计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):170−174 [doi: 10.15888/j.cnki.csa.007357]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Cotraining-LSTM 空气质量校准算法①

祁柏林[2], 张　欣[1,2], 刘　闽[3], 魏景锋[4], 杜毅明[3], 金继鑫[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](辽宁省沈阳生态环境监测中心, 沈阳 110000)
[4](辽宁省医疗器械检验检测院, 沈阳 110000)
通讯作者: 张　欣, E-mail: zhangxin172@mails.ucas.ac.cn

摘　要: 空气环境问题越发成为人们关注的焦点. 除了工厂排放的各种废气, 私家车的普及都导致了当前令人担忧的空气环境状况. 国家相关部门也开始加大对空气环境的治理, 提出了环境质量网格化监测的相关政策. 在此背景下, 市场涌现出很多微型监测仪器, 但由于自身内部的传感器精准度不够, 存在数据偏差的问题. 为了解决这一问题, 本文通过利用神经网络技术中的长短期记忆网络 (Long Short-Term Memory, LSTM) 模型结合半监督学习方法, 达到提高监测数据的精准度的目的. 通过与其它模型进行对比分析, 该方法达到了一定的效果.

关键词: 半监督学习; 校准; 长短期记忆网络; 传感器; 监测数据

引用格式: 祁柏林,张欣,刘闽,魏景锋,杜毅明,金继鑫.基于 Cotraining-LSTM 空气质量校准算法.计算机系统应用,2020,29(4):170–174. http://www.c-s-a.org.cn/1003-3254/7357.html

Air Quality Calibration Algorithm Based on Catraining-LSTM

QI Bo-Lin[2], ZHANG Xin[1,2], LIU Min[3], WEI Jing-Feng[4], DU Yi-Ming[3], JIN Ji-Xin[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Shenyang Ecological Environment Monitoring Center of Liaoning Province, Shenyang 110000, China)
[4](Liaoning Medical Device Inspection and Testing Institute, Shenyang 110000, China)

Abstract: The problem of air environment has become the focus of attention. Apart from the exhaust emissions from factories, the popularity of private cars has led to worrisome air conditions. Related government agencis have also begun to strengthen the control of air environment, and put forward relevant policies for grid monitoring of environmental quality. In this context, many micro-monitoring instruments have emerged into the market, but due to the inadequate accuracy of internal sensors, there is a problem of data deviation. In order to solve this problem, this study uses the Long Short-Term Memory (LSTM) model of neural network technology and semi-supervised learning method to improve the accuracy of monitoring data. By comparing with other models, this method achieves a sound effect.

Key words: semi-supervised learning; calibration; Long Short-Term Memory (LSTM); sensor; monitoring data

空气环境的好坏和人们的生活密切相关, 越来越多的环境问题也开始引起我们的关注. 早在 70 年代, 日, 美, 西欧国家的环境监测技术已经达到了较高水平, 仪器种类齐全, 但内部器件成本较高. 通过几十年的不断发展, 发达国家的环境保护工业已经进入了成熟阶段. 相比较来说国内监测起步较晚. 之前我国重视工业发展, 忽视了空气质量这方面的问题. 如今空气质量问题提上日程. 2014 年国务院下发的《国务院办公厅关

① 基金项目: 辽宁省“兴辽英才计划”项目 (XLYC1808004)
Foundation item: Talent Program of Revitalizing Liaoning, Liaoning Province (XLYC1808004)
收稿时间: 2019-08-20; 修改时间: 2019-09-09, 2019-10-21; 采用时间: 2019-10-22; csa 在线出版时间: 2020-04-05

于加强环境监管执法的通知》中就提到要全面加强环境监管执法, 并在 2015 年年底落实对环境的网格化监管. 而在国外, 对于空气环境质量的监测早已过渡到了逐点位指标评价, 但是国外的逐点指标评价无法体现各点位的具体参考数据. 所以目前来看, 国家对环境实现全面网格化监管必然是一个艰巨的任务.虽然国家投放的标准站点监测的数据准确, 但是公建费用较高, 监测设备仪器昂贵, 而且在点位布施上不够灵活, 远远达不到当前网格化监管的政策需求, 而且在局部污染以及污染细节监测方面的能力稍显不足[1]. 对应这些问题, 市场上涌现出很多方便布施成本又低廉的监测空气质量污染指数 $PM_{2.5}$, PM_{10}, NO, SO_2, CO, O_3 的微型监测仪器. 在北京, 石家庄等地都有该微型监测仪器的应用. 但是这些监测仪器存在一个共同的缺点: 传感器自身的物理特性导致了监测到的数据存在一定的偏差. 低精准度的传感器在监测污染物浓度时会很容易受到污染物结构形状等一些客观因素的影响. 研究人员尝试用标准气体对仪器进行校准, 但是传感器的物理特性的缺点仍是我们需要解决的一个难题. 中国环境科学研究院副研究员高健表示: 目前各地网格监控取得了很大进步, 下一步需在精细化方面做出突破. 目前国内关于这种微型监测仪器数据校准的技术领域存在很大的缺口. 本文就是从优化微型监测仪器的精确度的角度出发, 提高微型监测仪器的测量精度, 从而使这些微型监测仪器可以更为广泛的为社会服务.

1 研究方法

根据传统校准方法以及数据特征, 本文采用了基于半监督学习的协同训练的长短期记忆网络 (Long Short-Term Memory, LSTM). 半监督学习方式相比于监督学习方式可以解决数据遗弃问题. 监督学习中会丢弃大量的未标定的数据, 这会造成很大一部分的数据损失, 本文利用半监督学习的方法在一定程度上可以避免数据中存在大量未标定的数据而造成数据的浪费的问题, 从而提高数据的利用率. 半监督学习将大量未标记数据利用起来, 避免了以往数据浪费的问题. 通过协同训练, 本文把标记数据 (x,y) 复制成两个完全相同独立的数据集 1, 2, 同时将未标记 (x_μ,y_μ) 的数据集也分成两部分数据集 1, 2, 然后分别利用有标记的数据对 LSTM 模型进行训练, 将训练好的模型应用到未标记的数据集. 未标记的数据在通过相对应的 LSTM 模型预测之后会得到一个相应的结果. 未标记的数据集中的数据经过模型训练后会得到预测的结果, 这样就实现了给未标记数据打上 label 的目的. 接下来对未标记的数据集中的数据逐条分析. 针对未标记数据集中的数据x_μ, 找出x_μ在标记数据中 K-邻近值, 将这些邻近值组合成新的邻近数据集合. 邻近数据集集合中的数据按照 y 与y_μ的差值进行降序处理. 将处理好的邻近数据集合进行置信度的检测, 即: 在该集合中找到一条可以是该模型的损失函数 (loss function) 最小, 则将此条数据加入到标记数据集 2 中 筛选出 K 个组成新的邻近值数据集. 整体的算法流程图如图 1 所示.

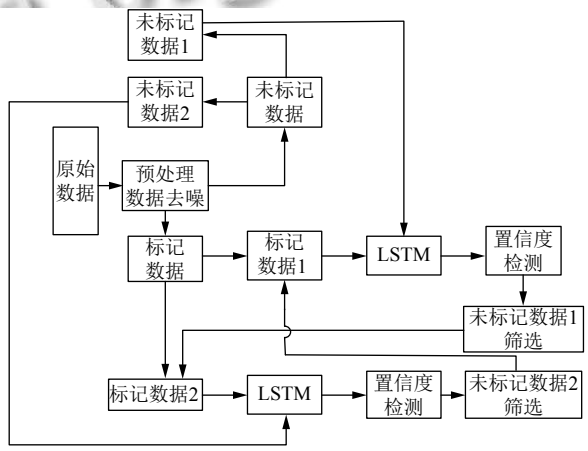

图 1 算法流程图

1.1 数据预处理

本文主要研究的数据对象是小时数据. 来源是国家标准站市控小时数据. 国家标准站的标准数据是我们最终想要学习到的一个结果. 通过学习得到的结果越接近国家标准站的数据就表明实验中模型学习的越好. 本文截取了从四月底到五月底之间的小时数据, 将它们作为训练数据的标签, 为了减少环境因素对仪器的影响, 这里本文选取了和国家市控站仪器处于同一环境的微型监测仪器设备, 并截取相同时间段的数据作为训练的输入数据. 通过分析发现: 数据在监测时也存在大量的噪声数据, 这些噪声数据会导致我们学习模型的好坏, 所以在前期要进行数据的去噪处理.

去噪即去除噪声数据 (异常数据). 微型监测仪器正常监测数据的变化应该是平滑有过渡的, 通过观察部分数据可以发现某些时间段的数据出现大幅度的波动, 出现这种情况可能是仪器设备在进行自我的校准. 所以, 为了保证实验中训练的学习模型能够有更好的

拟合性和泛化能力, 必须要去掉这些噪声数据. 原始数据如图 2 所示.

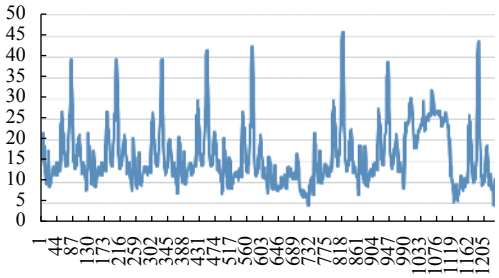

图 2 原始数据

通过图 2 的原始数据可以看出数据有很多高峰值.事实上, 在实际的监测中, 微型监测仪器采集的数据是平稳的, 数据之间的变化也是一个平稳的过渡. 去除噪声数据的方法很多. 比如针对电信号数据, 很多人会采用小波去噪, 傅立叶去噪, 针对分类问题, 会用到聚类等. 本文的数据与前述这些数据略有不同, 它的异常数据不是离散点, 是相邻数据之间存在大的波动. 由于我的数据类型比较单一简单, 太复杂的模型应用反而会达到适得其反的效果. 本文的数据去噪主要是处理那些浮动过大的数据, 避免其造成模型训练的不准确性. 实质上我们处理数据主要考虑相邻数据之间的差异大小, 从而对数据进行一定的处理.

针对监测数据是平滑过渡的特点, 参考小波阈值去噪的阈值思想[2]. 在去除异常数据之前先设定阈值, 然后读取数据进行相邻比较, 一旦两者差值超出设定的阈值, 通过加减阈值将异常数据拉回至正常范围. 本实验数据选择标准就是以第一条数据为基准, 之后数据之间的阈值差不超过之前设定的阈值. 如此不断往后滚动计算, 直至遍历完整个的数据集. 通过该方法处理后的数据相比之前数据的波动得到了一定的缓和. 之间比较尖锐的数据有所下降. 图 3 是处理后的数据与原始数据的对比.

1.2 置信度检测

这里的置信度检测是要在每次训练的时候, 选取一条最符合我们要求的数据. 通过每次对标记数据 1 的训练, 可以得到新的训练模型 1, 我们将每次训练得到的模型 1 用于未标记数据 1 中, 这样未标记数据集 1 中的数据就有了标签即数据形式为 (x_μ,y_μ). 对未标记数据集 1 中的数据x_μ我们在标记数据集中找到x_μ的 K-近邻, 让后选取这些近邻值重新组成新的集合, 记作: Z. 集合 Z 里面数据按照 y 与y_μ的距离差进行排序. 每次从集合中选取一条可以使训练模型的损失函数最小的一条数据, 该条数据就是我们认为的置信度最大的数据, 可以将此条数据加入到标记数据集 2 中. 标记数据集 2 中的数据进行同样的操作将选取的未标记数据集 2 中的数据添加到标记数据集 1 中, 如此交叉进行. 直到最后没有符合置信度要求的数据加入, 此时模型达到稳定状态. 置信度高的的数据加入到训练数据中可以使模型训练的损失函数降低, 模型训练结果会更加准确.

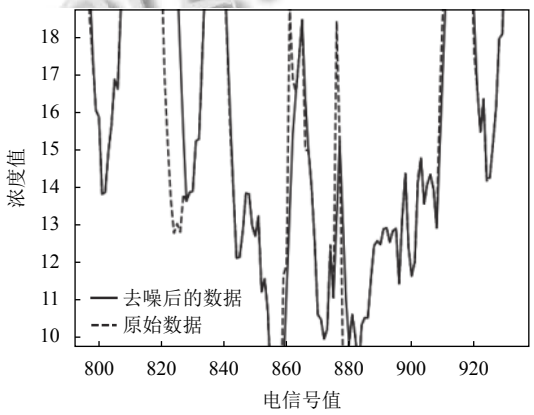

图 3 处理前后数据对比

2 Cotraining-LSTM 整体模型

2.1 模型整体框架

Cotraining-LSTM 是一种结合了协同训练和 LSTM 模型的混合模型. 它的特点在于对数据进行训练时采用了协同训练的方式. 从不同角度对模型各参数进行优化, 同时又充分将未标记的数据利用起来. 在不断交叉训练的过程中, 增加了训练的数据量同时也在不断动态的优化模型[3]. 最终的模型具有一定的泛化能力. 本文研究的课题涉及的数据不存在多维度, 所以过于复杂的模型反而得不到理想的效果, 采用上述协同训练的方式进行训练在运行效率上也有一定的优势.

2.2 协同训练

协同训练 (contrainning) 算法是半监督学习的一种.半监督学习顾名思义, 即可监督可不监督. 它集合二者的优点, 能够充分利用未标记数据和已标记数据来提升学习性能[4]. 协同训练方法采用标记数据分别在两个学习器上进行学习训练, 再利用训练好的学习规则对未标记数据进行预测, 选取置信度较高的数据, 然后

将选中的数据加入已有的标记数据集, 重新对分类器进行训练[5]. 协同训练的方法可以有效利用未标记数据来提高模型精准度. 虽然现在人们处在一个信息化丰富的数据社会, 但是, 想要获得真正能够为我们利用的数据并不容易, 带有标记的数据事实上并不是很多. 如果只用极少的标记数据进行模型的训练, 那么训练出来的模型势必存在准确度不高的问题. 所以, 面对这些不可逆因素, 本篇论文选取了协同训练方法. 它是当前比较流行的一种算法, 利用标记数据进行模型训练, 将训练好的学习规则应用到未标记数据集中, 然后计算未标记数据集中筛选出的邻近数据集的置信度, 将置信度大的数据添加到另一个训练模型的标记数据集中, 不断迭代, 直到训练的模型参数稳定. 协同训练从多角度对模型进行反复训练, 充分利用了已有数据, 在提高了数据利用率的同时对问题的解决也有很大的帮助.

2.3 LSTM 模型

LSTM 最早由 Hochreiter 和 Schmidhuber 在 1997 年提出, 设计初衷是希望能够解决神经网络中的长期依赖问题[6]. LSTM (长短期记忆) 模型是 RNN 的典型代表, 本质上来看是一种 RNN 结构的变形. 图 4 是 RNN 结构展开图, 从图中可以看出从 RNN 可以说在每次输入会结合之前的输出, 相当于拥有了记忆功能. 但是 RNN 不能记忆太久远的信息, 所以会存在一定的梯度消失和梯度爆炸的问题. 这一问题导致了 LSTM 的盛行.

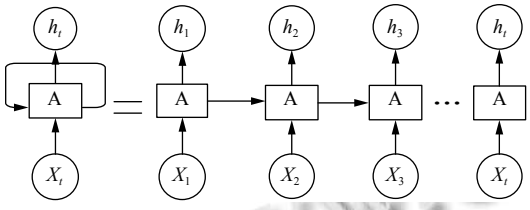

图 4 RNN 结构展开图

如图 5 的 LSTM 结构图可以看出, 它在传统 RNN 的隐藏层各神经单元中增加了记忆单元, 然后通过可控门控制之前信息和当前信息的记忆和遗忘程度, 这样可以让时间序列上的记忆信息可以选择性地保留下来, 从而使 RNN 网络具备了长期记忆功能[7].

LSTM 结构中通过设计两个门来控制记忆单元状态的信息量, 他们分别是遗忘门 (forget gate) 和输入门 (input gate). forget gate 的功能就是”丢弃”. 因为我们不可能将所有信息特征全部记住, 必须有所取舍, forget gate 就实现这一功能. 它决定了上一时刻的单元状态有多少“记忆”可以保留到当前时刻; input gate 决定了当前的时刻输入有多少被保存到单元状态.这两个门都是通过一个权重来决定留下信息的多少. LSTM 在最后设计了一个输出门 (output gate) 来控制单元状态有信息输出. 这三个门的功能特点就是 LSTM 相比传统 RNN 的优势所在.

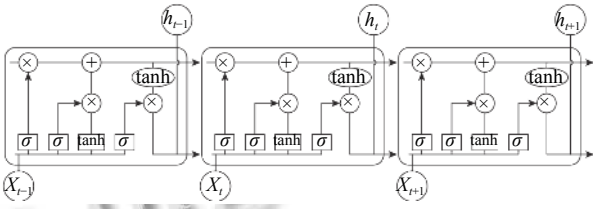

图 5 LSTM 结构图

1) forget gate: 遗忘门是以上一个单元的输出h_{t-1}和本单元的输入X_t为输入的 Sigmoid 函数, 为C_{t-1}中的每一项产生一个在[0,1]内的值, 来控制上一单元状态被遗忘的程度.

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

2) input gate: 输入门结合 tanh 函数来控制有哪些新的数据信息可以被加入. tanh 函数会产生一个新的候选向量$\widetilde{C}_t$, 输入门为$\widetilde{C}_t$中的每一项产生一个在[0, 1]内的值来控制新信息有多少可以加入. 在这之前, 我们得到了 forget gate 的输出f_t, 用来控制上一单元被遗忘的程度, 也有了输入门的输出i_t用来控制新信息被加入的多少, 我们就可以更新本记忆单元的单元状态了.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{2}$$

$$i_t = \sigma(W_i\left[h_{t-1}, x_t\right] + b_i) \tag{3}$$

$$C_t = \tanh(W_c\left[h_{t-1}, x_t\right] + b_c) \tag{4}$$

3) output gate: 输出门的作用是用来控制当前的单元状态有多少被过滤掉. 先将单元状态激活, 输出门为其中每一项产生一个在[0, 1]内的值, 控制单元状态.

被过滤的程度.

$$O_t = \sigma(W_o\left[h_{t-1}, x_t\right] + b_o) \tag{5}$$

$$h_t = O_t * \tanh(C_t) \tag{6}$$

3 实验结果

3.1 评价指标

在这里本文利用均方根误差来衡量测试值和估计值之间的差异程度. 本篇论文研究的目的就是想得到

与标准站数据更接近的数据值. 选择均方根误差在本次实验中更能贴近应用的需求. 均方根误差可以反映观测值与预测值之间的接近程度, 并且对测量数据中差异明显的数据非常敏感. 所以, 均方根误差能够很好地反映出测量的精密度. 一般情况下均方根误差越小越好.

$$RMSE(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2} \quad (7)$$

3.2 实验结果分析

为了能够体现出模型的效果, 验证模型的可行性和有效性, 在进行实验得出结果的同时, 利用同样的数据通过其他模型进行了训练, 对比结果如表 1.

表 1 不同模型训练结果

模型类型	*RMSE*	*LOSS*
LSTM	16.20	0.0678
Regression	30.15	0.1451
Cotraining-LSTM	9.43	0.0478

1) LSTM

单纯只利用 LSTM 进行模型训练, 通过表中的结果可以看出它的均方根差是 16.20, 损失值为 0.0678. 仅仅只利用 LSTM 模型, 造成大量未标记的数据浪费, 事实上参与到模型训练的数据不是很多, 自然模型训练起来在准确度上相比较而言会有一定的偏差.此处在实验室经过多次探索, 利用两层 layer 实现如表中最优的结果.

2) Regression

在已经应用的项目中采用的算法是一元回归. 从实际应用效果来看, 有很多偏差的数据需要人为的手动校准, 必须给数据设定上限值. 这种设置上限值的做法存在一定的主观性, 对此想到利用神经网络进行多元回归分析. 但由于标记数据不是很多, 数据很离散, 多维空间提升了数据复杂度, 反而使其效果不符合我们实际需求.

3) Cotraining-LSTM

基于协同训练的半监督 LSTM 训练模型, 相比较单纯的 LSTM 模型, 我在此基础上加了一个协同训练. 这种半监督的学习方法非常适合那种标记数据比未标记数据少的数据集. 它兼具了 LSTM 模型的记忆功能, 同时基于协同训练的半监督模型可以从多视角上充分利用两种类型的数据进行训练, 可以说是兼顾得更全面. 通过表格数据对比我们可以看出, 从 *RMSE* 这一评价指标可以看出, Cotraining-LSTM 模型训练效果更好. 协同训练模块的加入提升了 LSTM 模型的训练精度.

3.3 实验结果应用

从上述实验结果的分析上来看, 由于我们获得的带标记数据有限并且数据大量离散所以单纯使用 LSTM 模型和 Regression 模型进行训练存在数据利用率不高, 可参与训练数据量不足, 训练结果多维分散, 误差大等问题. Cotraining-LSTM 模型解决了上述方法存在的缺点: 提高了数据利用率, 同时, 使用 Cotraining-LSTM 模型算法校准之后的数据与国家标准数据十分接近. 同时我们将该算法应用到空气质量微型监测实时项目中, 通过对微型站仪器设备的校准所得到的结果数据其误差在应用级范围之内, 在运用此数据进行后续预测等相关操作结果也是一样的.

4 结论

本文针对当前很受大家关心的空气环境问题出发,针对目前市场上关于监测仪器存在由于传感器本身精度不高而存在的测量精度不准确的问题进行了改进, 在基于当前现有的测量数据的情况下, 提出了一种 Cotraining-LSTM (基于协同训练的半监督 LSTM) 模型. 实验结果通过和其他模型的对比可以看出该模型在处理仪器由于自身传感器物理特性而导致监测的结果数据存在偏差这类相关问题上有更好的处理效果. 为今后相似领域的问题解决方案提供一定的参考价值, 并且在此类设备投入生产之后, 将会带来可观的商业价值.

参考文献

1 王春迎, 潘本峰, 吴修祥, 等. 基于大数据分析的大气网格化监测质控技术研究. 中国环境监测, 2016, 32(6): 1–6. [doi: 10.19316/j.issn.1002-6002.2016.06.01]

2 张振凤, 威欢, 谭博文. 一种改进的小波阈值去噪方法. 光通信研究, 2018, (2): 75–78.

3 孟想. 基于群智感知的空气质量传感器校准方法设计与实现[硕士学位论文]. 北京: 北京邮电大学, 2018.

4 蔡毅, 朱秀芳, 孙章丽, 等. 半监督集成学习综述. 计算机科学, 2017, 44(S1): 7–13.

5 邱云飞, 刘聪. 基于协同训练的意图分类优化方法. 现代情报, 2019, 39(5): 57–63, 73. [doi: 10.3969/j.issn.1008-0821.2019.05.008]

6 Wang SH, Zhuo QZ, Yan H, *et al*. A network traffic prediction method based on LSTM. ZTE Communications, 2019, 17(2): 19–25.

7 周志华. 机器学习. 北京: 清华大学出版社, 2016.