摘要

城市的发展对人才的吸引力两者之间是相互联系与制约的，因此需要政府决策与措施影响，使城市增强对人才的吸引力。本文结合武汉的发展特点，利用**爬取**后的数据较全面的选择合适的指标进，通过**因子分析**的方法为武汉市不同年份人才吸引力打分；再对人才进行产业类别分类，通过**灰色关联分析**筛选出合适的因子后再横向比较各城市综合得分，并结合当地人才政策评价对人才吸引力的水平提出合理化建议。

针对问题一，通过**因子分析模型**分析武汉 2013—2017 年相关数据，对武汉各年的人才吸引力进行**量化打分**。本组首先根据题目要求和以往人才吸引要素研究，确立要素框架并**爬取**合适的人才吸引力因素。在标准化数据通过 **KMO 检验**后，利用 SPSS 对各个因素进行**主成分分析**以确定公因子数目，求解旋转后因子荷载矩阵并对公因子命名，计算武汉各年人才吸引力综合得分。综合得分结果为**-77.99**、**-30.08**、**-11.90**、**42.35**、**77.63**。综合得分表明武汉 2013—2017 人才吸引力逐年上升，且 2015—2016 年上升幅度最大，与城市实际发展情况相符合。

针对问题二，本文在问题一模型的基础上进行改进，将人才分为第一、第二和第三产业人才并用**灰色关联分析**筛选出每类人才相应的人才吸引力因素。然后将筛选后的数据用 SPSS 进行**因子分析**得到成都、天津、西安、南京和武汉 2013—2017 年的三类人才吸引力因子分析模型，并计算每个城市各年针对三类人才的人才吸引力**综合得分**。横向比较各个城市的综合得分，相较与其他城市，武汉市近几年对第二、三产业人才的吸引力有较高水平，相对于西安和成都，武汉市具有较强的人才吸引力，符合地区的经济发展特点。

针对问题三, 基于问题一中的武汉市人才吸引力评价模型以及人才政策对人才吸引力的量化评价, 并结合问题二中武汉相较于其他四个同类城市在人才吸引力上的优势与不足，给武汉市人力资源管理部门的领导写一篇建议报告。

本文中所提到的模型优点主要有两点：一、选取的指标从多维度、全方面考虑，收集的数据真实可靠；二、利用灰色关联度筛选出三类人才的人才吸引力要素，相较于人工选取指标具有客观性、严谨性。

关键词： **因子分析模型　主成分分析　灰色关联分析　数据库爬虫**

目录

一、 问题重述 3

1.1 问题背景 3

1.2 问题提出 3

二、 模型假设 3

三、 符号说明 4

四、 问题一模型的建立与求解 5

4.1 问题的描述与分析 5

4.2 模型的建立 5

4.2.1 图像基本参数定义 5

4.2.2 轮廓特征因素 5

4.2.3 边缘特征因素 6

4.3 模型的求解 6

五、 问题二模型的建立与求解 6

5.1 问题的描述与分析 6

5.2 模型的建立与求解 7

5.2.1 DNN 网络搭建 7

5.2.2 PSO-DNN 网络初始值确定 8

5.2.3 参数的选取与确定 9

5.3 实验结果与分析 9

六、 写给武汉市人力资源部门的建议报告 10

七、 模型的评价 10

7.1 模型的优点 10

7.2 模型的缺点 10

附录 A 代码 12

A.1 数据预处理–python 源代码 12

一、 问题重述

1.1 问题背景

植物的种类繁多, 要了解和掌握如此多的植物, 必须进行一个科学的分类。人们常常根据植物的用途，或根据植物的一个或几个明显的形态进行分类，植物的识别与分类对于区分植物种类, 探索植物间的亲缘关系, 阐明植物系统的进化规律具有重要的意义。因此植物分类学是植物科学甚至整个生命科学的基础学科。目前对于树叶识别与分类主要由人完成, 但是树叶种类庞大, 依赖人工地进行树叶识别与分类是不现实的。所以树叶的研究对于植物总体的研究能提供很大的帮助。

从树叶的各个方面, 纹理, 硬度, 离心率等方面都可作为主要方向研究, 现对树叶的研究主要通过采集树叶图形, 利用数字图像处理来对树叶进行分类识别, 这种方法只停留在处理形态特征, 有很大的局限性, 忽略了树叶的生理特征和其他特征, 所以研究方法来综合处理树叶平面图像特征, 形态特征和生理特征很有必要性。

1.2 问题提出

围绕植物分类进行树叶识别与分类, 以树叶二值化图片为依据, 依次提出以下问题:

(1) 结合附件中的二值化的图片数据, 建立合适的图片数据提取方案, 量化处理图片数据, 并具体分析说明所提取数据信息的量化指标体系。

(2) 基于问题一中提取的数据信息, 建立合适的数学模型由数据出发判断叶子的种类, 研究判别分类的核心指标, 并估计出模型的性能以及核心指标对模型判别性能的影响。

(3) 基于二值化图片数据, 结合附件中叶子纹理的数据信息, 对原有模型进行改进, 并对新旧模型进行比较分析。

二、 模型假设

(1) 假设相同的行业在不同的城市里吸引力影响因素相同, 每个行业在不同城市里的发展模式近似相同。

(2) 人才会考虑未来一段时间内自身对于发展前景、收入、环境的需求的变化。

(3) 政策对人才吸引力的影响转换成三个影响因素——地方财政支出与收入和固定资产投资总额。

(4) 人才的迁移是在追求效用最大化, 人才的行为仅受到城市因子的影响, 忽略人的非理性行为。

三、 符号说明

符号	说明
F	城市各年人才吸引力综合得分
F_1	工业发展与薪酬因子
F_2	医疗卫生环境因子
F_3	经济贸易因子
F_4	拥挤程度因子
X	原始指标
$\overline{X}$	指标平均值
$\widetilde{X}$	同向化指标
δ_X	指标标准差
Z	标准化指标
R	相关系数矩阵
λ_p	相关系数矩阵特征值
η_p	标准正交化特征向量
Λ	因子载荷矩阵
σ_i	方差
α_{ij}	载荷因子
k	两极最小差
K	两极最大差
$\Delta_i(t)$	特征序列与因素序列的序列差

四、 问题一模型的建立与求解

4.1 问题的描述与分析

针对问题一，本题要求建立合适方案提取二值化图片中的数据，并对所提取数据的量化指标进行分析说明。本组通过解析几何计算和时间序列展开，将目标图像转化成两个特征向量。本组根据题目要求和近年的图像识别研究，确定了针对二值化图像的两个重点识别因素——轮廓特征因素和边缘特征因素。针对轮廓特征因素，首先利用 matlab 做解析几何运算，计算与图像轮廓有关的特征量，并将计算结果作为元素，组成轮廓特征向量。针对边缘特征因素，首先将图像边缘通过极化投影展开为时间序列，计算每支时间序列的特征量，并将计算结果作为元素，组成边缘特征向量。最后合并两个向量得到总体特征向量。

4.2 模型的建立

4.2.1 图像基本参数定义

定义 I 表示目标树叶所对应的图像，∂I 表示图像边界，$D(I)$ 表示图像最小外接圆直径，$d(I)$ 表示图像最大内切圆半径，$A(I)$ 表示研究对象面积,$L(\partial I)$ 表示研究对象的轮廓线长度，$H(I)$ 表示研究对象的凸包域，$C(I)$ 为图像几何中心坐标，运算符 $d(.)$ 代表欧式距离。

4.2.2 轮廓特征因素

定义轮廓特征向量为 $ID1_1 = [id_1, id_2, \cdots, id_n](n = 6)$ 其中 $id_k(k = 1, 2, \cdots, 6)$ 为二值化矩阵的轮廓特征，其具体计算公式如下：

id_1 长宽比 (Aspect Ratio): 定义 X_I 为图像最上方非零行与最下方非零行的行数差 (长)，Y_I 为图像最左方非零列与最右方非零列的列数差 (宽)，长宽比 $id_1 = X_I/Y_I$。

id_2 离心率 (Eccentricity): 定义 $E(I)$ 是与研究图像具有相同的二阶矩的椭圆,a 和 b 分别为 $E(I)$ 对应的长轴与短轴，离心率 $id_2 = \sqrt{1 - (\frac{b}{a})^2}$，变化范围为 (0,1)。

id_3 密实度 (Solidity): 实密度 $id_3 = \frac{A}{A(H(I))}$。其反映研究对象的仿射特征, 即研究对象区域的固靠性程度。

id_4 等周因子 (Isoperimetric Factor): 等周因子 $id_4 = \frac{4\pi \cdot A}{L(\partial I)^2}$ ，其变化范围为 (0,1)，描述目标树叶轮廓规整度，叶子边缘越规则, 其值越接近于最大值 1。

id_5 伸长率 (Elongation): 伸长率 $id_5 = 1 - \frac{2d_I}{D(I)}$, 变化范围为 (0,1), 树叶越趋于圆形, 相应的伸长率越小, 比率 $\frac{2d_{max}}{D(I)}$ 其描述目标树叶轮廓向外伸展的趋势。□

id_6 最大压痕深度 (MaximalIndentationDepth): 定义 $C_{H(I)}$ 为研究对象凸型区域的几何中心,$L(\partial I)$ 表示为 $H(I)$ 的轮廓线长度,$\forall X \in H(I)$ 和 $\forall Y \in \partial I$, 计算距离 $d(X, C_{H(I)})$

和 $d(Y, C_{H(I)})$。定义函数 $\frac{d(X,C_{H(I)})-d(Y,C_{H(I)})}{L(H(I))}$, 该函数的最大值即最大压痕深度 id_6。

4.2.3 边缘特征因素

为得到研究对象的边缘特征，首先以 $C(I)$ 为坐标原点建立笛卡尔坐标系，对于曲线 ∂I 上任意一点 P 可以在该坐标系下表示为 $P(x_p, y_p)$，将其投影至以 $C(I)$ 为极点的极坐标系得 $P'(r_p, \theta_p)$，其中：

$$r_p = d(P(x_p, y_p).C(I)) \tag{1}$$

$$\theta_p = y_p/x_p, \tag{2}$$

(图 1) 通过将点集 $P'(r_p, \theta_p)$ 降维可得到时间序列 $arrays(P')$。将 arrays 滤波处理后，计算每支时间序列上的极大值点数 id_7，极小值点数 id_8。得到边缘特征向量 $ID2 = [id_7, id_8]$，合并轮廓特征向量与边缘特征向量得：

$$ID = [ID1, ID2] \tag{3}$$

其中 ID 为总体特征向量。

4.3 模型的求解

首先使用 matlab 围绕 regionprops 函数对研究对象进行解析几何计算，用遍历式算法逐一算出研究对象长宽比、离心率、实密度、等周因子、伸长率和最大压痕深度，得到轮廓特征向量 (见代码？)。再将图片通过 numpy 工具箱标准正立化，以其几何中心为极点将其边缘坐标转换为极坐标，然后使用 ndarrays 函数将所得极坐标降维展开成时间序列，搜寻滤波后时间序列的极点数得到边缘特征向量。

五、 问题二模型的建立与求解

5.1 问题的描述与分析

问题二要求针对所提取数据信息，建立数学模型判别叶子的种类并研究其核心指标。本组通过问题一中所提取的 11 维度的特征，采用 $Keras$ 工具库搭建深度神经网络（DNN）解决树叶多分类问题，并输出各维度特征的权重作为指标贡献率。DNN 神经网络采用 BP 算法基于梯度信息来调整连接权值，因而初始权值和阈值选取的随机性会导致网络稳定性差。为克服此缺点，引入收敛速度快、全局搜索能力强的 PSO 算法来优化 DNN 网络的连接权值和阈值，建立一种新的 $PSO-DNN$ 网络模型实现对树叶分类。当考虑到步长等参数的选取时，使用网格搜索的方式对参数进行优化，其流程图如下：

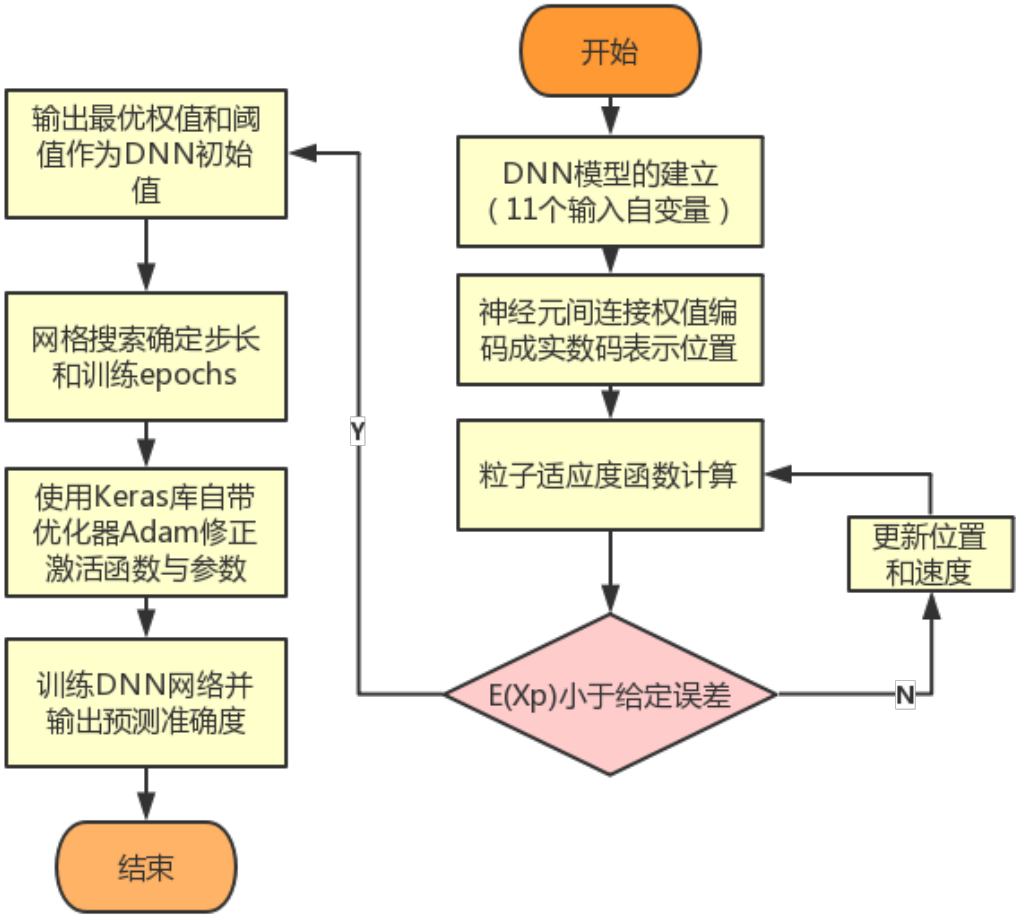

图 1　**PSO** 算法优化 **DNN** 网络流程图

5.2 模型的建立与求解

5.2.1 DNN 网络搭建

DNN 的正向传播主要依靠众多神经元的计算来完成的，此处设计的网络结构如图 2 所示，工作过程中可以用下式来表示：

$$u_k = \sum_{i=1} w_{ki} x_i \tag{4}$$

$$y_k = f\left(u_k - b_k\right) \tag{5}$$

其中:x_i 表示第 i 个输入; w_{ki} 表示与第 i 输入量相连的权值; u_k 表示所有输入的加权和; b_k 为神经元阈值;f 为激活函数; y_k 为神经网络的输出。

激活函数的种类有很多，如 $sigmoid$，$tanh$ 及 $Relu$，本文应用的是 $ReLU$ 作为激

活函数搭建两层隐含层，如式 6 所示：

$$f_{\text{Relu}} = \max(0, z)$$
$$\frac{d}{dz} f_{ReLU} = \begin{cases} 1, & z > 0 \\ 0, & z \leqslant 0 \end{cases} \tag{6}$$

网络的输出层使用 $softmax$ 函数作分类器，式 7 为第 i 个神经的输出：

$$f_{\text{softinax}} = \mathrm{e}^{i} / \sum_{j} \mathrm{e}^{j} \tag{7}$$

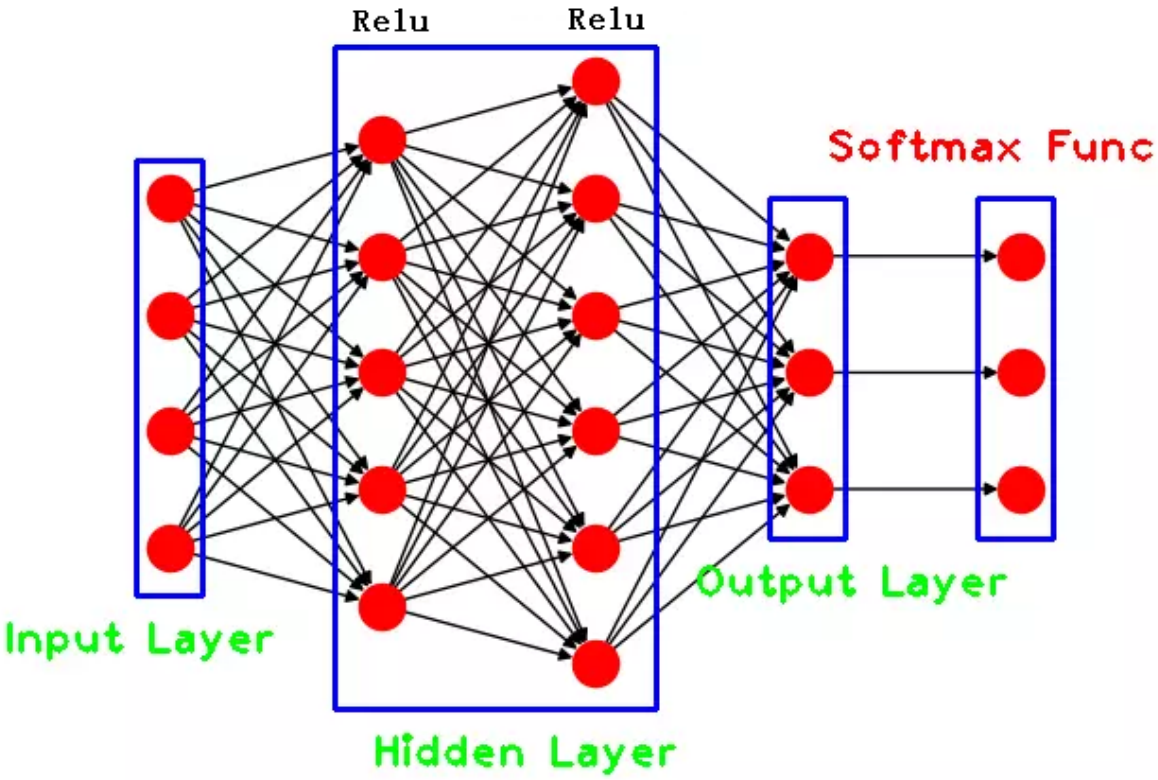

图 **2**　人工搭建 **DNN** 结构示意图

5.2.2 PSO-DNN 网络初始值确定

算法的具体实现步骤如下：

(1) 将 DNN 网络结构中所有神经元间的连接权值和各个神经元的阈值编码成实数码串表示的个体作为 PSO 算法要寻优的位置向量。

(2) 在编码空间中随机生成一定数目的个体组成种群，其不同个体代表神经网络不同权值。

(3) DNN

5.2.3 参数的选取与确定

由于训练 epochs 与步长选取较为困难，实验采取网格搜索法（Grid Search）来确定参数的选取，网格搜索是指定参数值的一种穷举搜索方法，通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。

在迭代修正的过程中，实验采用 $Keras$ 自带优化器 $Adam$ 函数进行迭代修正，如式 8 所示，其中 β_1 一般取 0.9，β_2 一般取 0.999。

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_w f(w_t) \\ v_t = \beta_2 v_{t-1} + (1-\beta_2)\nabla_w f(w_t)^2 \\ \widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t} \\ w_t = w_{t-1} - \eta \frac{\widehat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} \end{cases} \tag{8}$$

5.3 实验结果与分析

随机选取 80% 的数据作为训练集，训练输入样本为 $1280 * 11$，测试输入样本为 $320 * 11$。DNN 网络输入层神经元 11 个，输出层神经元 100 个，包含 2 个隐含层，适应度函数选择均方误差，取 $c_1 = c_2 = 2$，选择粒子个数为 30，最大迭代次数 500 次。按照 PSO 优化 DNN 网络模型的步骤不断地迭代寻找最优网络参数，进行树叶分类预测的仿真实验。其迭代过程中损失与准确率如下图所示：

图 **3**　迭代损失与准确率

六、 写给武汉市人力资源部门的建议报告

七、 模型的评价

7.1 模型的优点

(1) 从国家统计数据库、武汉市统计年鉴和其他同类城市的统计年鉴中爬取大量数据，并挑选符合实际的 33 个指标与最近五年的数据进行评价。其吸引人才指标从多维度、全方面的考虑，具备科学性、客观性。

(2) 针对具体人才根据相应产业进行分类。相较于现有城市人才吸引力水平评价模型，我们采用了灰色关联分析法，对三种人才类别分析计算出与指标的相关系数，求出不同人才对 33 种影响因素的的偏好程度，选出前十个合适的指标进行综合打分。

7.2 模型的缺点

未能具体地量化政府政策对人口吸引力的影响，只选取了地方财政收入、支出，固定资产投资总额三个影响因素作为政策影响的指标，与现实情况出现一定偏差，存在一定的局限性。

参考文献

[1] 张培利宋鸿. 城市人才吸引力的影响因素及提升对策. 湖北社會科學, 2010(2):43–45, 2010.

[2] 武汉统计局. http://tjj.wuhan.gov.cn/newslist.aspx?id=2018061817071264. Website. 武汉市国民经济和社会发展统计公报.

[3] Meric S Gertler, Kate Geddie, Carolyn Hatch, and Josephine Rekers. Attracting and retaining talent: Evidence from canada’s city-regions. *Seeking talent for creative cities: The social dynamics of innovation*, pages 3–30, 2014.

[4] 张瑞红. 河南省产业集群环境人才吸引力评价研究. 科技管理研究, 32(10):180–184, 2012.

[5] Josh Lepawsky, Chrystal Phan, and Rob Greenwood. Metropolis on the margins: talent attraction and retention to the st. john’s city-region. *The Canadian Geographer/Le Géographe canadien*, 54(3):324–346, 2010.

[6] EPS 数据平台. http://olap.epsnet.com.cn/auth/platform.html. Website. 中国宏观经济数据库.

附录 A 代码

A.1 数据预处理–python 源代码

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from factor_analyzer import FactorAnalyzer

#导入数据
df = pd.read_excel("wut.xls")
# print(df.head(5))

#转置
df = pd.DataFrame(df.values.T, index=df.columns, columns=df.index)
# print(df['商品房平均销售价格（元/平方米）'])

数据正向处理
df['商品房平均销售价格（元/平方米）'] =
    -(df['商品房平均销售价格（元/平方米）']-sum(df['商品房平均销售价格（元/平方米）'])/5)
df['工业废水排放量（万吨）'] = -(df['工业废水排放量（万吨）']-sum(df['工业废水排放量（万吨）'])/5)
df['工业二氧化硫排放量（吨）'] =
    -(df['工业二氧化硫排放量（吨）']-sum(df['工业二氧化硫排放量（吨）'])/5)
print(df['商品房平均销售价格（元/平方米）'])
print(df.head(5))
df.to_excel("武汉.xls")

#数据标准化
df = StandardScaler().fit_transform(df)
# print(df)
df=pd.DataFrame(df)
df.to_excel("武汉.xls")
#
#主成分分析
pca = PCA(n_components=4)
newX = pca.fit_transform(df)
print(newX)

#返回所保留的n个成分各自的方差百分比
print("它代表降维后的各主成分的方差值占总方差值的比例，这个比例越大，则越是重要的主成分")
print(pca.explained_variance_ratio_)
print("它代表降维后的各主成分的方差值，方差值越大，则说明越是重要的主成分")
print(pca.explained_variance_)
```