机器学习在数据分析中的实践与应用

幸锋[1]，刘兴旭[2]

(1 中国移动通信集团云南有限公司，昆明 650228；2 中国移动通信集团设计院有限公司，北京 100080)

摘　要　机器学习技术能够使机器从大量的数据中学习规律，从而对新的样本做出分类识别，或者对未来做出合理的预测。本文应用鸢尾花数据集介绍了机器学习应用于数据分析的一般流程，分析与比较了典型的机器学习数据分析方法，比如主成分分析、线性判别分析和K-Means聚类等方法，阐述了机器学习在数据分析中的实践与应用。

关键词　机器学习；数据分析；鸢尾花数据集

中图分类号　TN915　　**文献标识码**　A　　**文章编号**　1008-5599（2021）12-0082-03

DOI:10.13992/j.cnki.tetas.2021.12.016

随着当今时代信息技术的快速发展，数据已经成为各行各业越来越重要的生产要素，也是传统行业进行数字化转型升级的抓手。如何充分高效的利用数据并发挥数据的经济价值，是当今时代的一个关键任务。人工智能技术的兴起给我们提供了丰富的数据分析方法，机器学习在数据分析领域起到了越来越重要的作用。

本文应用鸢尾花数据集对机器学习技术中的主成分分析、线性判别分析和 K-Means 聚类进行了论证比较，从而对鸢尾花数据集进行分析，利用鸢尾花的 4 个属性来预测鸢尾花属于哪一类亚属。鸢尾花数据集包含 150 个数据样本，分为山鸢尾、变色鸢尾和维吉尼亚鸢尾 3 类属性，每类 50 个数据，每个数据元素包含萼片长度、萼片宽度、花瓣长度和花瓣宽度 4 个属性。因此，我们将其视为 150×4 的矩阵并将这些信息作为标签机制的基础。关于这个数据集，我们使用了 Jupyter Notebook 来做数据处理与分析。鸢尾花数据集内置于 python 的 Sklearn 库中，使用 Jupyter Notebook，python 代码可以分段运行，这对进行数据分析任务来说非常灵活方便。

1　数据预处理

首先，对数据集进行描述性统计，结果见表 1。其次，绘制直方图和箱形图，如图 1 和图 2 所示。从这些数字可以看出，数据中的测量尺度是合适的，不需要再对数据进行进一步的预处理。

表1　描述性统计

数据元素	萼片长度	萼片宽度	花瓣长度	花瓣宽度
数量（个）	150	150	150	150
平均值（cm）	5.84	3.05	3.76	1.20
标准差（cm）	0.83	0.43	1.76	0.76
最小值（cm）	4.30	2.00	1.00	0.10
25% 值（cm）	5.10	2.80	1.60	0.30
50% 值（cm）	5.80	3.00	4.35	1.30
75% 值（cm）	6.40	3.30	5.10	1.80
最大值（cm）	7.90	4.40	6.90	2.50

收稿日期：2021-10-19

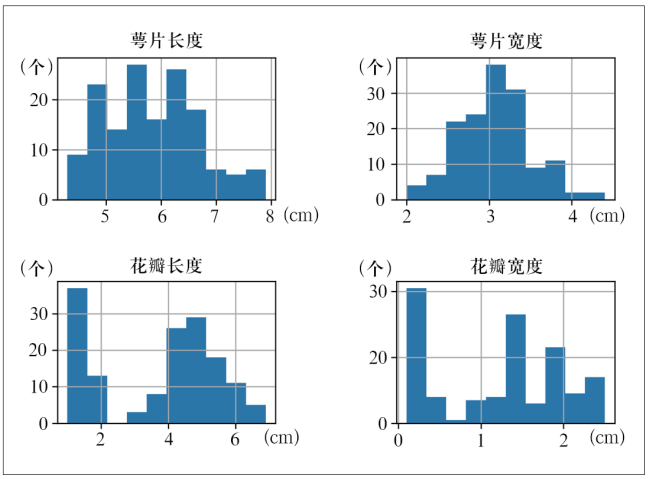

图1　直方图

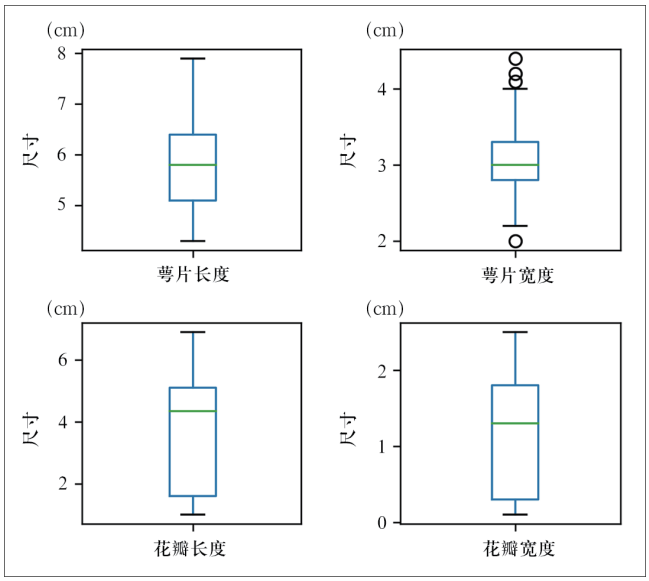

图2　箱型图

2　可视化

通过利用 pandas 工具，我们可以绘制多种图来进行数据分析，包括散点图、密度图和安德鲁斯曲线等。

选择一对现有的维度来绘制散点图，如图 3 所示。可以看出，在散点图上萼片的长度和宽度被划分为两个簇，它们分别有一定的关系。此外，绘制散点矩阵以获得更多的信息，如图 4 所示。可以看出，萼片的长度和宽度、花瓣的长度和宽度与鸢尾花的种类有一定的相关性。在这 3 种类别的分布中，山鸢尾在任何分布中都比其它两种更集中，萼片长度最长，花瓣宽度最短。对于同一属性的平均值来说，除萼片宽度外，其余 3 种的平均值按从大到小的顺序为维吉尼亚鸢尾、变色鸢尾、山鸢尾。

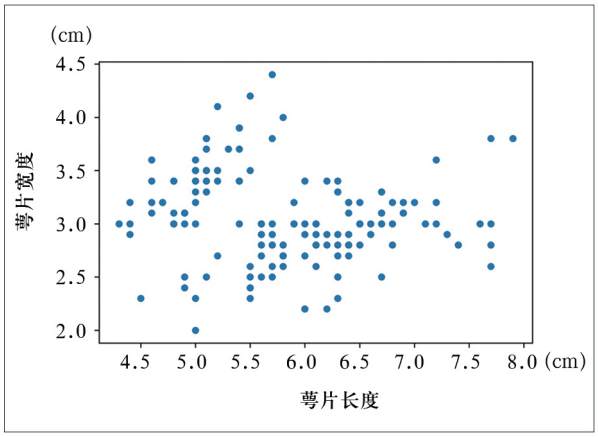

图3　在一对维度上的散点图

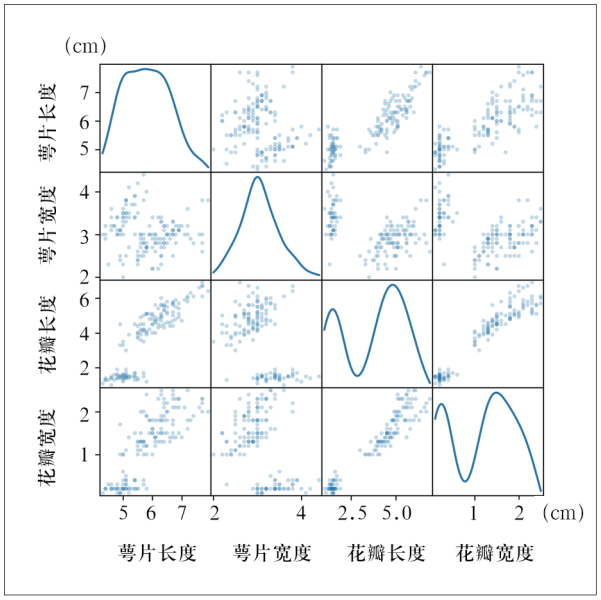

图4　散点矩阵

3　主成分分析（PCA）

PCA 是由因子分析演变而来的降维方法，通过正交变换将原始特征转化为线性无关的特征，得到的特征称为主成分。PCA 可以将原始维数降为 n 维。在特殊

情况下，通过 PCA 将维数降为 2 维。这样可以将多维数据转换为平面上的点，达到多维数据可视化的目的。如图 5 所示，对鸢尾花数据集应用 PCA，绿色代表山鸢尾，红色代表变色鸢尾，蓝色代表维吉尼亚鸢尾。矩阵是 150×2 维的主成分分析，提取的两个主成分共携带 97.77% 的信息，这说明主成分具有较好的解释效果。

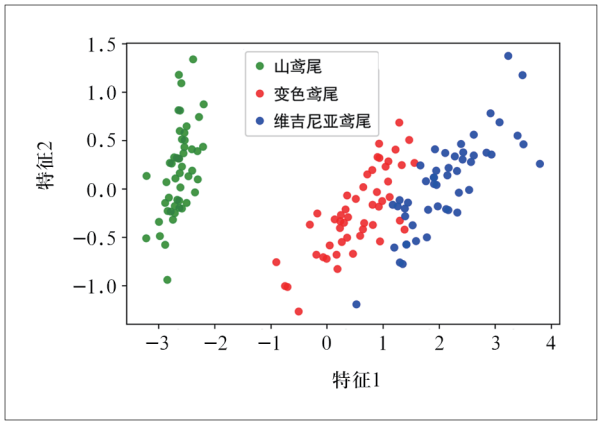

图5　对鸢尾花数据集应用PCA

4　线性判别分析（LDA）

LDA 的基本思想是通过投影降低标注数据的维数，使不同类型点的距离远，相似点的离散程度小。比较 PCA 和 LDA，PCA 是无监督降维，LDA 是有监督降维；PCA 期望预测数据的方差尽可能大（最大可分性），假设的方差越大，它包含的信息就越多。LDA 则期望同一类的组内方差小，组间方差大。LDA 可以合理利用标签信息，使投影维数具有判别性，尽量分离不同类型的数据。对鸢尾花数据集应用 LDA 如图 6 所示。

5　K–Means 聚类

K–Means 聚类算法是一种迭代求解的聚类分析算法。将数据分成 k 组，从 k 个对象中随机抽取数据作为初始聚类中心，然后计算每个对象到每个聚类中心的距离，并从其最近的聚类中心开始分配给每个对象。对鸢尾花数据应用 K–Means 聚类如图 7 所示。

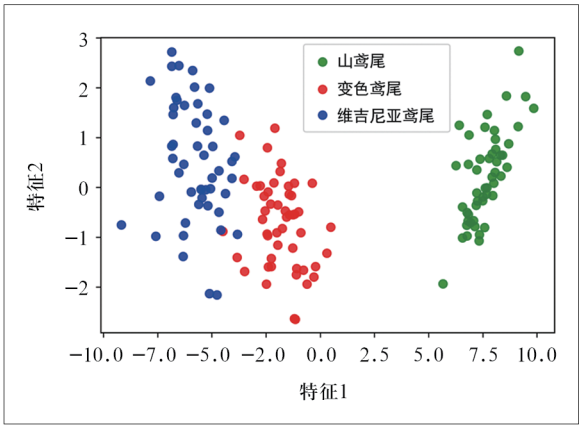

图6　对鸢尾花数据集应用LDA

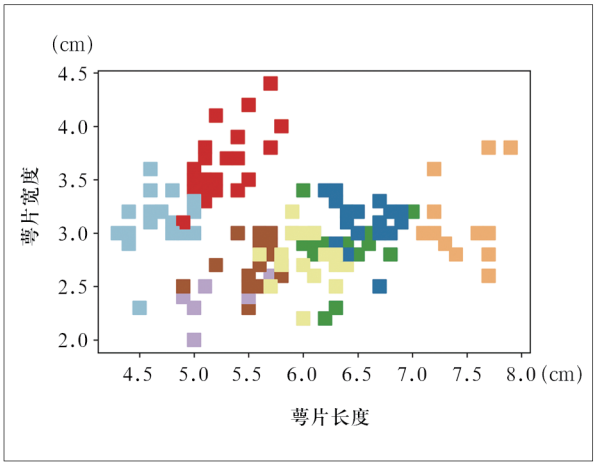

图7　对鸢尾花数据应用K–Means聚类

6　结束语

综上所述，本文从具体的案例阐述了机器学习在数据分析中的应用，并对典型的算法如主成分分析、线性判别分析和 K–Means 聚类进行了实际的应用。由该案例可见，机器学习在数据分析领域有着广阔的发展前景，能够为不同公司和机构带来各个方面的效益。

参考文献

[1] 王芳．主成分分析与因子分析的异同比较及应用[J]．统计教育，2003(5)．

[2] 董虎胜．主成分分析与线性判别分析两种数据降维算法的对比研究[J]．现代计算机(专业版)，2016(7)．

（下转第 88 页）

3 结束语

基于上述的分析及实验论证，NSA 单锚点网络连续覆盖时，无需配置双锚点模式，不会影响到锚点覆盖率及 5G 网络覆盖率，同时锚点间只涉及同频切换，切换简单且避免双层锚点网间的频繁切换，5G 下载速率会更高。NSA 单锚点网络非连续覆盖时，需要配置双锚点模式进行单锚点覆盖空洞的补充，以提升锚点覆盖率及 5G 网络覆盖率，进而大幅提升 5G 下载速率。

单双锚点配置组网策略均存在各自的适用场景及适用范围条件，会直接影响到 5G 网络覆盖率及下行速率，配置时需要结合场景特性及锚点网络覆盖情况，全面评估后选择合适配置，从而保证 5G 网络覆盖及感知的最优化。

参考文献

[1] 蒋声铭，裴蕾，罗通武．5G网络NSA组网及锚点选择分析[J]．通信电源技术，2020(6)．

Research and application of single and double anchor networking strategy in 5G non-standalone network

YU Fei[1], ZHAO Chun-yang[1], LIU Yue[2]

(1 China Mobile Group Design Institute Co., Ltd. Shanghai Branch, Shanghai 200060, China; 2 China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract 5G in non-standalone network architecture mode, the configuration of 4G anchor site is particularly important, which directly affects the coverage and performance of 5G network. This paper compares and analyzes the characteristics of the two networking strategies from the perspective of single anchor configuration and double anchor configuration, and summarizes the application scenarios of different anchor configuration based on the actual application of existing network, so as to improve the coverage performance and user perception of 5G network experience.

Keywords 5G; non-standalone; single anchor; double anchor

（上接第 84 页）

Practice of machine learning in data analysis

XING Feng[1], LIU Xing-xu[2]

(1 China Mobile Group Yunnan Co., Ltd., Kunming 650228, China; 2 China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract Machine learning enables machines to learn rules from large amounts of data, so as to classify new samples or make reasonable predictions about the future. This paper introduces the general process of machine learning applied to data analysis by using iris plants database, analyzes and compares typical machine learning methods in data analysis, such as principal component analysis, linear discriminant analysis and K-means clustering, and expounds the practice of machine learning in data analysis.

Keywords machine learning; data analysis; iris plants database