**基于鸢尾花数据的分析与建模**

**（目标：基于鸢尾花数据进行数据探索和分析，并建立合理的分类模型）**

**摘要**：本文对鸢尾花数据集进行探索性数据分析，以可视化、特征相关程度分析等方式对不同的特征属性进行分析。针对该类非线性可分多分类问题，本文基于多层前馈神经网络训练了一个高效的分类器，实验结果显示该分类器准确率为98%。

1. 数据集与数据探索
   1. 数据集介绍

（介绍你所用数据集的基本信息，来源以及需要解决的问题）

鸢尾花是一种在世界各地广泛发现的开花植物，包含超过300个不同品种，每个品种表现出不同的的物理特征，如花和叶的形状和大小。鸢尾花数据集最早由Ronald Fisher在他关于判别分析的开创性著作《在分类学问题中使用多个度量值》（Fisher，1936）中提出。

花瓣是花的内部部分，颜色比较鲜明，萼片是花的外部，通常是绿色的。然而，在鸢尾花中，萼片和花瓣都是紫色的，但是可以通过形状差异彼此区分，鸢尾花属性如图1-1。

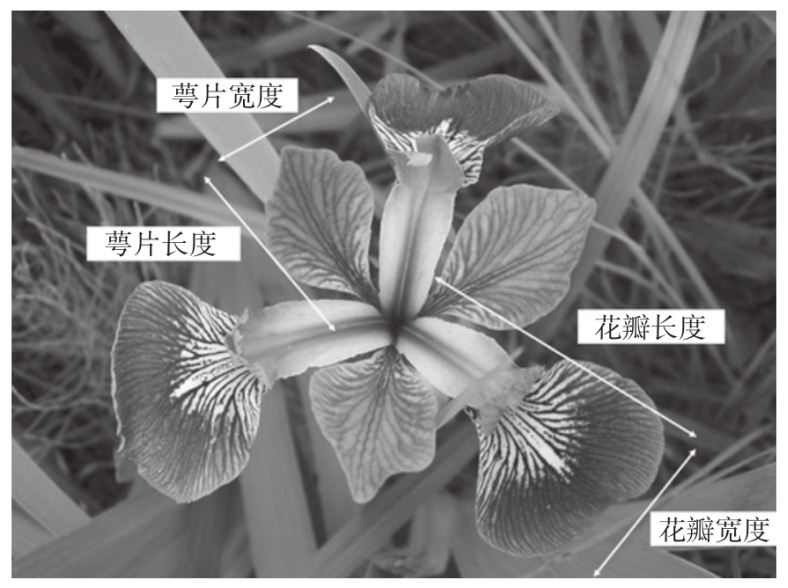


图 1.1：鸢尾花属性示意图

Iris数据集包含三个不同种类（Iris setosa、Iris virginica和Iris versicolor）的150个观察数据，每种有50个观察数据。每个观察数据包括四个属性：萼片长度（Sepal Length）、萼片宽度（Sepal Width）、花瓣长度（Petal Length）和花瓣宽度（Petal Width）。第五个属性是观察到的物种的名称，分别为Iris setosa、Iris virginica和Iris versicolor。

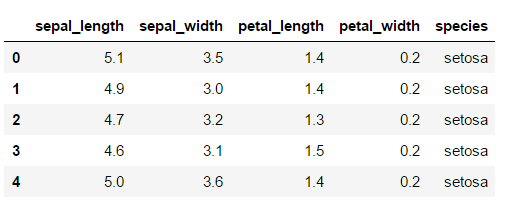


图 1.2：鸢尾花数据集

1.2 数据探索

本文从单变量的一次数据调查图表开始视觉探索，显示属性值如分布形状。为了获得更多的观测信息，本文以带有类型标签的增强柱状图显示数据集由三个不同分布组成，如图1.3为三种鸢尾花的花瓣长度属性分布。Iris setosa的分布在1至2cm范围内1.25的周边，Iris versicolor和Iris virginica的分布与Iris setosa有交叉和分离。从分析可知，Iris setosa在属性分布上与另外两种差异明显，可以明确的区分出来，但是Iris versicolor 和 Iris virginica 的交叉部分导致分辨有些困难。

（相同方法可以用来探索其他不同属性的分布情况）

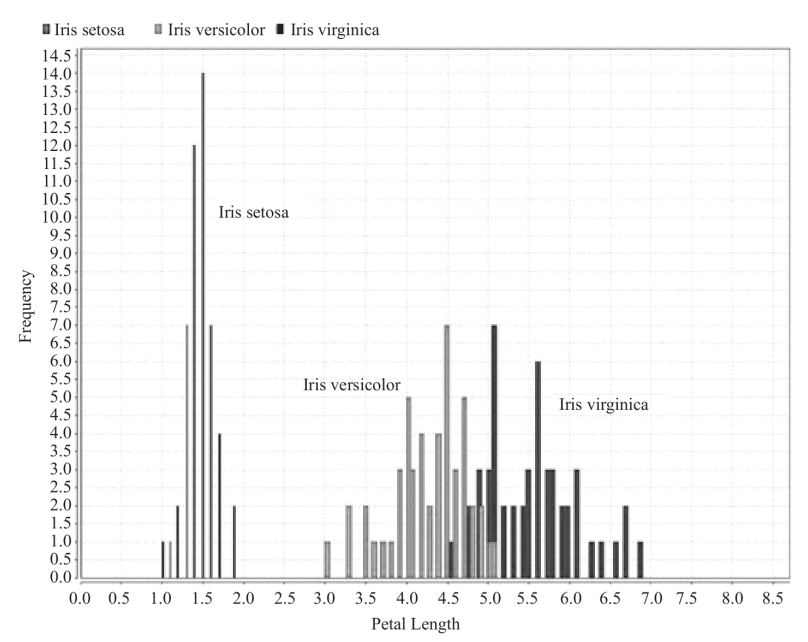
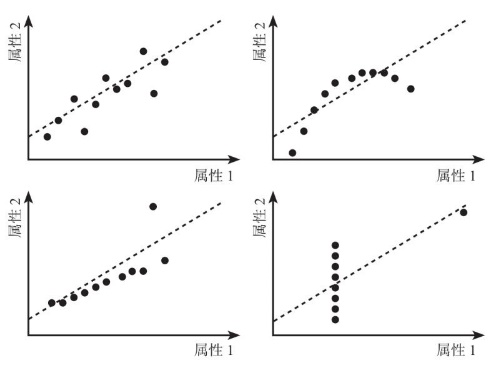


图1.3：三种鸢尾花的花瓣长度属性分布

（属性相关性分析，分析不同属性之间的相关性，作为特征选择的依据）

（）

将样本中的4个特征两两组合（任选2个特征分别作为横轴和纵轴，用不同的颜色标记不同品种的花），可以构建12种组合（其实只有6种，另外6种与之对称），如图 1.4所示：

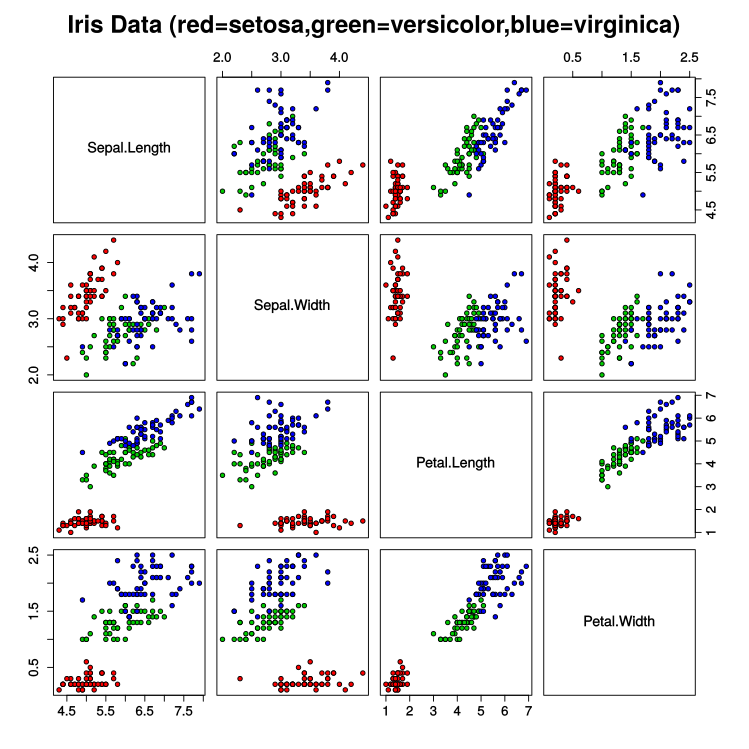


图 1.4：特征组合后的二维可视化

1. 数据建模

(在数据建模阶段，你可以利用数据探索的有用信息，基于某些算法设计合理的模型，这一部分需要写明你的设计细节以及创新点。针对你选定的具体数据集和不同应用场景（比如生物数据、交易数据、交通数据等等），设计细节可以是你具体的算法流程和模型结构 。实验结果、实验分析等内容放到单独的一节)

本文需要针对鸢尾花数据建立一个有效的分类模型，经过数据探索分析可知，鸢尾花数据是一个多分类问题，而且不是线性可分的问题。因此，本文基于神经网络模型进行数据建模，构建鸢尾花种类判别器。

神经元模型是神经网络中的基本组成部分，[McCulloch and Pitts,1943]将神经元的工作过程抽象为一个简单的模型，该模型被称为“M-P神经元模型”，模型结构如图2.1，其中为来自第个神经元的输入，是第个神经元的连接权重，为神经元阀值，为经过激活函数映射的输出，其计算公式为：

|  |  |  |
| --- | --- | --- |
|  |  | （2.1） |

其中为神经元中的激活函数，目前的神经网络模型通常使用非线性函数做激活函数。

…

…

…

…

图2.1 M-P神经元模型

具有单层功能神经元的神经网络模型只能解决线性可分的分类问题，但鸢尾花分类任务是一个复杂的非线性可分的分类问题，要解决此类问题，则需要考虑使用多层功能神经元。常见的神经网络模型具有层级结构，包括输入层、输出层以及两者之间的若干隐含层。每一层中包含若干个神经元，但输入层的神经元只接收数据，不进行激活函数映射，而隐含层和输出层的神经元都是含有激活函数的功能神经元，这种神经网络模型被称为“多层前馈神经网络”，其结构如图2.2。

…

…

…

…

…

…

图2.2 多层前馈神经网络

本文采用简单的三层全连接神经网络来建模，输入层有4个神经元，接收鸢尾花数据里的四个特征，模型只有一个隐含层，隐含层神经元数量为6，输出层为3个神经元，对应样本属于3种类别的概率。

三、实验结果和分析

（展示你所做实验的内容，包括训练过程、分类测试结果，比如下面展示为训练好的神经网络产生的决策边界，这里是二维的展示）

（鸢尾花数据是一个比较简单的例子，在比较复杂的数据分析建模问题中，你应该对实验结果做出具体的解释）

（比如，你可以在建模阶段选用不同的方法，例如用支持向量机或者聚类方法，在这里展示不同方法的实验结果，并分析和对比实验结果，说明哪些地方好，哪些地方不好，以及展示）

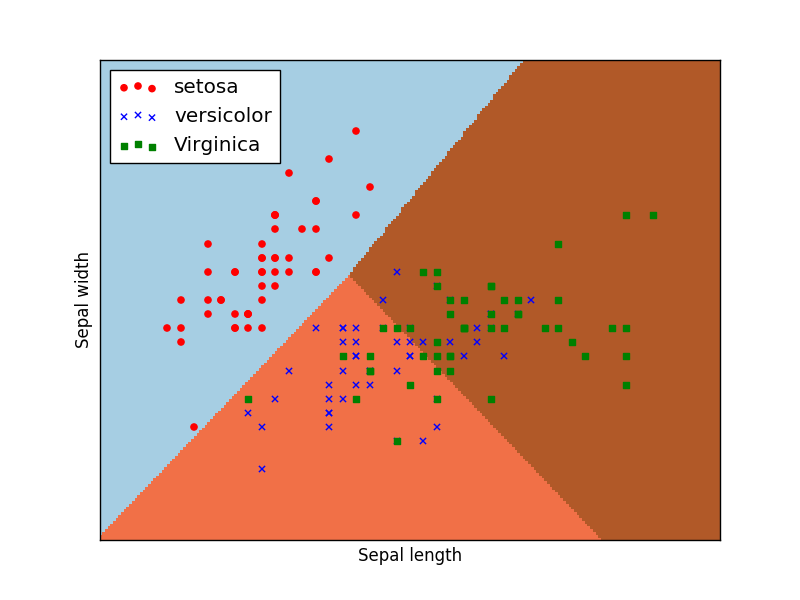


图 3.1 分类决策边界