文章编号：2095-2163(2021)09-0156-05 中图分类号：TP399 文献标志码：A

PDTR 模型对城市流动人口的预测

吴 宇，孙宏宇，孙明辰，王洪君
（吉林师范大学 计算机学院，吉林 四平 136000）

摘 要：当下所有人口大国都面临严峻的人口问题，人口流动预测成为其发展过程中的重要议题。因此，对中国城市流动人口进行准确预测具有重要的意义。本文利用主成分分析（PCA）与决策树（Decision Tree）相结合的模型，预测中国流动人口情况。实验证明，本文所使用的 PDTR 模型具有良好的预测精度。

关键词：流动人口；主成分分析；决策树；PDTR 模型

Urban floating population prediction with PDTR model

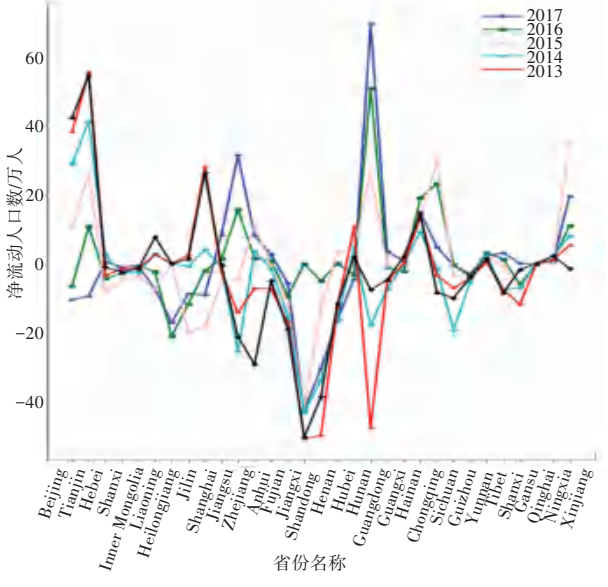
WU Yu，SUN Hongyu，SUN Mingchen，WANG Hongjun
（College of Computer Science，Jilin Normal University，Siping Jilin 136000，China）

【Abstract】At present，all the populous countries are faced with severe population problems，and population flow prediction has become an important issue in their development process. Therefore，it is of great significance to accurately predict the urban floating population in China. In this paper，a model combining principal component analysis（PCA）with Decision Tree is used to predict the situation of floating population in China. Experiments show that the PDTR model used in this paper has a good prediction accuracy.

【Key words】floating population；principal component analysis；decision tree；PDTR model

0 引 言

人口迁移这一社会现象目前已经引起多学科交融研究领域学者的关注，从图 1、图 2 可以看出，各个省份的净人口流动数量与自然增长率趋势截然不同。数据表明各地区的人口数量变化情况是受条件影响的，并不完全取决于人口基数，受人口迁移的影响也十分显著。进行人口迁徙预测可以更好的把握地区人口变化情况以及地区城市化情况，对社会经济发展具有重要指导意义。因此，进行人口迁徙预测研究势在必行。但目前研究上存在一些不足，一方面城市参数众多，应用现有技术将其统计可以轻易实现，但其中掺伴的无效数据，不仅无形中提高了实验的能耗，也造成了数据混淆；另一方面，传统的人口流动预测方法大多是根据经济、政策等理论来总结人口流动规律加以预测。如：流动人口的规模总量和结构形式随经济体发展变迁的规律、城市收入水平和公共服务能力差异，是吸引外来人口流入的首要因素等等[1]。但无论使用什么方法，其根本在于分析人口流动情况和其影响因素之间的关系，并通过该关系构建模型或形成理论预测未来人口变化情况。



数据来源：国家统计局发布

图 1 2013~2017 年各省份净流动人口数

Fig. 1 The number of net population flow by province in 2013-2017

随着科技的发展，通过人工智能的方法进行大数据分析预测城市人口，可以节省大量的时间以及资源的消耗。数据的获取以及预测算法的选择在很大程

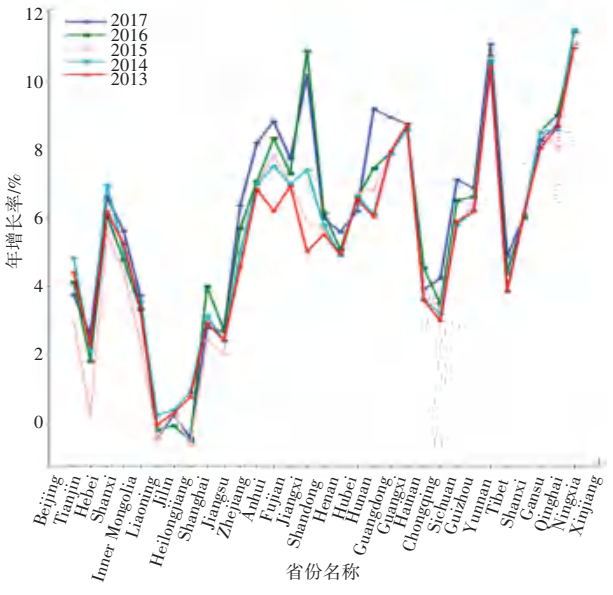
基金项目：吉林省教育厅科学技术研究项目（JJKH20210457KJ）；吉林省大学生创新创业训练项目（2020JLSFDX-JSJ03）。

作者简介：吴 宇（1999-），女，硕士研究生，主要研究方向：人工智能安全；孙宏宇（1986-），女，博士，讲师，主要研究方向：无线网络与智能计算；孙明辰（1998-），男，硕士研究生，主要研究方向：人工智能；王洪君（1965-），男，博士，教授，硕士生导师，主要研究方向：信息安全。

通讯作者：孙宏宇 Email：hongyu@jlnu.edu.cn

收稿日期：2021-06-18

度上影响着预测结果的精确性，不同模型对于人口的预测结果也不同[2]。本文旨在提出一种 PDTR 预测模型，通过使用人工智能算法，总结出人口流动与影响其发生变化的城市参数之间的关系并形成模型，以进行对各省份未来人口流动情况的预测。



数据来源：国家统计局发布

图 2　2013～2017 年各个省份自然增长率

Fig. 2　Natural growth rates for each province (2013～2017)

1　PCA 原理

PCA 方法可以利用降维思想抓住所要研究问题的主要矛盾，简化复杂问题，使研究效率得到提高[3]。

本文从燃气、供水、供热、公共交通、城市市容、绿地园林等 7 个方面中，选取 46 项城市参数指标，由于在选择训练样本时，各个样本指标之间的可能相关性较高，所以可能导致样本信息过度重复的情况，这时就需要借助 PCA 方法来概括诸多信息的主要方面，对样本指标信息进行降维。通过这些综合指标相互独立地代表某一方面的性质，从而改进训练样本的有效性[4]。

将现有 m 个城市指标参数组成的原始数据集，分别用 I_1，I_2，…，I_m 表示，由这 m 个城市参数指标组成了 m 维随机向量 $\boldsymbol{I}=(I_1, I_2, \ldots, I_m)$，设 α 为随机向量 $\boldsymbol{I}$ 均值；随机向量 $\boldsymbol{I}$ 线性变换成新的综合变量，用 $\boldsymbol{D}$ 表示。新综合变量 $\boldsymbol{D}$ 与原始变量 $\boldsymbol{I}$ 线性关系由公式(1)表示[5]：

$$\begin{cases} D_1 = \alpha_{11} I_1 + \alpha_{12} I_2 + \ldots + \alpha_{1m} I_m \\ D_2 = \alpha_{21} I_1 + \alpha_{22} I_2 + \ldots + \alpha_{2m} I_m \\ \cdots \\ D_n = \alpha_{n1} I_1 + \alpha_{n2} I_2 + \ldots + \alpha_{nm} I_m \end{cases} \tag{1}$$

式中：系数 α_{ij} 可以根据下面几个原则来确定：

(1) ${\alpha_{11}}^2 + {\alpha_{12}}^2 + \ldots + {\alpha_{1m}}^2 = 1 (i = 1, 2, \cdots, m)$；

(2) D_i 与 $D_j (i \neq j; i, j = 1, 2, \cdots, n)$ 线性无关；

(3) D_1 为 I_1，I_2，…，I_m 所有线性组合中方差最大者；D_2 为与 D_1 不相关的 I_1，I_2，…，I_m 的所有线性组合中方差最大者；D_n 为 $D_1, D_2, \ldots, D_{n-1}$ 都不相关的线性组合中方差最大者。

这样确定的新变量指标 $D_1, D_2, \ldots, D_n$ 分别称为原变量指标 I_1，I_2，…，I_m 的第 1 主成分，第 2 主成分，…，第 n 主成分。其中，$D_1, D_2, \ldots, D_n$ 的方差依次减小。实际问题分析时，常挑选前面几个最大的主成分，这样既可以减少变量的数目，又抓住了问题的主要矛盾，简化了各变量之间的关系[6]。

本文最终使用 PCA 的 fit 方法，对全部训练数据进行训练，得到训练好的 PCA 模型。输入格式为 $fit(X)$，其中 X 是预处理后的训练集数据样本。通过 PCA 的 transform 方法将全部训练数据进行变换，得到经过主成分分析后的特征。输入格式为 $transform(X)$，其中 X 是待转换的数据，也是后续决策树分析的输入数据。

2　PDTR 模型构建

决策树是一种树形结构的分类与回归方法[7]，其目的是通过对训练集进行学习，找出特征和类别之间的关系。一旦这种关系被找出，就能用其来预测未知类别数据的类别。本文使用决策树回归分析方法进行回归分析，所谓“决策”就是进行一次选择，每进行一次选择实质上就是对特征空间进行一次划分，每划分出一个单元该单元就会有一种特定的输出[8]。而划分或做“决策”的过程就是建立决策树的过程。本文使用标准差标准化和主成分分析(PCA)进行数据与处理，对预处理后的数据使用决策树回归模型(Decision Tree Regression)进行回归分析，以得到预测模型。具体流程如图 3 所示。

实现步骤如下：

(1)对输入数据进行预处理，其中包括数据清洗和标准化；

(2)对处理后的数据进行主成分分析，得到降维后数据；

(3)使用降维后数据训练决策树模型；

(4)对测试数据进行预测得到结果，若结果达到标准则保存模型对真实数据进行预测，否则修改主成分分析和决策树回归模型的参数，返回步骤(3)继续进行第三步操作。

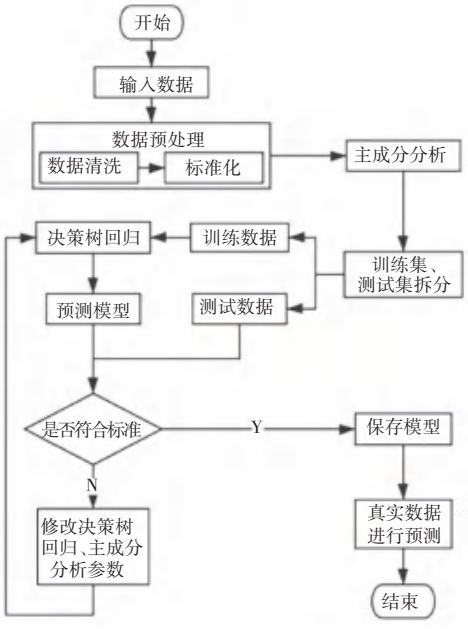


图 3 PDTR 模型的总体设计方案流程图

Fig. 3 Overall design scheme flow char of PDTR model

3 实验与结果分析

3.1 数据集

本文实验数据来源于 2006~2017 年《中国统计年鉴》,从各年的数据中选取供热、供水、燃气、城市市容、公共交通、绿地园林等 6 大类城市参数指标,共 45 小项数据类别进行分析,将各年的出生率、死亡率、年增长率和 6 大类城市参数指标进行了集成用于预测实验。详细情况见表 1。

3.2 实验及结果

数据的完整性很重要,会影响到后续的数据处理。本文对于重要的数据,使用的是相对于丢弃更常用的补全。首先利用 Pandas 的 fillna 方法,将原始数据集中的缺省值部分填充为相应特征下样本的平均值(df.fillna(df.mean()['chas':'rm'])); 再利用 StandardScaler 对上一步处理后的数据,采用公式(2)进行数据去均值和方差,实现数据归一化,以便更好地对数据进行特征提取。

表 1 城市参数数据集

Tab. 1 City parameter data set

指标	指标项	单位
供热	蒸汽	吨/小时
	热水	兆瓦
	蒸汽、热水	万吉焦
	管道长度	公里
	供热面积	万平方米
供水	生产能力	万立方米/日
	供水管道长度	公里
	全年供水总量、生活用水、生产用水	万立方米
	用水人	万人
	人均日生活用水	升
燃气	液化石油气	吨/日
	燃气产力、人工煤气、天然气、液化石油气	万立方米/日
	人工煤气、天然气	万立方米
	液化石油气、人工煤气、天然气	吨
城市市容	生活垃圾、粪便清运	万吨
	车辆设备	台
	公共厕所、三类以上	座
	清扫保洁	万平方米
绿地园林	绿地面积、公园绿地	公顷
	公园	个
	公园面积	公顷
	建成区绿化覆盖率	百分数
公共交通	轨道交通	辆
	运营线路总长度、公共汽、电车、轨道交通	公里
	公共交通客运总量、公共汽、电车、轨道交通	万人次
	出租汽车、公共交通车辆、公共汽、电车	辆

数据来源:中国统计年鉴

$$x^* = \frac{x - \mu}{\sigma} \tag{2}$$

式中：μ 为所有样本数据的均值，σ 为所有样本数据的标准差。

将归一化后的 6 个指标（$x = (x_1, x_2, \ldots, x_6)$）作为 PDTR 模型的自变量，将流动人口（10 万人）y 作为因变量。

本文共采集 46 项城市参数指标，为了更好的保存数据信息且提高实验效率，使用 PCA 时选取了前 24 项主成分，将数据从 46 维降维 24 维；在使用 Decision Tree Regression 时，本文针对 2016 年数据，将 max_depth 参数即决策回归树的最大深度设置为从 1 开始，通过不断迭代直至达到极限，得到图 4 所示结果。将 min_weight_fraction_leaf 参数，即最小权重系数设置为从 0 开始，通过不断迭代直至达到极限，得到图 5 所示结果。

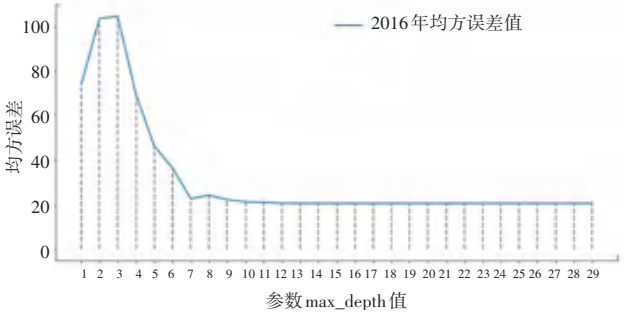


图 4　2016 年均方误差变化情况

Fig. 4　Change of the mean annual square error in 2016

图 5　2016 年均方误差变化情况

Fig. 5　Change of the mean annual square error in 2016

图 4 中蓝色折线代表 2016 年份的原始数据经数据预处理后，对设置了不同 max_deep 值的决策树回归模型进行训练，得到的均方误差值。从图 4 中可以看出，将 max_deep 值设置为 14 时，预测的绝对误差相对较小。因此，本文在使用决策树回归模型时将该参数设置为 14。

图 5 中蓝色的折线代表 2016 年份的原始数据经数据预处理后，使用处理后的数据对设置了不同 min_weight_fraction_leaf 值的决策树回归模型进行训练，得到的均方误差值。当 min_weight_fraction_leaf 值设置为 0 时，代表不使用权重。从图 5 中的趋势可以看出，当该参数值设置为 0.01 时，均方误差达到最小。因此，本文将该参数的值设置为 0.01。

本文从研究总体中选择 2013 年的数据作为训练集，将 2014～2017 年的数据作为测试集。将预测值与真实值进行比较，并计算平均绝对误差（*MAE*）、均方误差（*MSE*）、中值绝对误差（*MDAE*）、可解释方差值（*EVS*）和 R 方值（R^2），与进行过数据标准化和 PCA 处理的 SVR 算法进行比较，实验结果见表 2。

表 2　模型评价

Tab. 2　Model evaluation

算法	SVR	PDTR
MAE	9.365 675 621 612 368	2.463 487 274 193 553 4
MSE	213.755 013 835 748 54	12.956 150 561 932 136
MDAE	4.967 198 900 343 163	1.354 955 000 000 018 1
EVS	0.116 545 324 240 816 1	0.938 856 577 703 883 1
R^2	0.101 701 582 821 607 68	0.938 824 728 491 319 5

由于本文进行对比分析的数据样本数量相同，因此 R^2 值可以很好地反映出本文所使用的回归模型拟合程度效果的好坏。从表 2 可以看出，本文提出的算法与 SVR 相比，平均绝对误差、均方误差、中值绝对误差的值更接近于 0，可解释方差和 R 方值更接近于 1，证明 PDTR 模型性能良好。从图 6～图 9 可看出，模型对 2014～2017 这 4 年预测的结果变化趋势与真实值近乎相同。

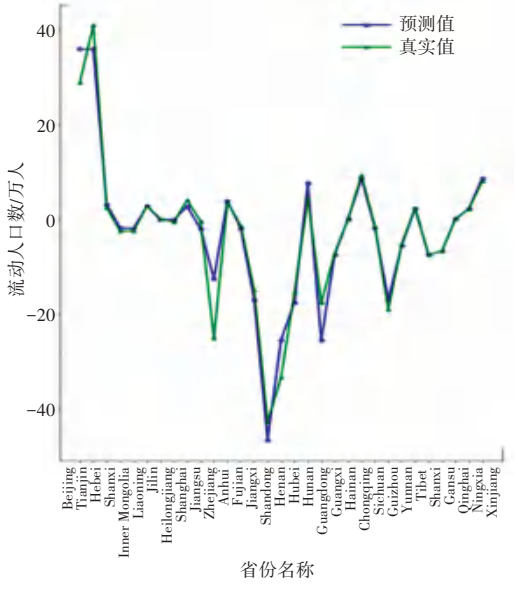


图 6　2014 年对比图

Fig. 6　Comparison chart of the data in 2014

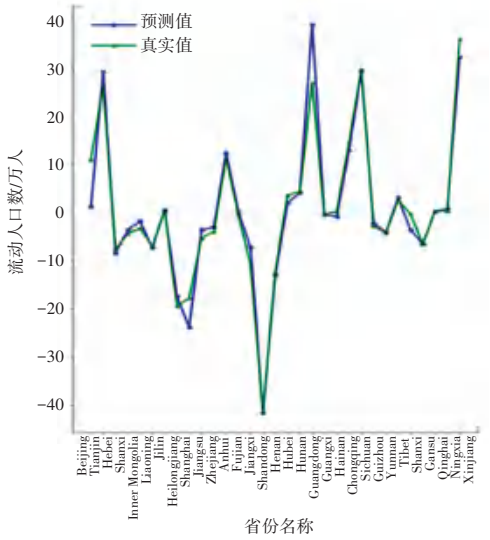


图 7 2015 年对比图

Fig. 7 Comparison chart of the data in 2015

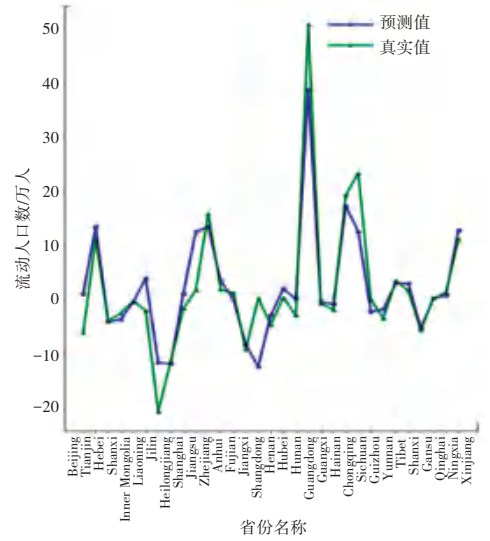


图 8 2016 年对比图

Fig. 8 Comparison chart of the data in 2016

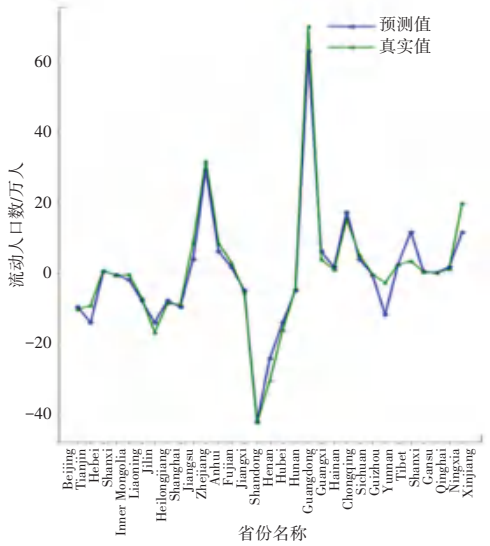


图 9 2017 年对比图

Fig. 9 Comparison chart of the data in 2017

4 结束语

本文探究了国内各城市人口流动情况与城市参数之间的关系，选取了 6 种城市参数（燃气、供水、供热、城市市容、公共交通、绿地园林）共 45 项指标。由于不是所有指标都对人口流动有影响，因此对原始数据进行特征提取，选出有效特征进而进行数据分析。本文提出的 PDTR 模型通过实验分析，可以很好地解决上述问题。本文采用主成分分析方法（PCA）进行特征提取。该方法可以在很大程度不损失数据信息的条件下，对原始高维度数据进行降维，即通过变换映射到低维空间中。通过实验验证，PCA 的 n_components 超参数为 24，对原始数据进行去燥和降维；Decision Tree Regression 的 max_depth 超参数为 14、min_weight_fraction_leaf 超参数为 0.01 时，可以很好的进行预测，得到的结果相对准确。

参考文献

[1] 沈洁，张可云. 人口区域流动的一般逻辑：理论、经验及预测[J]. 社会科学辑刊，2019(5)：128-137.

[2] 徐丽丽，李洪，李劲. 基于灰色预测和径向基网络的人口预测研究[J]. 计算机科学，2019，46(S1)：431-435.

[3] 魏金萧，周步祥，唐浩，等. 综合 RapidMiner 与改进粒子群极限学习机算法的变压器故障诊断[J]. 电力系统及其自动化学报，2019，31(3)：133-138.

[4] 徐国祥，杨振建. PCA-GA-SVM 模型的构建及应用研究——沪深 300 指数预测精度实证分析[J]. 数量经济技术经济研究，2011，28(2)：135-147.

[5] 朱志洁，张宏伟，韩军，等. 基于 PCA-BP 神经网络的煤与瓦斯突出预测研究[J]. 中国安全科学学报，2013，23(4)：45-50.

[6] 何晓群. 多元统计分析[M]. 北京：中国人民大学出版社，2012：135.

[7] 程渊. 基于集成学习的土壤含水量遥感反演研究[D]. 成都：电子科技大学，2020.

[8] 黄建琼，郭文龙，李秋缘. 基于决策树的城市环境空气质量评价模型实证研究[J]. 科技和产业，2019，19(9)：104-108，160.