Computer Science and Application 计算机科学与应用, 2020, 10(12), 2388-2396
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012253

Hans汉斯

基于DLR模型的PM10－能见度－湿度相关性研究

程茂华[1]，李　敏[2]，马　丹[1]，许双峰[1]

[1]广西科技师范学院，广西　来宾
[2]广西农业科学院农业科技信息研究所，广西　南宁
Email: maohuacheng04130@sina.com

收稿日期：2020年11月28日；录用日期：2020年12月23日；发布日期：2020年12月30日

摘　要

针对当前大数据在单机运算时间过长，对硬件设备要求高的问题，为此提出基于云环境下使用分布式逻辑回归算法DLR (Distributed Logistic Regression)模型对PM10与能见度以及湿度之间的相关性问题，根据二分类思想，将能见度、湿度作为特征值，PM10作为标签值使用逻辑回归算法构建模型对其进行分析，实验结果表明，在同一湿度范围下能见度值越小，大气气溶胶PM10浓度偏大，在同一能见度范围下湿度值越大，大气气溶胶PM10浓度偏低。并且DLR算法模型在时间性能方面要优于传统回归模型，具有更好的鲁棒性以及实时性。

关键词

逻辑回归，云计算，机器学习，PM10，能见度，湿度

Correlation Study of PM10-Visibility-Humidity Based on DLR Model

Maohua Cheng[1], Min Li[2], Dan Ma[1], Shuangfeng Xu[1]

[1]Department of Mathematical and Computer Sciences, Guangxi Science & Technology Normal University, Laibin Guangxi
[2]Institute of Agricultural Science and Technology Information, Guangxi Academy of Agricultural Sciences, Nanning Guangxi
Email: maohuacheng04130@sina.com

Received: Nov. 28th, 2020; accepted: Dec. 23rd, 2020; published: Dec. 30th, 2020

文章引用: 程茂华, 李敏, 马丹, 许双峰. 基于 DLR 模型的 PM10－能见度－湿度相关性研究[J]. 计算机科学与应用, 2020, 10(12): 2388-2396. DOI: 10.12677/csa.2020.1012253

Abstract

Considering the problem that the current big data has a long stand-alone operation time and high requirements for hardware devices, this paper proposes the use of the Logistic Regression (DLR) model in the cloud environment for the correlation between PM10 and visibility and humidity. According to the idea of two classifications, visibility and humidity are used as feature values, and PM10 is used as a tag value to construct a model using a logistic regression algorithm. The experimental results show that under the same humidity range, the smaller the visibility value is, the higher the PM10 concentration of atmospheric aerosol is. The higher the humidity value in the same visibility range, the lower the concentration of PM10 in atmospheric aerosols. And the DLR algorithm model outperforms the traditional regression model in terms of temporal performance, and has better robustness and real-time performance.

Keywords

Logistic Regression, Cloud Computing, Machine Learning, PM10, Visibility, Humidity

1. 引言

在目前的研究中，经常使用数学领域中的统计学方法对湿度、大气能见度与气溶胶 PM10 进行研究，如文献[1]用标准化降水指数作为为干旱指标，分析贵州省年度和季节的干旱发生频率的变化特征；刘艳萍等[2]应用统计学和 Arc GIS 软件分析中国四大工业基地主要城市大气颗粒物污染的时间及空间分布特征，并利用 SPSS 软件对大气污染物 PM10、SO_2、NO_2、CO、O_3 和 PM2.5 的相关性进行分析；文献[3]分析显示 PM10 浓度增大以及颗粒物吸湿性增长可导致能见度数值降低。但这类方法存在人为影响的客观因素。近年来，更多的学者寻找其他的替代方法。20 世纪 80 年代以来人工智能领域兴起，人工神经网络成为研究的热点。石灵芝等人提出基于 BP 人工神经网络[4]的大气颗粒物 PM10 质量浓度预测，建立人工神经网络模型捕捉污染物浓度与气象因素间的非线性影响规律，能够准确预测大气 PM10 质量浓度的实时变化。D. K. Papanastasiou 等人提出使用神经网络[5]和多元回归模型[5]以预测一个中型的地中海城市 PM10 水平，两个模型都能够较好的预测 PM10 实时日均值。

然而，气象领域在处理气象数据的问题上使用的多是投入耗费大的传统方法。对于硬件设备的要求过高，普通的单机运算已经无法满足大数据时代的要求。云计算，一种按量付费模式，融合了分布式计算、并行计算、网络存储等技术的新型产物。能够方便、快速、按需分配的完成客户端的运算任务。云计算的低成本运算快让更多研究者用这种方法解决大数据的运算问题。郑湃等[6]提出云计算环境下面向数据密集型应用的数据布局策略与方法，介绍了一种三阶段数据布局策略以解决跨数据中心数据传输、数据依赖关系和全局负载均衡三个问题，实验分析所提出的方法在时间性能上得到有效提高。王晓燕等[7]提出基于云计算环境中面向 OLTP 应用的数据分布研究，提出以数据分片、数据分片和负载执行为变量对基于计算环境中大数据 OLTP 应用的数据分布问题进行详细归纳分析。张石磊/武装提出一种基于 Hadoop 云计算平台的聚类算法优化的研究[8]，对首先选定初始聚类中心的并行 K-means 聚类算法进行相关实验验证优化后的算法在时间性能上更优。在气象数据方面，云计算技术也有相应的研究。如潘吴斌

提出的基于云计算的并行 K-means 气象数据挖掘研究与应用，并行 K-means 算法可以有效解决分布式数据问题，而大规模数据对运算负荷暴增，为此提出在 Hadoop 平台上实现 K-means 算法的 MapReduce 并行化[9]，实验表明 K-means 聚类效果较好，运行时间较传统方法更优，具有一定的应用意义。如文献[10]基于改进贝叶斯网络的气象数据预测算法研究，提出 Hadoop 平台上使用贝叶斯网络方法对气象数据进行预测，所提出的改进方法与现有的气象预测算法有效地提高了预测精度。以上研究表明，云计算可以有效的解决大规模数据的运算问题，并且可以有效应用在气象领域方面。

为此，本文基于中国气象局气象数据中心和南宁市环保局环境监测站历年气象数据提出了基于云环境下 DLR (Distributed Logistic Regression)模型用于能见度、湿度与气溶胶 PM10 相关性的研究，实验分析表明，湿度区间一致大气气溶胶 PM10 浓度越大能见度就越小，能见度区间一致大气气溶胶 PM10 浓度越低湿度越大。实验结果还发现湿度介于 40%~90%，能见度介于 8~19 km DLR 模型预测效果最好。

2. 逻辑回归定义

在介绍逻辑回归前，我们先简单了解一下线性回归。线性回归的核心思想是：相当于拟合一条直线，设定界限中间值 0.5，归属一类，归属为另一类。线性回归可以用于一些分类问题[11]，但存在的缺点是，线性回归的鲁棒性太差，导致训练的性能较差。

逻辑回归主要用于二分类问题，简单地可以理解为：分类结果要么是，要么是。本实验将 PM10 溶度大于 0.15 mg/m^3 设定为，将 PM10 溶度小于 0.15 mg/m^3 设定为。大气能见度指标、湿度指标为自变量，对应 y 的数学期望如式(1)：

$$E\left(y\,|\,x_i\right)=p\left(y=1\,|\,x_i\right) \tag{1}$$

其中，$p\left(y=1\,|\,x_i\right)$ 表示在 $x=x_i$ 时 $y=1$ 的概率。那么，建立逻辑回归问题可以简单理解为：y 在[0,1]内取某个值的概率与 x 的函数关系。本文采用极大似然估计利用特征数据集以及分类数据集构建逻辑回归模型如式子(2)：

$$p=p\left(y=1\,|\,x\right)=\frac{\exp\left(\alpha+\beta x\right)}{1+\exp\left(\alpha+\beta x\right)} \tag{2}$$

变换得式(3)：

$$\frac{p}{1-p}=\exp\left(\alpha+\beta x\right) \tag{3}$$

其中 p 表示溶度大于 0.15 mg/m^3，$1-p$ 表示溶度小于 0.15 mg/m^3，α、β 是模型参数。式子(2)为一元逻辑回归模型，对取得式子(4)：

$$\operatorname{logit}\left(p\right)=\ln\left(\frac{p}{1-p}\right)=\alpha+\beta x \tag{4}$$

式(4)是关于的线性函数，式子(2)是关于的非线性函数。实际应用中，变量会有多个，此时就需要引用多元逻辑回归，通过式子(3)和式子(4)可以得到式(5)、式(6)。

$$p=p\left(y=1\,|\,x_1,\cdots,x_k\right)=\frac{\exp\left(\alpha+\sum_{i=1}^{k}\beta_i x_i\right)}{1+\exp\left(\alpha+\sum_{i=1}^{k}\beta_i x_i\right)} \tag{5}$$

$$\operatorname{logit}\left(p\right)=\ln\left(\frac{p}{1-p}\right)=\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k \tag{6}$$

DLR 模型设计思想

云计算环境下，DLR 模型设计思想主要分以下过程：

1) 训练样本集，

2) 对测试集分类结果进行结果预测。

根据上述逻辑回归模型定义可以设定属于正类概率的 DLR 模型回归方程如式(7)：

$$f(x,\beta)=p(y=1\mid x_1,x_2)=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2)} \tag{7}$$

其中 DLR 模型回归方程的参数 $\beta=(\beta_0,\beta_1,\beta_2)$ 可以从训练结果中得到，向量 $x=(x_1,x_2)$ 为能见度和湿度。于是可以得出 y 的概率如式(8)：

$$P(y\mid x,\beta)=f(x,\beta)^y\left[1-f(x,\beta)\right]^{1-y} \quad (y=0 \text{ 或 } 1) \tag{8}$$

采用极大似然估计建立似然函数 L 如下(9)：

$$L=P(y\mid x,\beta)=\prod_{i=1}^{n}f(x_i,\beta)^{y_i}\left[1-f(x_i,\beta)\right]^{1-y_i} \tag{9}$$

令 $\theta=\ln L=\ln P(y\mid X,\beta)=\sum_{k=1}^{N}\left(y_i\ln f(x_i,\beta)+(1-y_i)\left(1-f(x_i,\beta)\right)\right)$，其中，$x_i$ 为数据集 X 得第 i 个记录，y_i 是第 i 个记录的分类结果，计算 θ 的最大值可以得到参数：

$$\beta=\max\ln P(y\mid x,\beta) \tag{10}$$

本文采用梯度下降法解极大似然估计：

$$\beta^{i+1}=\beta^i+\varepsilon\frac{\partial\ln P(X,y,\beta)}{\partial\beta}=\beta^i+\varepsilon X^{\mathrm{T}}\left(y-f(X,\beta)\right) \tag{11}$$

云计算环境下分布式 DLR 模型基本步骤如下：

1) 初始化参数，若干运算节点同时训练训练数据集。

2) 各运算节点分别计算对应样本的以及样本各特征值与的积。

3) 各节点求和得出行向量并转置得到的值。

4) 判断是否收敛，是则运算结束，否则执行(1)、(2)，直至收敛。

3. DLR 模型实验结果与分析

3.1. 实验数据集

数据集为广西南宁环保局环境监测站以及中国气象局气象数据中心 1980~2014 年气象数据，包括气溶胶、能见度、湿度等特征值。

实验过程将湿度、能见度各划分三个区间范围，湿度的三个区间分别是：湿度值<40%；40%≤ 湿度值 ≤ 90%；湿度值>90%。能见度的三个区间分别是能见度值<8 km；8 km≤ 能见度值 ≤19 km；能见度值>19 km。然后使用数据库 SQL 语句将这些区域进行两两组合构成 9 种组合方式，再分别通过 SQL 语句连接 PM10 数据表，共 9 个实验数据表源，分别是数据表 1 为湿度值<40%，能见度值<8 km 对应的气溶胶 PM10 值；数据表 2 为湿度值<40%，8 km ≤ 能见度值 ≤ 19 km 对应的气溶胶 PM10 值；数据表 3 为湿度<40%，能见度值>19 km 对应的气溶胶 PM10 值；数据表 4 为湿度值>90%，能见度值<8 km 对应的气溶胶 PM10 值；数据表 5 为湿度值>90%，8 km ≤能见度值 ≤19 km 对应的气溶胶值；数据表 6 为湿度>90%，能见度值>19 km 对应的气溶胶 PM10 值；数据表 7 为 40%≤ 湿度值 ≤ 90%，能见度值<8 km 对应的气溶胶 PM10 值；数据表

8 为 40%≤ 湿度值 ≤90%，8 km≤ 能见度值 ≤19 km 对应的气溶胶 PM10 值；数据表 9 为 40%≤ 湿度值 ≤90%，能见度值>19 km 对应的气溶胶 PM10 值，以下数据源表格以数据表 2 为例，数据源内容如表 1 所示(即湿度值<40%，8 km≤ 能见度值 ≤12 km 组合的表格数据源)，实验结果截图以数据表 2 和数据表 8 为例。

Table 1. Part data of data table 2
表 1. 数据表 2 部分数据

年	月	日	能见度值	气溶胶 PM10 值	湿度值
1989	11	22	12	0.155	31.3
1989	11	25	18	0.137	29.8
1989	12	24	13	0.133	27.6
1989	1	4	10	0.110	37.6
1989	10	10	18	0.099	36.0
1989	1	12	12	0.155	24.8
1989	10	18	14	0.113	36.9
1989	12	13	15	0.110	34.3
1989	12	16	10	0.099	30.3
1989	12	22	8	0.155	34.8
1989	10	5	15	0.112	37.6

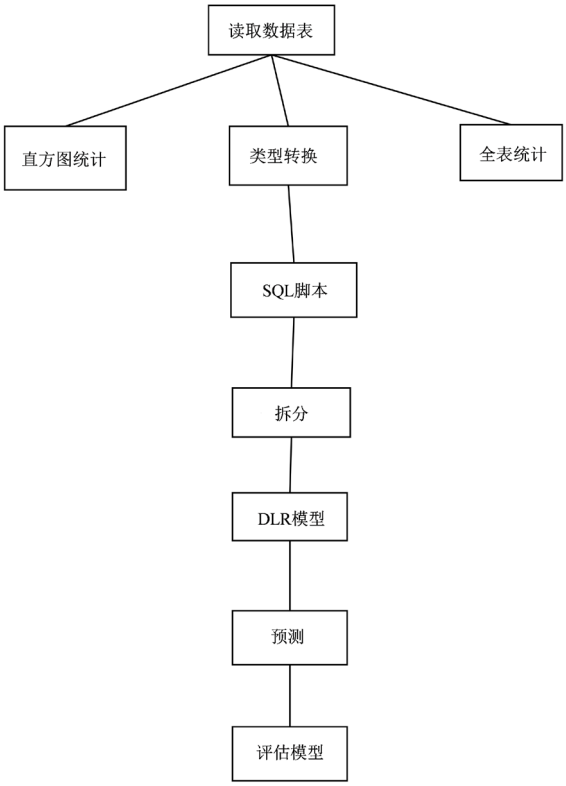

Figure 1. Experiment procedure of DLR model
图 1. DLR 模型实验过程

3.2. 实验过程

DLR 模型是通过二分类的思想来对输入的数据源进行预测分析，将能见度、湿度作为特征列，对气溶

胶 PM10 标签列进行预测分析。在实验当中，我们分为四个部分，分别是读取数据源以及对数据预处理；使用直方图、全表统计对数据表进行统计分析；DLR 模型的训练以及预测分析。最后通过评估模型分析 DLR 模型的预测准确率，以及哪个区间范围内的数据更适合用来预测 PM10 的值，实验过程如图 1 所示。

3.3. 实验结果分析

以数据表 2、数据表 8 作为分析对象，训练前按照 0.3 比例划分数据，即 30%作测试集，70%作训练集。设置迭代次数 100；正则系数 1；最小收敛误差 0.00001。DLR 模型输出如图 2、图 3 所示。

在输入数据为稀疏的时候，不显示weight全是0的特征

字段名	权重	
	1	0
values1	-0.1555705576654346	-
values3	-0.010142593518896	-
常量	0.8889573419259068	0

Figure 2. DLR model output results of data Table 2
图 2. 数据表 2 DLR 模型输出结果

在输入数据为稀疏的时候，不显示weight全是0的特征

字段名	权重	
	1	0
values1	-0.3490370382035013	-
values3	-0.08337938371483497	-
常量	5.710147458279673	0

Figure 3. DLR model output results of data Table 8
图 3. 数据表 8 DLR 模型输出结果

其中系数值越大影响越大，+表示正相关，−表示负相关。图 2 中表明湿度、能见度与 PM10 呈负相关。数据表 2 和数据表 8 为数据源的 DLR 模型准确性分别为 0.6649 和 0.8279 即 ROC 面积，值越大表明分类效果越好，如图 4、图 5 所示。

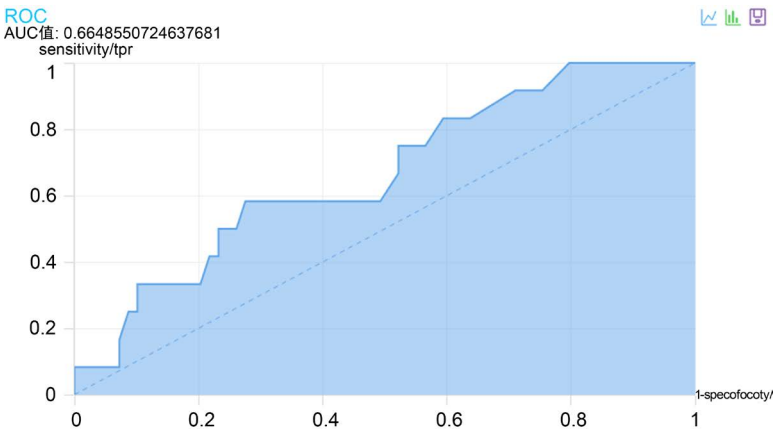

Figure 4. DLR model roc curve of data Table 2
图 4. 数据表 2 DLR 模型 ROC 曲线

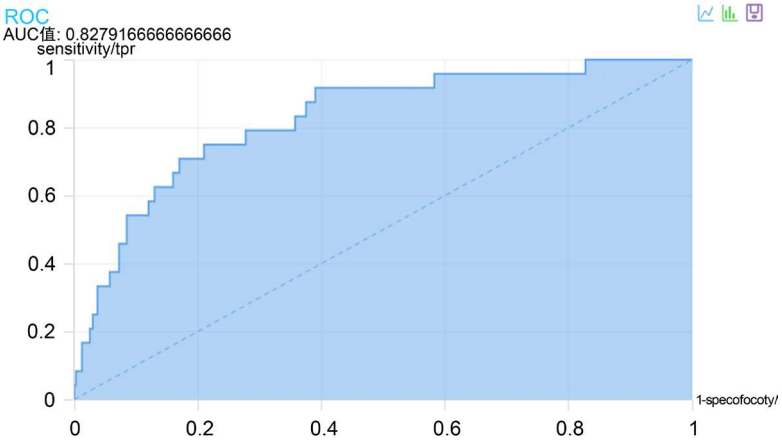

Figure 5. DLR model roc curve of data Table 8
图 5. 数据表 8 DLR 模型 ROC 曲线

下面给出了 DLR 模型数据表 2、数据表 8 的预测值与原值拟合图(见图 6，图 7)，进一步说明本文提出的 DLR 模型较好的预测性能。

数据源的相关结果为表 1 (湿度值<40%，能见度值<8 km) AUC 值为 0.6671；表 2 (湿度值<40%，8 km ≤ 能见度值 ≤ 19 km) AUC 值为 0.6648；表 3 (湿度值<40%，能见度值>19 km) AUC 值为 0.8695；表 4 (湿度值>90%，能见度值<8km) AUC 值为 0.5518；表 7 (40% ≤ 湿度值 ≤ 90%，能见度值<8 km) AUC 值为 0.7876；表 8 (40% ≤ 湿度值 ≤ 90%，8 km ≤ 能见度值 ≤ 19 km) AUC 值为 0.8279，表 9 (40% ≤ 湿度值 ≤ 90%，能见度值>19 km) AUC 值为 0.5。实验结果分析主要依据预测准确率即 AUC 的值以及特征列的相关系数。通过实验证明湿度值在 40%~90%之间，能见度值在 8~19 km 之间效果最好即数据表 8，其次是湿度值在 40%~90%之间，能见度值小于 8 km 即数据表 7。效果最差的是湿度值小于 40%，能见度值在 8 km~19 km 之间即数据表 2。并且针对上述各个数据表的实验，特征列 values1 的相关系数都比 values3 的相关系数大。表九的分析结果出现反常，此表的范围为 40% ≤ 湿度值 ≤ 90%，能见度值>19 km，这个区间被称为“非常好”能见度，本文不做参考对象，同时表 3 实验结果 AUC 值也偏大，考虑样本数的情况，样本数太小，不能作为评估对象。表 5、表 6 两个数据表中符合条件的样本数太少，并且只有一个类别，故无法得到实验结果。

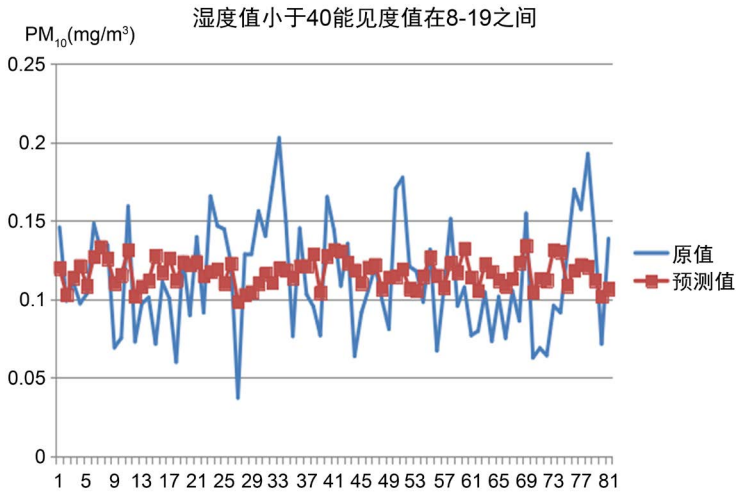

Figure 6. Fitting the predicted and original values of DLR model in data Table 2
图 6. 数据表 2 DLR 模型预测值与原值拟合图

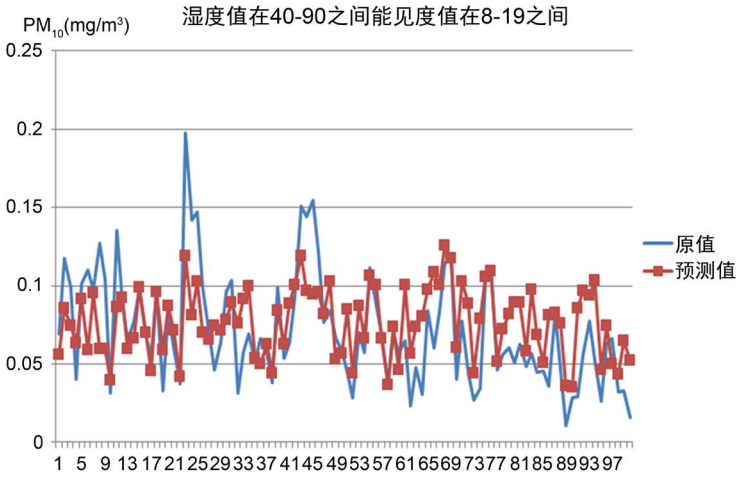

Figure 7. Fitting the predicted and original values of DLR model in data Table 8
图 7. 数据表 8 DLR 模型预测值与原值拟合图

从实验结果分析可知，湿度指标在 40%~90%之间，能见度指标在 8 km~19 km 之间，预测 PM10 值正确率最高，并且在同一湿度范围下，PM10 值与能见度值成反比；在同一能见度范围下，PM10 值与湿度值成反比。相比湿度指标，能见度指标相关性更高。针对本文研究的数据集，DLR 模型在预测正确率上比较高。

4. DLR 模型与传统回归模型时间性能的比较

如图 8 所示，DLR 模型在时间性能上要优于传统的回归模型，基本在所有划分的数据源表上运行速度都明显提高，从运行时间性能上比较进一步证明了本文提出的 DLR 算法在 PM10－能见度－湿度相关性研究上具有较好的性能，鲁棒性更强。

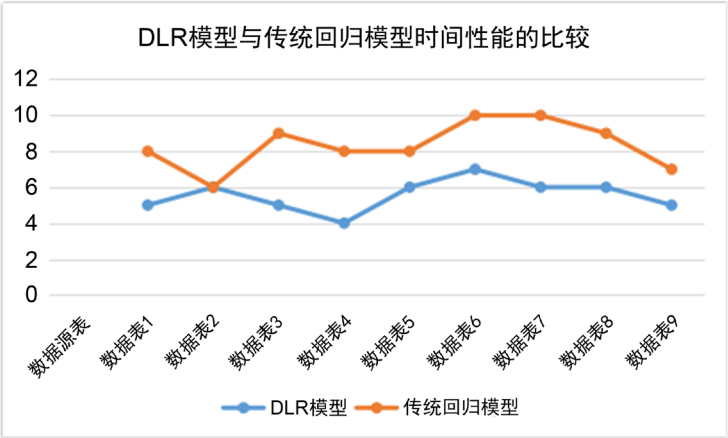

Figure 8. Comparison of running time between DLR model and traditional regression model
图 8. DLR 模型与传统回归模型运行时间比较

5. 结论

本文提出一个基于云计算环境的 DLR 模型，所用模型用于 PM10－能见度－湿度相关性研究模型总体得分较高，鲁棒性强，模型得分存在差异是由于符合条件数据样例数不同造成。从实验结果分析可知，湿度指标在 40%~90%之间，能见度指标在 8~19 km 之间，预测 PM10 值正确率最高，并且在同一湿度范

围下，PM10 值与能见度值成反比；在同一能见度范围下，PM10 值与湿度值成反比。相比湿度指标，能见度指标相关性更高。此外 DLR 算法模型在时间性能方面要优于传统回归模型。云计算环境下 DLR 模型的验证对于实际应用具有一定可行性。本文研究存在如下不足：实验气象因子(如风速、降水等)考虑欠缺，后续工作考虑加入更多的影响因子进行实验研究。

基金项目

2018 年广西高校中青年教师基础能力提升项目(NO.2018KY0699)：云环境下基于机器学习分析广西 PM10 的相关性研究；2020 年度广西高校中青年教师科研基础能力提升项目(2020KY23019)：基于深度学习的桂中地区 PM2.5 浓度预测模型研究；2020 年广西高校中青年教师基础能力提升项目(NO.2020KY23026)；2020 年校级本科教学改革工程一般项目 A 类(NO.2020GKSYGA01)；来宾市智慧农业及农业大数据应用研究团队(GXKS2020QNTD02)。

参考文献

[1] 吴建峰, 张凤太, 卢海芬, 等. 基于标准化降水指数的贵州省近 54a 干旱时空特征分析[J]. 科学技术与工程, 2018, 18(15): 207-214.

[2] 刘艳萍, 王明仕, 曹景丽, 等. 中国工业基地城市群 PM2.5 时空分布特征及相关性分析[J]. 科学技术与工程, 2018, 18(15): 184-189.

[3] Cheung, H.-C., Wang, T., Baumann, K., *et al.* (2005) Influence of Regional Pollution Outflow on the Concentrations of Fine Particulate Matter and Visibility in the Coastal Area of Southern China. *Atmospheric Environment*, **39**, 6463-6474. https://doi.org/10.1016/j.atmosenv.2005.07.033

[4] 石灵芝, 邓启红, 路婵等. 基于 BP 人工神经网络的大气颗粒物 PM10 质量浓度预测[J]. 中南大学学报, 2012(5): 1969-1974.

[5] Papanastasiou, D.K., Melas, D. and Kioutsioukis, I. (2007) Development and Assessment of Neural Network and Multiple Regression Models in Order to Predict PM10 Levels in a Medium-sized Mediterranean City. *Water*, *Air*, *and Soil Pollution*, **182**, 325-334. https://doi.org/10.1007/s11270-007-9341-0

[6] 郑湃, 崔立真, 王海洋, 等. 云计算环境下面向数据密集型应用的数据布局策略与方法[J]. 计算机学报, 2010, 33(8): 1472-1480.

[7] 王晓燕, 陈晋川, 杜小勇. 云计算环境中面向 OLTP 应用的数据分布研究[J]. 计算机学报, 2016, 39(2): 253-269.

[8] 张石磊, 武装. 一种基于 Hadoop 云计算平台的聚类算法优化的研究[J]. 计算机科学, 2012, 39(10): 115-118.

[9] 潘吴斌. 基于云计算的并行 K-means 气象数据挖掘研究与应用[D]: [硕士学位论文]. 南京: 南京信息工程大学, 2013.

[10] 王昊. 基于改进贝叶斯网络的气象数据预测算法研究[D]: [硕士学位论文]. 太原: 太原理工大学, 2016.

[11] 强宝华, 唐波, 王玉峰, 等. 基于线性回归和属性集成的分类算法[J]. 计算机科学, 2017, 44(6): 212-215.