18 组有代表性的超参数组合来安排实验,实验次数可降低为 252(18×14)。

表 2 给出了通过正交设计选择得到的 18 组超参数配置组合(见表 2 的 1 ~5 列),每组超参数组合按照 7×2 正则化交叉验证进行 14 次实验计算得到信噪比(见表 2 的最后一列)。结果显示,当 LSTM 层的节点数为 300、初始学习率为 0.02、学习率衰减系数为 0.01、迭代次数为 300 时,信噪比达到最大。因此将这组超参数作为文中方法选到的最优超参数组合。

表 2　文中方法得到的调优结果

序号	hidden	lr_init	decay	epoch	信噪比
1	100	0.01	0.01	100	14.28
2	100	0.015	0.05	200	14.38
3	100	0.02	0.1	300	14.39
4	200	0.01	0.01	200	15.01
5	200	0.015	0.05	300	14.84
6	200	0.02	0.1	100	14.42
7	300	0.01	0.05	100	14.20
8	300	0.015	0.1	200	14.62
9	**300**	**0.02**	**0.01**	**300**	**15.05**
10	100	0.01	0.1	300	13.87
11	100	0.015	0.01	100	14.28
12	100	0.02	0.05	200	14.34
13	200	0.01	0.05	300	14.69
14	200	0.015	0.1	100	14.29
15	200	0.02	0.01	200	14.74
16	300	0.01	0.1	200	14.17
17	300	0.015	0.01	300	15.04
18	300	0.02	0.05	100	14.63

3.2　基于全部超参数组合的调优结果

为了进一步和网格搜索进行比较,该文遍历所有的超参数组合(3^4组)进行网格搜索,每组超参数组合同样在小语料上进行 7×2 正则化交叉验证,以信噪比为指标选择最优超参数组合,结果见表 3。

表 3　全部超参数组合的网格搜索结果

hidden	lr_init	decay	epoch	信噪比
300	0.01	0.01	300	15.12

由表 3 结果显示,文中方法选到的最优超参数组合与基于全组合的网格搜索得到的最优超参数组合虽然不同,但二者对应的信噪比基本相当,进一步分析,两种方法得到的超参数组合的 F1 值并没有显著差异(见 3.3 节)。

3.3　显著性检验

将上述两种最优超参数组合以及在 2.3 中给出的基线组合应用到完整的数据集上,即在 CoNLL 2003 命名实体识别数据集标准切分下的训练集上训练模型、测试集上测试,得到的最终性能指标 F1 值见表 4。

表 4　CoNLL 2023 数据集上的测试结果　%

超参数组合	精确率	召回率	F1 值
文中方法	90.53	88.23	89.37
网格搜索	90.54	88.39	89.45
基线组合	90.56	87.31	88.90

与基线组合相比,网格搜索和文中方法调优得到的超参数组合在完整的数据集上测试得到的 F1 值更高。为了进一步验证三种超参数组合得到的 F1 值是否存在显著差异,该文采用了 McNemar 检验[22]进行显著性检验,结果见表 5。

表 5　McNemar 检验结果

不同超参数组合的比较	χ^2 值
文中方法↔基线组合	5.31 *
网格搜索↔基线组合	10.87 *
文中方法↔网格搜索	1.07

注:* 表示有显著差异。

检验结果显示,使用文中方法得到的超参数组合在 NER 任务上的性能与使用基线组合得到的性能有显著差异,χ^2 值为 5.31,大于显著水平为 0.05 时的 χ^2 临界值 3.84。同样,网格搜索得到的超参数组合与基线组合也有显著差异,χ^2 值为 10.87。而文中方法与网格搜索得到的超参数组合没有显著差异,χ^2 值为 1.07。换言之,使用文中方法或者网格搜索得到的超参数组合与基线组合相比,在性能上都有显著提升,而文中方法或者网格搜索二者之间并没有显著差异,这也符合文中方法的出发点:使用小数据集而不是完整的数据集调优,可避开大数据集上调优的耗时难题,同

时在完整的大数据集上还可以得到和网格搜索相当的性能。

3.4 调优效率对比

传统调优方法如网格搜索是在完整数据集上进行调优,模型训练非常耗时。如对表 1 中的 4 个超参数使用网格搜索调优,需要评估 3^4 组不同的超参数组合,在山西大学高性能计算平台的单个计算节点(Intel Xeon CPU E5-2620 2.10 GHz)上耗时约 33 小时,而文中方法在选取的小的数据集上一组超参数配置的调优时间约为 3.6 小时。虽然文中方法需要实施 $m \times 2$ 交叉验证,进行的实验次数较多,但在 24 个计算节点上并行实验的结果显示文中方法的调优时间(46.1 小时)和网格搜索(137.2 小时)相比,降低了约 66%。

3.5 各超参数对模型性能影响的定量分析

现有的超参数调优的工作,很少分析哪些超参数对模型性能至关重要,哪些超参数对模型性能影响较小。该文采用 $m \times 2$ 正则化交叉验证方法进行实验,对每一种超参数组合,能得到 m 组结果,可以计算得到模型性能评价指标估计的均值和方差,可再进一步使用方差分析方法(ANOVA)分析超参数对模型性能的具体影响。

3.5.1 超参数对性能指标 F1 值的均值的影响

以 7×2 正则化交叉验证得到的 F1 值的均值为应变量,以表 1 中列出的 4 个超参数为自变量进行方差分析,结果见表 6。

表 6 超参数对 F1 值均值影响的分析

	自由度	均方	F 值	Pr(>F)	
hidden	2	0.003 1	10.41	0.004 6	**
lr_init	2	0.001 0	3.49	0.075 4	
decay	2	0.003 2	10.74	0.004 1	**
epoch	2	0.001 5	4.91	0.036 2	*
残差	9	0.000 3			

注:** 和 * 分别表示在 0.01 和 0.05 显著水平下显著。

根据表 6 给出的方差分析结果可知,LSTM 层的节点数和学习率衰减系数这两个超参数对 F1 值的均值有非常显著的影响;在 0.05 显著水平下,迭代次数也对 F1 值的均值有显著影响;而初始学习率的设置对 F1 值的均值没有显著影响,即就实验中的超参数设置而言,初始学习率取 0.01,0.015 或 0.02 不会显著地改变 F1 值。

3.5.2 超参数对性能指标 F1 值的方差的影响

以 7×2 正则化交叉验证得到的 F1 值的方差为应变量,以表 1 中列出的 4 个超参数为自变量进行方差分析,结果见表 7。

表 7 超参数对 F1 值方差影响的分析

	自由度	均方	F 值	Pr(>F)	
hidden	2	0.036 2	6.16	0.020 7	*
lr_init	2	0.021 2	3.60	0.070 9	
decay	2	0.003 9	0.66	0.541 1	
epoch	2	0.005 1	0.87	0.449 9	
残差	9	0.005 9			

注:* 表示在 0.05 显著水平下显著。

表 7 的方差分析结果显示,在 0.05 显著水平下,只有 LSTM 层的节点数这一超参数对 F1 值的方差有显著影响。结合表 6 来看,在神经网络模型中,有的超参数同时对性能评价指标的均值和方差有显著影响,这说明在调优过程中,采用信噪比作为调优指标是有其科学依据的。而且通过方差分析可以更好地理解每个超参数对模型的性能影响,更好地指导调优过程。

4 结束语

超参数的调优问题是使用神经网络模型的一个关键问题,传统的方法如网格搜索或随机搜索直接在完整的大数据集上进行调优,计算量大,而且一般以模型性能指标的大小为调优目标,不考虑性能指标的方差,导致模型的稳定性、可复现性较差。该文首先从完整的数据集上选取少部分数据,进而采用均衡 $m \times 2$ 正则化交叉验证以信噪比为调优目标进行超参数调优,同时引入正交设计来提高调优效率。该方法适用于目前大模型、大数据集背景下神经网络超参数调优,该方法能够调优得到和网格搜索性能相当的超参数同时显著降低计算开销。

参考文献:

[1] CAI R, LAPATA M. Syntax-aware semantic role labeling without parsing[J]. Transactions of the Association for Computational Linguistics, 2019, 7(8): 343-356.

[2] WU G, TANG G, WANG Z, et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition[J]. IEEE Access, 2019, 7: 113942-113949.

[3] YANG J, LIANG S, ZHANG Y. Design challenges and misconceptions in neural sequence labeling[J]. arXiv: 1806.04470, 2018.

[4] REIMERS N, GUREVYCH I. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks[J]. arXiv: 1707.06799, 2017.

[5] BERGSTRA J, BARDENET R, BENGIO Y, et al. Algorithms for hyper-parameter optimization[C]//International conference on neural information processing systems. Granada: Curran Associates Inc., 2011: 2546-2554.

[6] ZHENG S, SONG Y, LEUNG T, et al. Improving the robust-

ness of deep neural networks via stability training[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:4480-4488.

[7] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[8] BERGSTRA J,BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research, 2012,13(2):281-305.

[9] REIMERS N,GUREVYCH I. Reporting score distributions makes a difference:performance study of LSTM-networks for sequence tagging[J]. arXiv:1707.09861,2017.

[10] BERG-KIRKPATRICK T,BURKETT D,KLEIN D. An empirical investigation of statistical significance in NLP[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island: Association for Computational Linguistics,2012:995-1005.

[11] 宋毅君,王瑞波,李济洪. 基于条件随机场的汉语框架语义角色自动标注[J]. 中文信息学报,2014,28(3):36-47.

[12] 王瑞波,王　钰,李济洪. 面向文本数据的正则化交叉验证方法[J]. 中文信息学报,2019,33(5):54-65.

[13] 曹学飞,李济洪,王瑞波,等. 基于稳健设计的双向长短期记忆神经网络模型的调优方法[J]. 应用概率统计,2022, 38(3):317-332.

[14] 杜　博. 基于正交设计的神经网络超参数调优方法[D]. 太原:山西大学,2022.

[15] SANG E F, DE MEULDER F. Introduction to the CoNLL-2003 shared task:language-independent named entity recognition[C]//Proceedings of conference on natural language learning at HLT-NAACL. Edmonton:Association for Computational Linguistics,2003:142-147.

[16] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[17] HAFIDI B,MKHADRI A. Repeated half sampling criterion for model selection[J]. The Indian Journal of Statistics,2004,66 (3):566-581.

[18] RODRIGUEZ J D,PEREZ A,LOZANO J A. Sensitivity analysis of k-fold cross validation in prediction error estimation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,32(3):569-575.

[19] GILLICK L,COX S J. Some statistical issues in the comparison of speech recognition algorithms[C]//International conference on acoustics, speech, and signal processing. Glasgow: IEEE,1989:532-535.

[20] OOMMEN T, BAISE L G, VOGEL R M. Sampling bias and class imbalance in maximum-likelihood logistic regression [J]. Mathematical Geosciences,2011,43(10):99-120.

[21] 牛　倩,曹学飞,王瑞波,等. 基于稳健设计的 SGNS 算法的超参数调优方法[J]. 计算机应用研究,2021,38(2):510-516.

[22] PENNINGTON J,SOCHER R,MANNING C D. Glove:global vectors for word representation[C]//Proceedings of the conference on empirical methods in natural language processing. Doha:Association for Computational Linguistics,2014:1532-1543.

[23] 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.