Hans Journal of Data Mining 数据挖掘, 2019, 9(4), 153-158
Published Online October 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.94019

Hans汉斯

A Short-Term Traffic Forecasting Research Based on Didi Chuxing GAIA Open Dataset Using Echo State Network Optimized by PSO Algorithm

Jiashun Zhang

Department of Transportation, Hebei University of Technology, Tianjin
Email: jszhang@hebut.edu.cn

Received: Oct. 8th, 2019; accepted: Oct. 18th, 2019; published: Oct. 25th, 2019

Abstract

Short-term traffic forecasting is an important part of contemporary intelligent transportation systems. In this paper, based on echo state network, a procedure is proposed to provide a forecast of traffic state. For the invalid data in the Didi Chuxing GAIA Open Dataset, the random forest algorithm is used to preprocess the data and eliminate the invalid data. Because of the huge amount of real-time traffic data in GAIA data set, particle swarm optimization algorithm is employed to optimize the parameters of echo state network. Finally, the method is validated by using the local area trajectory data of the second ring road in Chengdu from October to November 2016 as the sample set. The result illustrates the effectiveness of this method.

Keywords

Echo State Network, Traffic Forecasting, PSO Algorithm

基于改进回声状态网络的盖亚大数据短时交通状态预测研究

张家顺

河北工业大学，交通运输系，天津
Email: jszhang@hebut.edu.cn

收稿日期：2019年10月8日；录用日期：2019年10月18日；发布日期：2019年10月25日

文章引用：张家顺. 基于改进回声状态网络的盖亚大数据短时交通状态预测研究[J]. 数据挖掘, 2019, 9(4): 153-158.
DOI: 10.12677/hjdm.2019.94019

摘　要

短期交通预测是现代智能交通系统重要的组成部分，本文设计了改进的回声状态网络基于盖亚开放数据来对短时交通状态进行预测。对于数据集中的无效数据，采用随机森林算法进行数据预处理，剔除其中的无效数据。由于盖亚数据集中的实时交通数据量巨大，采用粒子群算法对回声状态网络参数进行优化。最后采用2016年10月~11月成都市二环局部区域轨迹数据作为样本集对所设计的方法进行了验证，结果表明了该方法可以有效的对短时交通状态进行预测。

关键词

回声状态网络，交通预测，粒子群算法

1. 引言

近年来，随着城市化进程的加快，城市道路的拥堵问题日益突出。由于大量扩建道路会受到资金和城市空间的限制，所以发展智能交通，提高道路通行效率就成了解决城市道路交通问题的重要手段，而短时交通状态预测是实现智能交通系统的基础和前提。蒲斌等[1]提出了基于神经网络推论模型为主体的交通流量预测系统，通过实验验证了 ARIMA 乘积季节模型、BP 神经网络和 RBF 神经网络的多种训练函数的预测精度及适应性。张海鹏等[2]研究了基于公交车 GPS 数据的短时交通流预测问题。晏臻[3]等针对传统的预测方法只考虑到了交通流量的时序特征，忽略了其空间特征这一问题，提出卷积神经网络(CNN)和长短期记忆网络(LSTM)相结合的短时交通流量预测模型。李志帅等[4]提出一种基于图卷积神经网络和注意力机制的短时交通流量预测模型，该模型利用图卷积网络捕获路网流量空间特征，利用自注意力机制调整网络输出，提高最终预测结果的精确度。闫杨等[5]提出了一种基于时空相关性的短时交通流量预测方法，利用卷积－循环门控单元提取交通流量的时空特征，双向门控循环单元提取交通流量的周期特征，将提取的特征进行融合，得到交通流量的预测值。Lin [6]提出了一种基于深度学习的空中交通流量预测模型，对空中交通流量预测进行了研究。Azadeh Emami [7]使用卡尔曼滤波算法对连通车辆环境下的短时交通流预测做了研究，可以对连通车辆环境下的短时交通流提供实时预测。

回声状态网络是由 Jaeger [8] [9]在 2001 年提出并最初用于对无线通信系统的预测。由于该方法能有效的对混沌系统进行预测，采用回声状态网络进行交通预测成为了重要的研究热点。Zhang [10]基于改进的果蝇优化算法(ESN-IFOA)对回波状态网络进行优化，提出了一种 5 分钟交通量预测模型。王小洁[11]针对船舶交通事故的预测误差大，建模过程耗费时间长等难题，设计基于回声状态网络的船舶交通事故预测模型。李丁园[12]和张晋雁等[13]，对于采用回声状态网络进行预测的网络结构和参数优化做了大量的研究。Thiede [14]提出了一种基于梯度的优化算法，用于对回声状态网络的超参数进行调整。

本文基于回声状态神经网络，设计了一个基于滴滴出行盖亚大数据中车辆 GPS 轨迹数据进行短时交通状态预测的方法。本文的结构如下，第 2 节介绍了基于改进回声状态网络的预测方法。第 3 节以 2016 年 10 月~11 月成都市二环局部区域轨迹数据作为样本集验证了该方法的有效性。第 4 节对研究的结果做

了总结。

2. 基于改进回声状态网络的预测方法

在基于改进的回声状态网络对短时交通状态进行预测的问题中，主要包括提取特征集、数据预处理、学习训练、预测验证等步骤。其流程如图 1 所示。

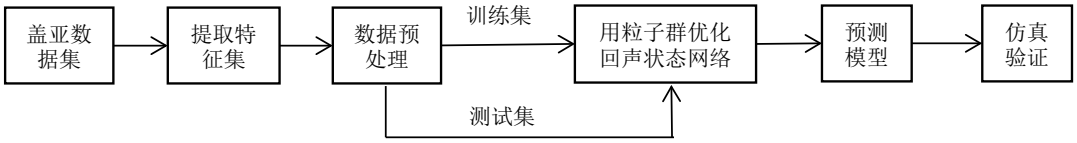

Figure 1. The procedure of structure of short-term traffic forecasting based on improved echo state network
图 1. 基于改进回声状态网络短时交通状态预测流程

2.1. 数据预处理

盖亚数据集中提供的车辆轨迹信息数据包括车辆的 GPS 信息和订单信息，其数据格式如下面两个表格(表 1~2)所示。

Table 1. The fields of GPS trajectory data of Didi Chuxing GAIA Open Dataset
表 1. 盖亚数据集车辆轨迹信息数据数据格式

字段	类型	示例	备注
司机 ID	String	glox.jrrlltBMvCh8nxqktdr2dtopmlH	已经脱敏处理
订单 ID	String	jkkt8kxniovIFuns9qrrlvst@iqnpkwz	已经脱敏处理
时间戳	String	1501584540	unix 时间戳，单位为秒
经度	String	104.04392	GCJ-02 坐标系
纬度	String	104.04392	GCJ-02 坐标系

Table 2. The fields of order data of Didi Chuxing GAIA Open Dataset
表 2. 盖亚数据集车辆订单信息数据数据格式

字段	类型	示例	备注
订单 ID	String	mjiwdgkqmonDFvCk3ntBpron5mwfrqvI	已经脱敏处理
开始计费时间	String	1501581031	unix 时间戳，单位为秒
结束计费时间	String	1501582195	unix 时间戳，单位为秒
上车位置经度	String	104.11225	GCJ-02 坐标系
上车位置纬度	String	30.66703	GCJ-02 坐标系
下车位置经度	String	104.07403	GCJ-02 坐标系
下车位置维度	String	30.6863	GCJ-02 坐标系

由于 GPS 数据采集受设备和环境的影响需要对数据中存在的无效数据进行剔除处理。由于实时采集的 GPS 轨迹数据量巨大，所以需要自动的对采集到的数据进行清理。对于盖亚开放数据集可以通过随机森林算法对车辆轨迹数据进行有效的清洗[15]。

2.2. 回声状态网络

回声状态网络是具有大量随机连接神经元的递归神经网络，该网络采用“储备池”代替传统神经网

络中的隐层。储备池由大量稀疏连接的神经元组成，并将输入信号从低维空间映射到高维空间，唯一需要训练的参数即为输出权值矩阵。这些特点大大简化了回声状态网络的训练算法和求解过程。其基本结构可以采用下图(图 2)来描述。

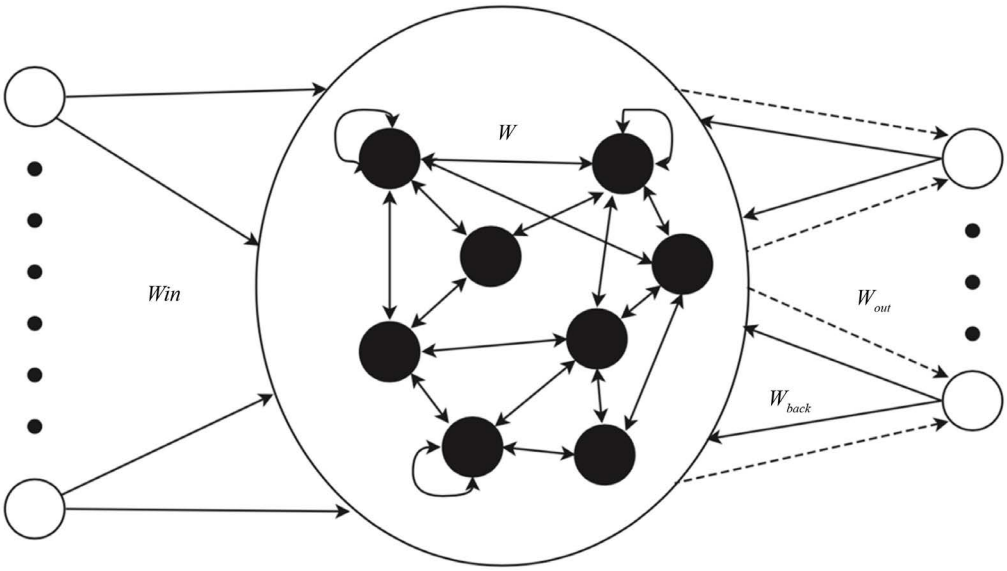

Figure 2. The topology structure of a typical echo state network
图 2. 回声状态网络拓扑结构

其中，回声状态网络的基本方程为：

$$x(n+1)=f\left(W_{\text{in}}u(n+1)+Wx(n)+W_{\text{back}}y(n)\right) \tag{1}$$

$$y(n+1)=f^{\text{out}}\left(W_{\text{out}}\left(u(n+1),x(n+1),y(n)\right)\right) \tag{2}$$

其中，$u(n)=\left(u_1(n),u_2(n),\cdots,u_K(n)\right)^T$ 和 $y(n)=\left(y_1(n),y_2(n),\cdots,u_L(n)\right)^T$ 分别表示 K 个输入单元和 L 个输出单元，$x(n)=\left(x_1(n),x_2(n),\cdots,x_N(n)\right)^T$ 表示储备池 N 个内部神经元的状态，W_{in} 为输入单元与储备池的内部连接权值矩阵，W_{out} 为储备池与输出单元连接权值矩阵，W_{back} 为输出单元与储备池的连接权重矩阵，f 和 f^{out} 分别为储备池单元和输出单元的激活函数。其中，需要训练的就是输出连接权值矩阵 W_{out}。

2.3. 改进的回声状态网络

与其它大多数机器学习算法一样，回声状态网络的几个超参数，例如输入单元尺度、储备池谱半径、储备池稀疏程度和储备池规模等，必须仔细调整才能获得最佳性能。通常，在对回声状态网络训练时，首先将样本输入和样本输出一次注入到回声状态网络的输入和输出单元，通过公式(1)迭代更新网络内部状态，直到系统输出尽可能逼近期望输出。由于盖亚开放数据集中样本量巨大，同时短时交通状态预测对于算法的实时性要求较高，需要更有效的对于回声状态网络参数的优化算法。

对于网络参数的优化，等价于下面的最小化问题：

$$\min\left(d(n)-\sum_{i=1}^{L}W_{\text{out}}x_i(n)\right) \tag{3}$$

于是，该参数问题转化为最优化问题。

粒子群优化算法是一种进化计算算法，每个粒子在搜索空间中单独的搜寻最优解，并将其记为当前

个体极值，并将个体极值与整个粒子群里的其他粒子共享，找到最优的那个个体极值作为整个粒子群的当前全局最优解，粒子群中的所有粒子根据自己找到的当前个体极值和整个粒子群共享的当前全局最优解来调整自己的速度和位置。在粒子群算法中，粒子的速度和下一个位置可以用如下的公式来获得：

$$v'_{id} = \omega * v_{id} \oplus \alpha\left(p_{id} - x_{id}\right) \oplus \beta\left(p_{gd} - x_{id}\right) \tag{4}$$

$$x'_{id} = x_{id} + v'_{id} \tag{5}$$

这里 α 和 β 是 0 到 1 之间的随机数，p_{id} 是已知粒子最佳位置，p_{gd} 是已知全局最佳位置，x_{id} 是粒子的当前位置，x'_{id} 是粒子的下一个位置，ω 是权重参数。

3. 仿真和分析

这里选取了盖亚开放数据集中 2016 年 10 月~11 月成都市二环局部区域轨迹数据作为样本集(数据来源：https://gaia.didichuxing.com)，该样本集包括约 180GB 数据，大约 1,993,642,054 条记录。其中 10 月的数据作为训练集，11 月的数据作为测试集。首先，对训练集和测试集中的数据进行数据清洗。随机森林算法进行数据清洗的关键在于特征集的提取和特征子集的维度的确定。对于数据集中的任意一点 i，选择该点和前面 8 个相邻点之间的速度、加速度共 16 个维度作为其特征子集。然后，输入训练数据采用粒子群算法对网络参数进行优化。最后，用所得到的回声状态网络对测试集数据进行预测，所得到预测数据与实际数据的比较如图 3 所示，其中折线为真实数据，黑点为预测数据。从图中可以看出预测数据与实际数据的整体趋势基本吻合，在车辆速度较高的区间误差相对较大，在车辆速度较低的区间预测值与实际值基本接近。对预测的结果计算其平均绝对误差和平均绝对百分比误差分别为 1.973%和 7.132%。结果表明，基于所设计的粒子群算法优化的改进回声状态网络，可以有效的对短时交通状态进行预测。

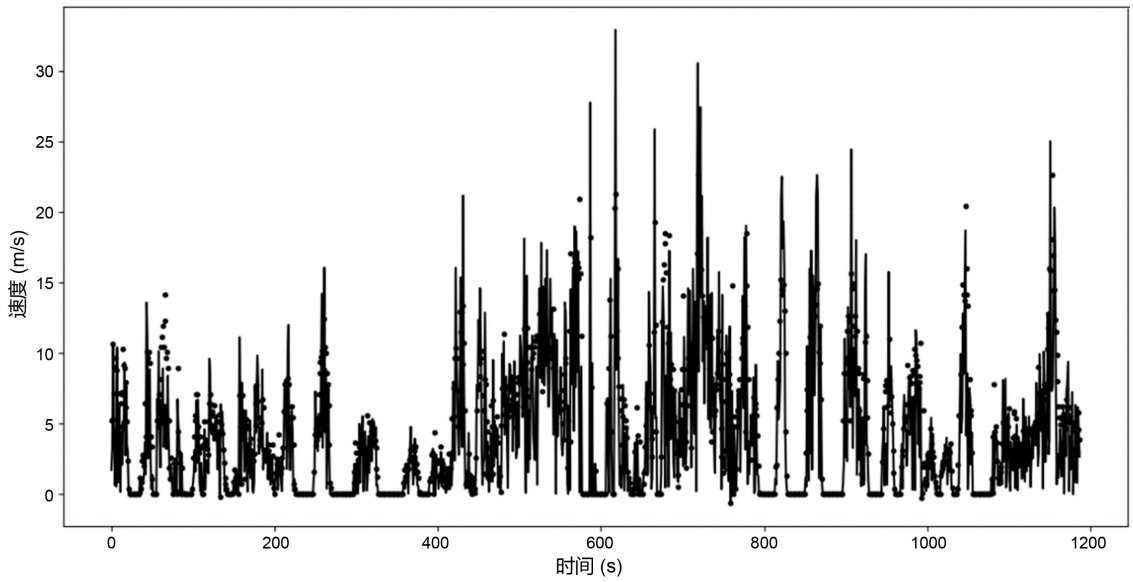

Figure 3. The comparison between predicted data and actual data
图 3. 预测数据与实际数据比较

4. 结论

本文利用滴滴出行盖亚开放数据中的 GPS 轨迹数据对短时交通流状态预测算法进行研究。由于滴滴出行的 GPS 轨迹实时数据量巨大，首先采用随机森林算法对数据自动预处理，然后采用粒子群算法对回

声状态网络参数进行快速优化。最后采用 2016 年 10 月~11 月成都市二环局部区域轨迹数据作为样本集对所设计的方法进行了验证。结果表明，该基于粒子群算法改进的回声状态网络可以有效的对短时交通状态进行预测。

致　谢

数据来自滴滴出行，数据出处：https://gaia.didichuxing.com。

基金项目

河北省科技计划项目 No.15456135。

参考文献

[1] 蒲斌, 李浩, 卢晨阳, 王治辉, 刘华. 基于神经网络的海量 GPS 数据交通流量预测[J]. 云南大学学报(自然科学版), 2019, 41(1): 53-60.

[2] 张海鹏, 杨宏业, 邬鑫珏, 王葆元. 基于公交车 GPS 数据的短时交通流预测研究[J]. 内蒙古工业大学学报(自然科学版), 2018, 37(1): 75-80.

[3] 晏臻, 于重重, 韩璐, 苏维均, 刘平. 基于 CNN+LSTM 的短时交通流量预测方法[J]. 计算机工程与设计, 2019(9): 2620-2624+2659.

[4] 李志帅, 吕宜生, 熊刚. 基于图卷积神经网络和注意力机制的短时交通流量预测[J]. 交通工程, 2019, 19(4): 15-19+28.

[5] 闫杨, 孙丽珺, 朱兰婷. 一种基于时空相关性的短时交通流量预测方法[J/OL]. 计算机工程, 1-8.

[6] Lin, Y., Zhang, J.-W. and Liu, H. (2019) Deep Learning Based Short-Term Air Traffic Flow Prediction Considering Temporal—Spatial Correlation. *Aerospace Science and Technology*, **93**, Article ID: 105113.

[7] Emami, A., Sarvi, M. and Bagloee, S.A. (2019) Using Kalman Filter Algorithm for Short-Term Traffic Flow Prediction in a Connected Vehicle Environment. *Journal of Modern Transportation*, **27**, 222-232. https://doi.org/10.1007/s40534-019-0193-2

[8] Jaeger, H. (2010) The "Echo State" Approach to Analysing and Training Recurrent Neural Networks—With an Erratum Note. German National Research Center for Information Technology GMD Technical Report Vol. 148, Bonn, 13.

[9] Jaeger, H. and Haas, H. (2004) Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Telecommunication. *Science*, **304**, 78-80. https://doi.org/10.1126/science.1091277

[10] Zhang, Q.Y., Qian, H., Chen, Y.P. and Lei, D.M. (2019) A Short-Term Traffic Forecasting Model Based on Echo State Network Optimized by Improved Fruit Fly Optimization Algorithm. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2019.02.062

[11] 王小洁. 基于回声状态网络的船舶交通事故预测[J]. 舰船科学技术, 2019, 41(16): 16-18.

[12] 李丁园. 回声状态网络结构设计及应用研究[D]: [博士学位论文]. 长春: 吉林大学, 2019.

[13] 张晋雁, 陶宏才. 回声状态网络研究[J]. 成都信息工程学院学报, 2015, 30(6): 546-550.

[14] Thiede, L.A. and Parlitz, U. (2019) Gradient Based Hyperparameter Optimization in Echo State Networks. *Neural Networks*, **115**, 23-29.

[15] 张家顺. 基于随机森林算法的盖亚大数据清洗的研究[J]. 计算机科学与应用, 2019, 9(9): 1747-1752.