摘要

植物的种类繁多, 对植物进行科学的分类至关重要。叶片的识别与分类是识别植物的重要组成部分。本文基于 100 种植物树叶的二值化图片数据, 通过**解析几何计算**和**时间序列展开**的方法转换成**轮廓特征向量**和**边缘特征向量**; 再通过**深度神经网络 (DNN)** 解决树叶识别分类问题, 引入**粒子群 (PSO)** 算法优化网络中的连接权值与阈值, 并结合树叶纹理数据信息对模型进行改进和分析比较。

针对问题一，通过**解析几何计算**和**时间序列展开**，分别提取每一张图片中的特征向量。根据题目要求，本组首先利用 matlab 计算与图像形状相关的解析几何特征量，得到由八个几何学、拓扑学特征值组成的**八维**形状特征向量。然后将图像轮廓进行极化投影，并通过 numpy 工具箱将极化投影轮廓展开成时间序列，挖掘分析时间序列的三个特征值，组成**三维**边缘特征向量。最后合并两个向量得到**十一维**总体特征向量。

针对问题二，本文

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

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) 不考虑叶片的枝干和叶柄的数据特征

(2) 假设二值化的树叶叶片数据没有残损卷曲。

(3) 为保证分类方法具有可推广性，假设现实中同种类树叶与题目所给的树叶二值化图像具有相似的形状轮廓。

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——形状特征因素和边缘特征因素。针对形状特征因素，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。其具体特征值如图 1 所示：

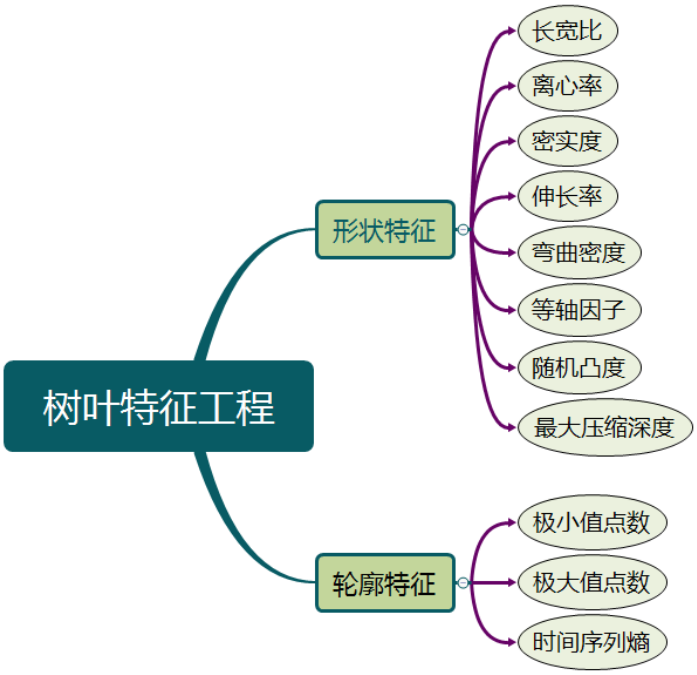


图 **1**　图形特征示意图

4.2 模型的建立

4.2.1 图像基本参数与运算符号定义

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 形状特征因素

定义轮廓特征向量为 $ID_{shape} = [id_1, id_2, \cdots, id_n](n = 6)$ 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：

(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)，即长宽比计算公式：

$$id_1 = X_I / Y_I \tag{1}$$

(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，即离心率计算公式：

$$id_2 = \sqrt{1 - (\frac{b}{a})^2} \tag{2}$$

(3) 密实度 (Solidity): 反映研究对象的仿射特征的变量, 即研究对象区域的固靠性程度，计算公式：

$$id_3 = \frac{A}{A(H(I))} \tag{3}$$

(4) 等周因子 (Isoperimetric Factor): 描述目标树叶轮廓规整度的变量，变化范围为 (0,1)，叶子边缘越规则, 其值越接近于最大值 1，计算公式：

$$id_4 = \frac{4\pi \cdot A}{L(\partial I)^2} \tag{4}$$

(5) 伸长率 (Elongation): 描述研究对象向某方向伸展趋势的变量, 变化范围为 (0,1), 树叶越趋于圆形相应的伸长率越小，计算公式：

$$id_5 = 1 - \frac{2d_I}{D(I)} \tag{5}$$

□

(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 $d(Y, C_{H(I)})$。即最大压痕深度计算公式为：

$$max\left\{\frac{d(X, C_{H(I)}) - d(Y, C_{H(I)})}{L(H(I))}\right\} \tag{6}$$

(7) 随机凸性 (Stochastic Convexity): 记随机给定两个端点 $P_1, P2 \in \partial I$，记 $P(G)$ 为线段 $[XY]$ 完全包含于图像区域 I 中的概率，即随机凸性计算公式为：

$$id_7 = P(G) \tag{7}$$

(8) 弯曲能量 (Bending Energy): 描述研究对象边界弯曲程度的值，定义 φ_n 为 ∂I 的曲率，即弯曲能量计算公式为：

$$id_8 = 1/P \sum_{i=0}^{n-1} |\varphi_n - \varphi_{n-1}| \tag{8}$$

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：

$$r_p = d(P(x_p, y_p).C(I)) \tag{9}$$

$$\theta_p = y_p / x_p, \tag{10}$$

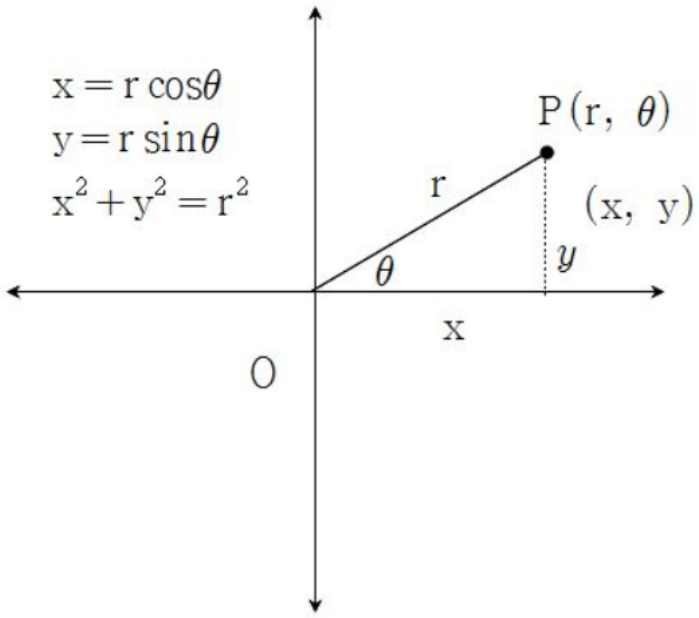


图 **2**　极化投影实意图

定义点集 $P'(r_p, \theta_p)$ 为研究样本，其中中包含 d' 个一元时间序列，Q 为单个样本一维的长度。将数据集中每个样本第 j 维 $(j \in \{1, 2, \cdots, d\})$ 数据组成一个一元时间序列数据集，记为 X。且其中表类样本 $y_i \in \{1, 2, \cdots, C\}, i \in \{1, \cdots, N\}$。

得到一元时间序列集 X 后，记 N 为序列集中的时间序列条数。在数据集 D 中有 n_i 个实例，并且 $n_1 + n_2, \cdots, n_n = N$ 计算 X 序列集熵值作为边缘特征值：

$$id_9 = -\sum_{i=1}^{C} \frac{n_i}{N} \log \frac{n_i}{N} \tag{11}$$

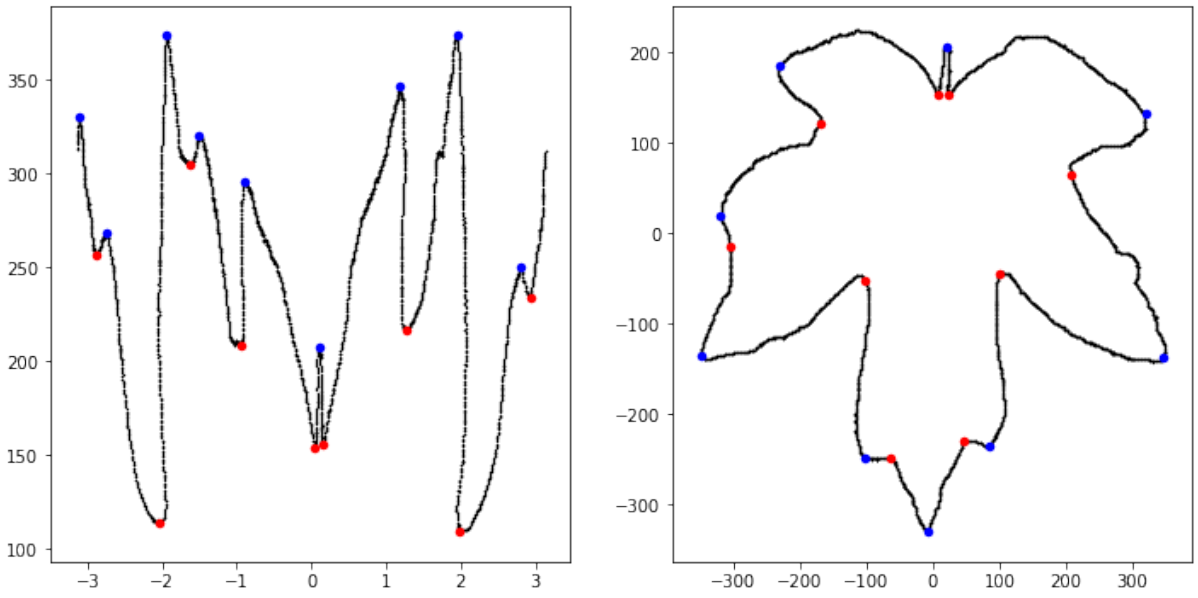
定义 $shapelet$ 为时间序列 X 中能够最大程度区分不同类别时间序列的子序列。在一元时间序列分类问题中选出 K 个最优 shapelet 并记为 $S \in R^{K \times M}$(其中 M 为 shapelet 的长度。一个长度为 M 的 shapelet

$$S_k = \{s_{k1}, s_{k2}, \cdots, s_{kn}\} (k \in \{1, \cdots, K\}) \tag{12}$$

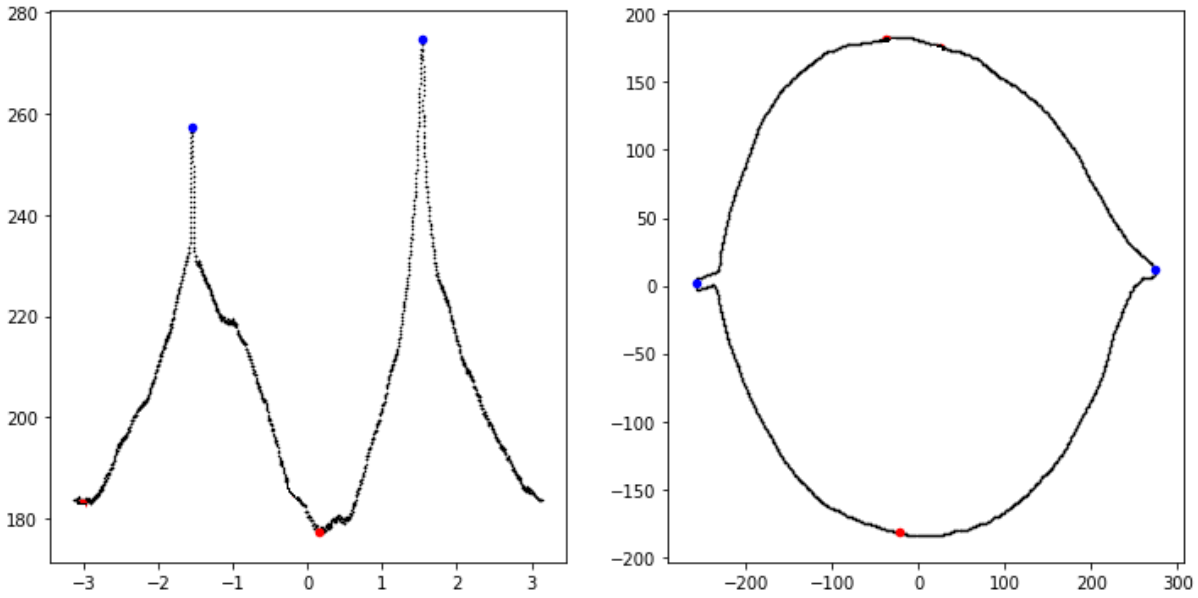
是时间序列 $X_i(i \in \{1,2,\cdots,d\})$ 的子序列，定义 X_i 与 S_k 之间的距离为：

$$D_{i,k} = \min_{j=1,\cdots,j} \frac{1}{M} \sum_{m=1}^{M} (X_{i,j+m-1} - S_{k,m})^2 \tag{13}$$

其中 M 表示所选定的 X_i 与 S_k 的子序列长度。再找到在找到 K 个最优 shapelet 后，计算 K 个最优 shapelet 与一个时间序列间的距离作为该时间序列的新特征，时间序列样本集被映射至新特征空间，最终将之间序列降至一维时间序列 arrays，展开效果如图 3 所示：



(a) Acer-Campestre



(b) Cornus-Controversa

图 **3**　时间序列展开效果图

将 arrays 滤波处理后，计算每支时间序列上的极大值点数 id_{10}，极小值点数 id_{11}。得到边缘特征向量 $ID_{margin} = [id_9, id_{10}, id_{11}]$，合并轮廓特征向量与边缘特征向量得：

$$ID = [ID_{shape}, ID_{margin}] \tag{14}$$

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。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。其算法流程图如图 4

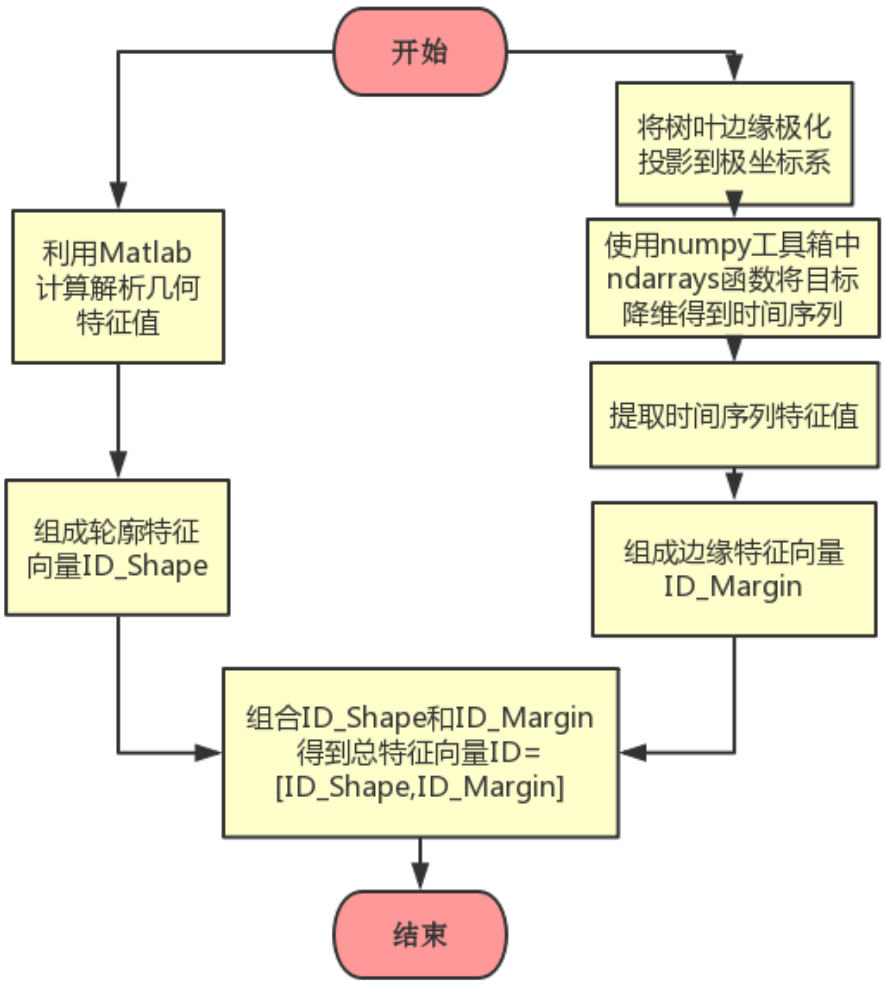


图 **4**　树叶特征工程示意图

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（Deep Neural Network）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、全局搜索能力强的粒子群 (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：

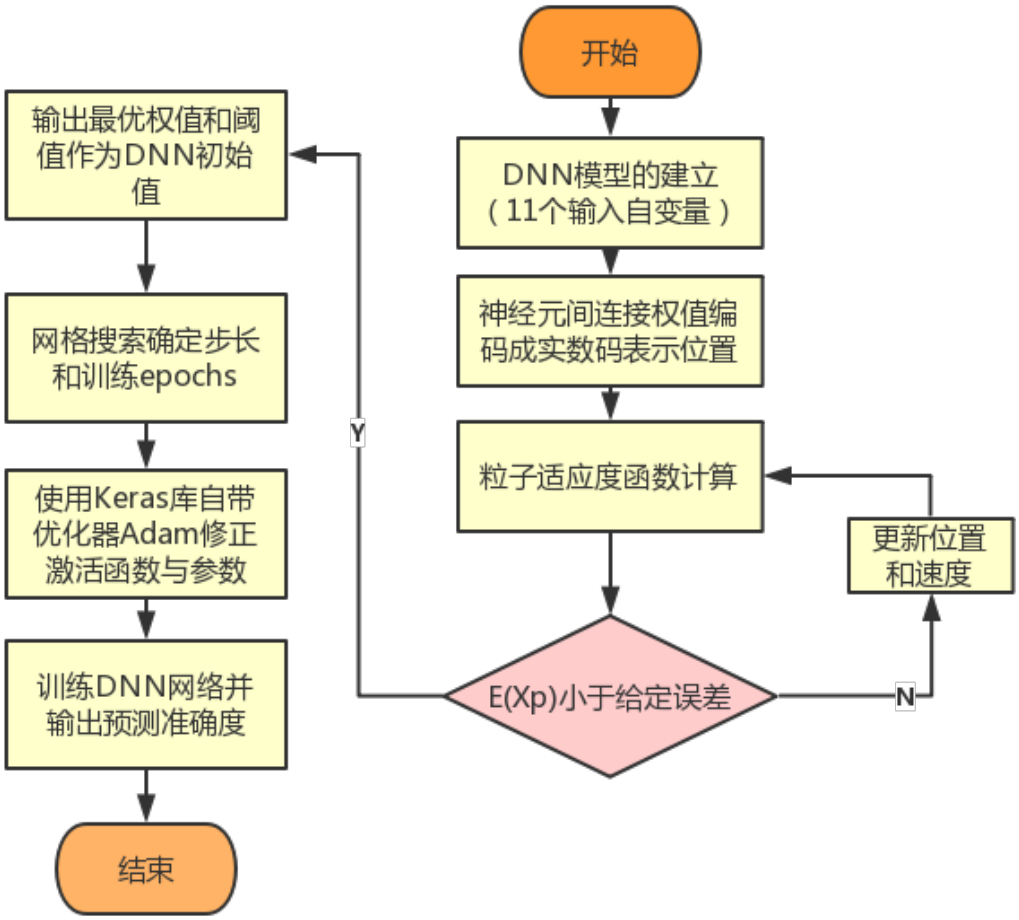


图 **5**　**PSO** 算法优化 **DNN** 网络流程图

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 6 所示，工作过程中可以用下式来表示：

$$u_k = \sum_{i=1} w_{ki} x_i \tag{15}$$

$$y_k = f\left(u_k - b_k\right) \tag{16}$$

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 17 所示：

$$\begin{gathered} f_{\text{Relu}} = \max(0, z) \\ \frac{d}{dz} f_{ReLU} = \begin{cases} 1, & z > 0 \\ 0, & z \leqslant 0 \end{cases} \end{gathered} \tag{17}$$

网络的输出层使用 $softmax$ 函数作分类器，式 18 为第 i 个神经的输出：

$$f_{\text{softinax}} = \mathrm{e}^i / \sum_j \mathrm{e}^j \tag{18}$$

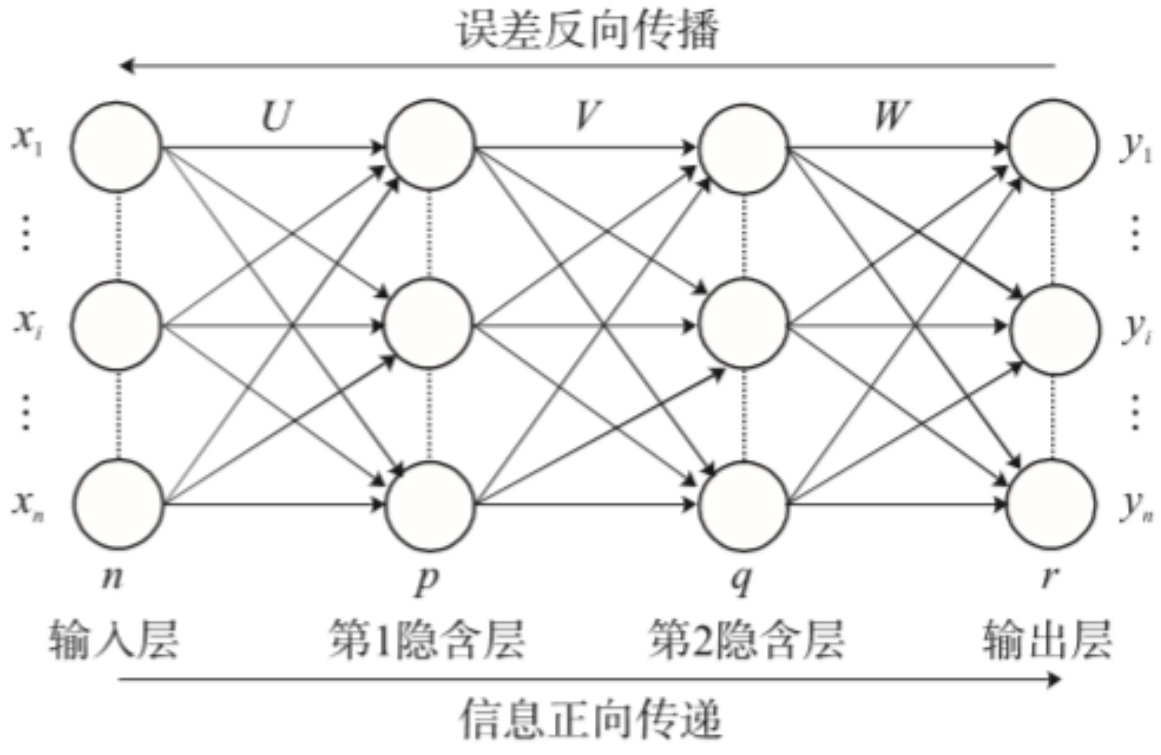


图 **6**　人工搭建 **DNN** 结构示意图

5.2.2 PSO 优化 DNN 网络

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 网络训练及个体的适应度评价。将微粒群中的每一个个体的分量映射为网络中的权值和阈值，从而构成个体对应的神经网络。首先划分训练样本和测试样本; 其次输入训练样本进行网络训练，通过反复迭代来优化网络权值，并计算每一个网络在训练集上产生的均方误差,，以此作为目标函数; 最后对每个个体进行适应度评价，从中找到最佳个体用来判断是否需要更新微粒的 Gbest 与 Pbest，构造如下的适应度函数:

$$E\left(X_p\right)=\frac{1}{n}\sum_{p=1}^{n}\sum_{k=1}^{c}\left[Y_{k,p}\left(X_p\right)-t_{k,p}\right]^2 \tag{19}$$

式中：n 为训练样本个数，c 为输出端个数，$t_{k,p}$ 为训练样本 p 在 k 端的给定输出，$Y_{k,p}\left(X_p\right)$ 为训练样本 p 在 k 端的实测值，他们俩个值的误差平方和越小，表示实际值和预测值越接近，网络的性能越好。

(4) 更新每个粒子的速度和位置，产生下一代的粒子群。更新公式如下:

$$v_{id}^{t+1}=v_{id}^{t}+c_1r_1(p_{id}^{t}-x_{id}^{t})+c_2r_2(g_{id}^{t}-x_{id}^{t}) \tag{20}$$

$$x_{id}^{t+1}=x_{id}^{t}+v_{id}^{t+1} \tag{21}$$

其中 $i=1,2,...,n;d=1,2,...,d;t$ 为当前迭代次数; v_{id}^{t} 为当前粒子速度 (t 时刻); x_{id}^{t} 为当前粒子位置 (t 时刻); $(p_{id}^{t}-x_{id}^{t})$ 为当前位置与自己最好位置之间的距离; $(g_{id}^{t}-x_{id}^{t})$ 为当前位置与群体最好位置之间的距离; v_{id}^{t+1} 为下一时刻粒子速度 ($t+1$ 时刻); c_1,c_2 为非负常数，称为加速因子； r_1,r_2 为均匀分布与 $[0,1]$ 区间的随机数。

(5) 当目标函数小于给定的误差或达到最大迭代次数时，算法结束。将 PSO 算法训练出来的最佳神经网络的权值和阈值作为 DNN 网络的初始值，并记录。

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 5.2.3 所示，其中 β_1 一般取 0.9，β_2 一般取 0.999。

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

5.3 实验与结果分析

随机选取 80% 的数据作为训练集进行数据**标准化处理**，再对树叶种类（species）进行 **One-Hot 编码**作为分类目标。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。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。具体参数设置间表 1 。

表 **1**　参数设置表

参数名称	参数符号	参数值
选择粒子个数为	T	30
计算精度	ε	0.00001
学习率	η	0.01
最大迭代次数	M	500
维度	S	11
加速因子 1	c_1	1.49
加速因子 2	c_2	1.49
惯性权重	β	0.9

通过上述参数设置，训练后各维度权重占比如图 8 所示，最小均方误差为 0.458，**预测准确率为** 91.237%。其迭代过程中损失与准确率如下图所示：

图 7　**PSO-DNN** 迭代过程中损失与准确率

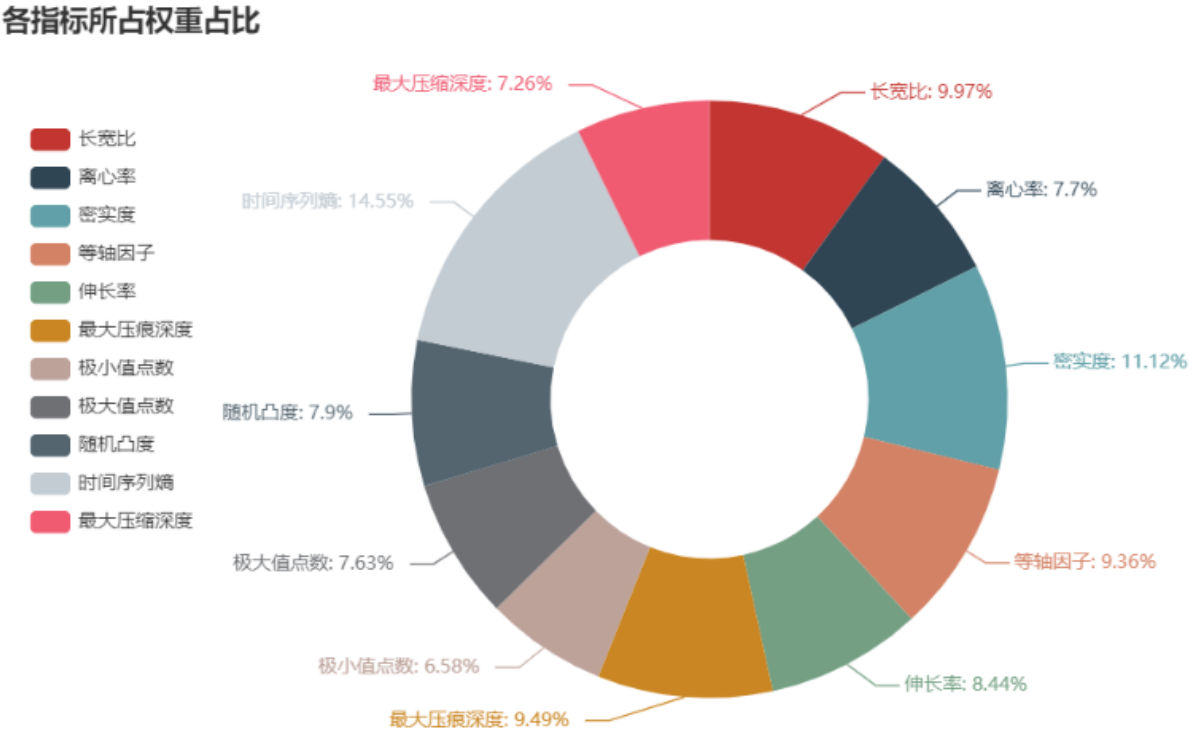


图 8　各维度权重占比图

5.3.1 模型性能评估

表 2 中，实验与三种神经网络进行纵向比较可知，四种神经网络训练后准确率都在 90% 左右，证实了神经网络非线性映射能力较强。Elman 网络和 DNN 网络的测试准确

率都大于 BP 网络，说明反馈动态网络的逼近能力要强于前反馈静态网络。使用 PSO 优化 DNN 网络后，模型的跟踪性能有所改善，测试集中损失 (Validation Loss) 仅为 0.458，说明 PSO-DNN 网络的泛化能力和稳定性明显提高，且全局收敛性增强。

表 **2**　常见神经网络的比较

网络类型	网络名称	测试集中准确率
静态网络	BP	87.424%
静态网络	Elman	90.541%
动态网络	DNN	89.912%
动态网络	PSO-DNN	91.237%

六、 问题三模型的建立与求解

6.1 问题描述与分析

针对问题三, 本题要求结合给出的叶子纹理的数据信息, 对问题二原有模型进行改进并进行分析比较。本组通过主成分分析法对树叶纹理数据信息降维, 基于问题二模型中原有的 11 个输入自变量, 把降维后的纹理信息添加成一个新的输入变量, 搭建深度神经网络 (DNN), 引入变异粒子群 (MPSO) 算法优化参数。新的 $MPSO-DNN$ 模型相较于问题二模型增加了树叶叶片的指标特征数量, 提高了叶片识别与分类的精度。

6.2 模型的建立与求解

6.3 结果分析

Epoch 498/500 1296/1296 [==============================] - 1s 689us/step - loss: 0.1006 - acc: 0.9861 - val-loss: 0.4902 - val-acc: 0.9653

七、 模型的评价

7.1 模型的优点

7.2 模型的缺点

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_)
```