Computer Science and Application 计算机科学与应用, 2019, 9(9), 1747-1752
Published Online September 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.99196

Hans汉斯

An Automatic Data Cleaning Method for GPS Trajectory Data on Didi Chuxing GAIA Open Dataset Using Random Forest Algorithm

Jiashun Zhang

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

Received: Aug. 29th, 2019; accepted: Sep. 13th, 2019; published: Sep. 20th, 2019

Abstract

A new data cleaning method for the GPS trajectory data on Didi Chuxing GAIA Open Dataset is developed. Random forests algorithm is employed to the identification of invalid, weak, and normal data of the Didi Chuxing GAIA Open Dataset raw data. Firstly, the feature set is selected according to the mathematical characteristics of three types of data, and then the optimal feature subset dimension is determined. Finally, to implement the proposed method, the Pandas and scikit-learn Python library are used to read and process the data and the result illustrates the effectiveness of this method.

Keywords

Data Cleaning, Machine Learning, Random Forest

基于随机森林算法的盖亚大数据清洗的研究

张家顺

河北工业大学，天津
Email: jszhang@hebut.edu.cn

收稿日期：2019年8月29日；录用日期：2019年9月13日；发布日期：2019年9月20日

摘　要

本文针对滴滴出行的盖亚开放数据集中的GPS轨迹数据，设计了一种自动数据清洗方法。该方法基于随机森林算法，用来识别盖亚开放数据集原始GPS轨迹数据中的无效数据，弱信号数据和正常数据。首先

文章引用: 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.
DOI: 10.12677/csa.2019.99196

根据三类数据的数学特征选择其特征集，然后确定其最优的特征子集维度。最后，基于python的pandas和scikit-learn实现所提出的方法，并以盖亚数据集中的2016年10~11月成都市二环数据集作为样本进行了实验来验证该数据清洗方法的效果，结果表明了该方法可以有效地完成弱信号数据和无效数据的数据清洗工作。

关键词

数据清洗，机器学习，随机森林

1. 引言

随着智能交通系统的发展，以 GPS 轨迹数据、订单历史等形式的交通数据迅速增加。目前，交通数据的采集，正在以前所未有的速度发展，滴滴出行在成都市仅二环线一条道路每天就会产生超过 3 GB 的数据。但是，由于采集到的 GPS 数据来自于司机的手机，再通过手机网络上传到云服务器，所以采集和存储的数据除了有正确的 GPS 数据以外，还会有由于 GPS 无信号或者 GPS 信号弱而由 AGPS 提供的定位数据，这些数据通常具有较大误差。如果不能有效地清理这些包含误差的数据，就很难对采集到的数据做有效的分析，这对数据处理技术提出了新的挑战。

Jennifer Baur [1]提出了一个有效的检测缺失数据并改良数据的方法。Patrick Röhm [2]针对投资数据设计了一种识别企业风险投资者的数据清理程序。Tomer Gueta [3]提出了用于量化用户级大数据清理价值的分布式模型。Ridha Khedri [4]则设计了一种基于代数的数据清洗方法。Salem [5]提出了基于条件函数依赖的数据清理规则。Saul Gilla [6]针对数据流的清洗设计了一个分布式的计算框架。

由于交通系统固有的复杂性和庞大的数据量，传统的数据清洗方法面对大数据量开始力不从心。随着人工智能和机器学习的发展，大量学者提出了很多基于机器学习的分类的自动数据清洗方法。随机森林算法，作为集成学习中的典型算法，以具有极好的准确率，能够有效地运行在大数据集上，能够处理具有高维特征的输入样本，而且不需要降维，对于缺省值问题也能够获得很好的结果等优秀特性，获得了大量的研究和应用。李传冰[7]研究了基于机器学习对电子回旋辐射成像信号的数据清洗问题，并分别使用了支持向量机和随机森林两种方法进行了研究。张西宁[8]针对随机森林算法不能处理异常检测问题的局限，提出了一种基于改进 Graham 扫描法的单类随机森林，实现了随机森林在只有单类样本时的分类应用。徐乔[9]为抑制相干斑噪声对极化 SAR 图像分类结果的干扰，提出一种综合多特征的极化 SAR 图像随机森林分类方法。郑建华[10]针对传统分类算法难以处理不平衡数据的问题，提出了一种基于混合采样策略的改进随机森林不平衡数据的分类算法。刘云翔[11]在分析随机森林算法基本原理的基础上，提出一种改进的基于随机森林的特征筛选算法，并将该系统应用于肝癌预后预测。尹儒[12]提出了一种模型决策森林算法以提高模型决策树的分类精度。林栢全[13]为进一步提高基于多准则评分的推荐算法的推荐性能，着重分析了基于张量分解的推荐算法和基于聚类与降维的模糊推理系统推荐算法并基于这两种算法在目标数据集上的运行结果，提出了基于矩阵分解与随机森林的多准则推荐算法。张宸宁[14]为解决 SMOTE 算法在生成数据时会弱化数据的真实分布，同时考虑到本福特法则在处理自然数据中可以弥补数据弱化这一特点，将 SMOTE 算法与本福特法则相结合，提出一种新的处理类别不平衡数据的算法，以

提高数据分布真实性和准确性。孙悦[15]、董娜[16]和关晓蔷[17]分别设计了基于 Spark、贝叶斯模型组合和类别随机化的改进随机森林算法。朱冰[18]提出了一种基于随机森林模型的驾驶人驾驶习性辨识策略来辨识驾驶人驾驶习性。

本文基于随机森林算法，设计了一个对滴滴出行，盖亚大数据中车辆 GPS 轨迹数据进行自动清洗的算法。本文的结构如下，第 2 节介绍了滴滴出行盖亚开放数据的格式和特征。第 3 节定义了数据的特征集，并确定了最优特征子集的维度。并以 2016 年 10 月~11 月成都市二环局部区域轨迹数据作为样本集进行了数据清洗实验。第 4 节对研究的结果做了总结。

2. 滴滴出行盖亚开放数据集

在过去六年的时间里，滴滴出行日订单已突破 3000 万，为超过 5.5 亿用户提供出租车、快车、专车、豪华车、顺风车、公交、小巴、代驾、企业级、共享单车、外卖等全面的出行和运输服务。盖亚数据开放计划，依托于滴滴出行的大数据，面向学术界提供真实的数据资源，其提供的海量出行数据，可以有效的帮助我们了解城市交通状况。

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，其数据格式如表 1、表 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

图 1 中展示了盖亚开放数据集中依据原始数据绘制的典型车辆速度变化曲线，从图中我们可以看到部分速度数据的剧烈变化显然是不正常的。

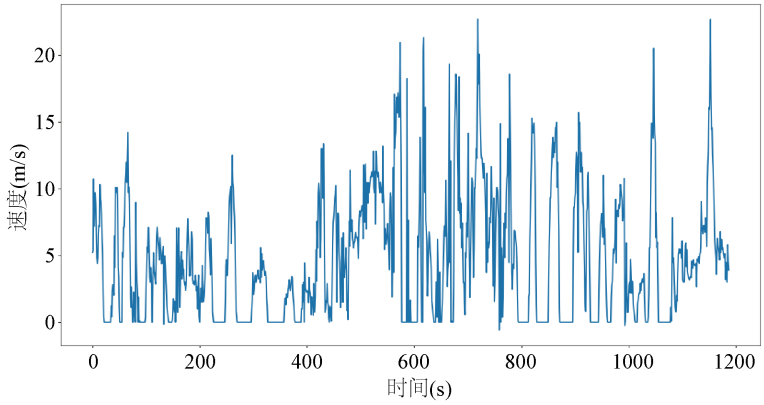

Figure 1. Typical patterns of vehicle speed from raw Didi Chuxing GAIA Open Dataset
图 1. 盖亚开放数据集中典型车辆速度变化曲线

3. 盖亚轨迹数据集的自动清洗方法

滴滴出行盖亚开放数据中的 GPS 轨迹数据来自于驾驶员的手机 GPS，其信号容易受到天气和环境的影响，精度很难得到保证。为了有效的利用这些数据，可以将盖亚数据集中的数据分为如下三类：正确数据，弱信号数据和无效数据。无效数据可以直接去除，弱信号数据可以进一步进行修正。为了识别这三类数据，首先需要找到数据的特征集。在盖亚轨迹数据集中，相邻两点的距离 l_i 可以定义如下：

$$l_i = 12756274 \times \arcsin\left(\sqrt{\sin^2\left(\frac{\pi\times(lati_i - lati_{i-1})}{360}\right) + \cos\left(\frac{\pi\times lati_i}{180}\right)\times\cos\left(\frac{\pi\times lati_{i-1}}{180}\right)\times\sin^2\left(\frac{\pi\times(long_i - long_{i-1})}{360}\right)}\right) \tag{1}$$

$i = 1,2,\cdots,n$，其中 $lati_i$ 和 $long_i$ 分别是点 i 在 GCJ-02 坐标系下的经度和纬度坐标，t_i 车辆在 i 时的时间，其数据为 unix 时间戳格式。

根据(1)，我们可以得到车辆在点 i 和点 $i-1$ 之间的平均速度 v_i。

$$v_i = \frac{l_i - l_{i-1}}{t_i - t_{i-1}} \tag{2}$$

以及加速度 a_i：

$$a_i = \frac{v_i - v_{i-1}}{t_i - t_{i-1}} \tag{3}$$

$$i = 1,2,\cdots,n$$

进一步的，我们可以得到车辆在点 i 和点 $i-j$ 之间的平均速度 v_{ij}：

$$v_{ij} = \frac{l_i - l_{i-j}}{t_i - t_{i-j}}, \tag{4}$$

$$j = 1,2,\cdots,M-1$$

对于任意的一点 i，可以得到其对应的一组 M 维特征集 $(a_i, v_{i1}, v_{i2}, \cdots, v_{iM-1})$。

在随机森林算法中，最重要的参数是选定的特征子集的维度 m。减小特征选择个数 m，树的相关性

和分类能力也会相应的降低；增大 m，两者也会随之增大。所以关键问题是如何选择最优的 m，为了解决这个问题需要对不同的 m 分别计算包外错误率。在选定的样本数据集中，m 的取值范围为$[1, M]$，从图 2 中可以看出，当 $m = 16$ 时，包外错误率就降到了比较低的水平，所以最终选择的特征子集的维度是 16。

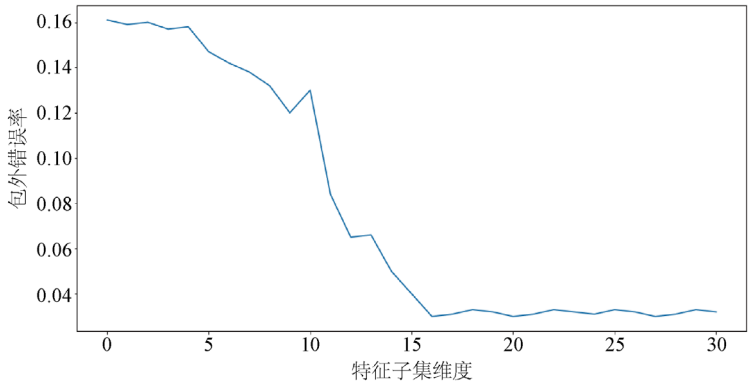

Figure 2. The influences of the number of feature variables on the out-of-bag classification error
图 2. 特征子集的维度对包外错误率的影响

这里选取了 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 180 GB 数据，大约 1993642054 条记录。该自动清洗程序基于 Python 中的 pandas 和 scikit-learn 来实现随机森林。图 3 显示了在样本数据集中，各部分数据的比例。其中，无效数据占总样本数的 3%，信号弱数据占样本总数的 2%，有效数据为 95%，该开放数据集质量较高。

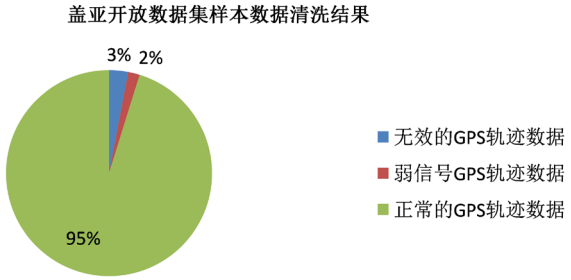

Figure 3. The results of the automatic cleaning of Sample Data Set
图 3. 盖亚开放数据集盖亚开放数据集成都市二环样本数据清洗结果

4. 结论

在对交通大数据的研究中，对于数据进行有效的清洗是一个重要的研究方向。本文针对滴滴出行的盖亚开放数据集中的 GPS 轨迹数据，设计了一种基于随机森林算法自动数据清洗方法用来识别盖亚开放数据集原始 GPS 轨迹数据中的无效数据，弱信号数据和正常数据。该方法根据数据集中三类数据的数学特征选择其特征集并确定其最优的特征子集维度并以盖亚数据集中的 2016 年 10~11 月成都市二环数据集作为样本进行了实验来验证该数据清洗方法的有效性。

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] Baur, J., Moreno-Villanueva, M., Kötter, T., Sindlinger, T., Bürkle, A., Berthold, M.R. and Junk, M. (2015) MARK-AGE Data Management: Cleaning, Exploration and Visualization of Data. *Mechanisms of Ageing and Development*, **151**, 38-44. https://doi.org/10.1016/j.mad.2015.05.007

[2] Röhm, P., Merz, M. and Kuckertz, A. (2019) Identifying Corporate Venture Capital Investors—A Data-Cleaning Procedure. *Finance Research Letters*.

[3] Gueta, T. and Carmel, Y. (2016) Quantifying the Value of User-Level Data Cleaning for Big Data: A Case Study Using Mammal Distribution Models. *Ecological Informatics*, **34**, 139-145. https://doi.org/10.1016/j.ecoinf.2016.06.001

[4] Khedri, R., Chiang, F. and Sabri, K.E. (2013) An Algebraic Approach towards Data Cleaning. *Procedia Computer Science*, **21**, 50-59. https://doi.org/10.1016/j.procs.2013.09.009

[5] Salem, R. and Abdo, A. (2016) Fixing Rules for Data Cleaning Based on Conditional Functional Dependency. *Future Computing and Informatics Journal*, **1**, 10-26. https://doi.org/10.1016/j.fcij.2017.03.002

[6] Gilla, S. and Lee, B. (2015) A Framework for Distributed Cleaning of Data Streams. *Procedia Computer Science*, **52**, 1186-1191. https://doi.org/10.1016/j.procs.2015.05.156

[7] Li, C., Lan, T., Wang, Y., Liu, J., Xie, J., Lan, T., Li, H. and Qin, H. (2018) An Automatic Data Cleaning Procedure for the Electron Cyclotron Emission Imaging on EAST Tokamak Using Machine Learning Algorithm. *Journal of Instrumentation*, **13**, P10029. https://doi.org/10.1088/1748-0221/13/10/P10029

[8] 张西宁, 张雯雯, 周融通, 向宙. 基于单类随机森林的异常检测方法及应用[J/OL]. 西安交通大学学报, 2019(12): 1-8.

[9] 徐乔, 张霄, 余绍淮, 陈启浩, 刘修国. 综合多特征的极化 SAR 图像随机森林分类算法[J]. 遥感学报, 2019, 23(4): 685-694.

[10] 郑建华, 刘双印, 贺超波, 符志强. 基于混合采样策略的改进随机森林不平衡数据分类算法[J]. 重庆理工大学学报(自然科学), 2019, 33(7): 113-123.

[11] 刘云翔, 陈斌, 周子宜. 一种基于随机森林的改进特征筛选算法[J]. 现代电子技术, 2019, 42(12): 117-121.

[12] 尹儒, 门昌骞, 王文剑. 一种模型决策森林算法[J/OL]. 计算机科学与探索, 1-11.

[13] 林栢全, 肖菁. 基于矩阵分解与随机森林的多准则推荐算法[J]. 华南师范大学学报(自然科学版), 2019, 51(2): 117-122.

[14] 张宸宁, 李国成. 基于 BL-SMOTE 和随机森林的不平衡数据分类[J]. 北京信息科技大学学报(自然科学版), 2019, 34(2): 23-28.

[15] 孙悦, 袁健. 基于 Spark 的改进随机森林算法[J]. 电子科技, 2019, 32(4): 60-63+67.

[16] 董娜, 常建芳, 吴爱国. 基于贝叶斯模型组合的随机森林预测方法[J]. 湖南大学学报(自然科学版), 2019, 46(2): 123-130

[17] 朱冰, 李伟男, 汪震, 赵健, 何睿, 韩嘉懿. 基于随机森林的驾驶人驾驶习性辨识策略[J]. 汽车工程, 2019, 41(2): 213-218+224.

[18] 关晓蔷, 庞继芳, 梁吉业. 基于类别随机化的随机森林算法[J]. 计算机科学, 2019, 46(2): 196-201.