肥料登记数据分析

摘要

本文主要针对肥料登记数据对各省肥料数据，利用 python 对肥料产品数据进行数据化分析与可视化操作。

针对任务一：首先，我们对所有的数据集进行缺失值、异常值、重复值等进行查验和处理。

针对任务二： 首先，我们确定组距，通过筛选复混肥料的产品，按照总无机养分百分比最大值和最小值之差按 10 等份的对数据进行分组，得出各分组的标签；其次针对氮、磷、钾养分的百分比利用 K-means 聚类分析方法进行聚类分析，并将结果可视化操作。

针对任务三：利用附件三以原料作为特征计算出杰卡德相似系数矩阵完成结果。

针对任务四：

关键字：列表推导式、K-means 聚类分析、杰卡德相似

目录

一、 背景介绍

肥料是农业生产中的一种非常重要的一种生产资料，生产销售必须遵守《肥料登记管理办法》这项法律法规，并需要按照法律在农业行政管理部门进行登记。各省、市、自治区、直辖市人民政府农业行政主管部门主要负责本行政区域内销售的肥料进行登记工作，相关数据可以从政府网站上自由下载。

二、 数据挖掘与分析工具

2.1 数据挖掘工具

2.1.1 TipDM 数据库建模平台和挖掘工具

TipDM 数据挖掘建模平台是基于 Python 引擎、用于数据挖掘建模的开源平台。TipDM 提供数量丰富的数据分析与挖掘建模组件，用户可在没有编程基础的情况下，通过拖拽的方式进行操作，将数据输入输出、数据预处理、挖掘建模、模型评估等环节通过流程化的方式进行连接，帮助用户快速建立数据挖掘工程，提升数据处理的效能。

主要特性：

- 基于 Python，用于数据挖掘建模。
- 使用直观的拖放式图形界面构建数据挖掘工作流程，无需编程。
- 支持多种数据源，包括 CSV 文件和关系型数据库。
- 支持挖掘流程每个节点的结果在线预览。
- 提供 5 大类共 40 种算法组件，包括数据预处理、分类、聚类等数据挖掘算法。
- 支持新增/编辑算法组件，自定义程度高。
- 提供众多公开可用的数据挖掘示例工程，一键创建，快速运行。
- 提供完善的交流社区，提供数据挖掘相关的学习资源（数据、代码和模型等）。

2.2 数据分析工具

2.2.1 MATLAB

MATLAB 是美国 MathWorks 公司出品的商业数学软件，用于数据分析、无线通信、深度学习、图像处理与计算机视觉、信号处理、量化金融与风险管理、机器人，控制系统等领域。

优势特点：

◆ 高效的数值计算及符号计算功能，能使用户从繁杂的数学运算分析中解脱出来；

◆ 具有完备的图形处理功能，实现计算结果和编程的可视化；

◆ 友好的用户界面及接近数学表达式的自然化语言，使学者易于学习和掌握；

◆ 功能丰富的应用工具箱（如信号处理工具箱、通信工具箱等），为用户提供了大量方便实用的处理工具，如图 1 所示。

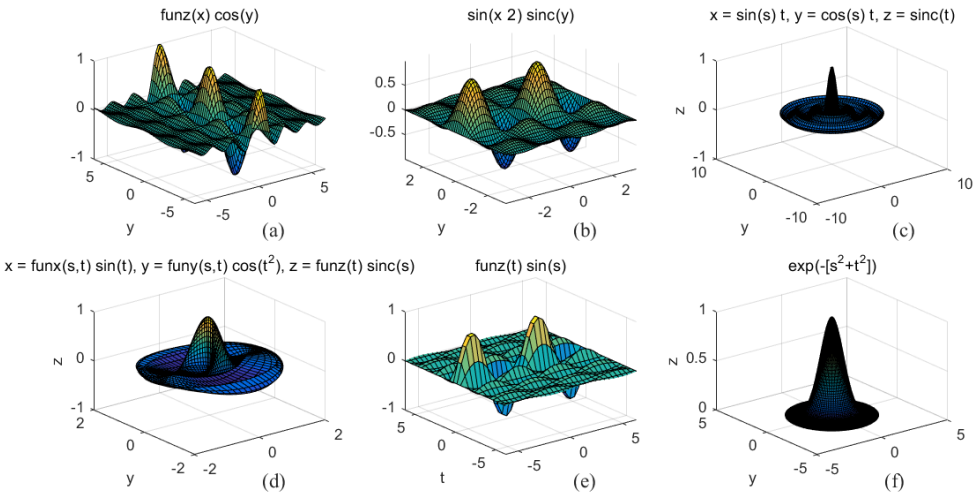

图 1 MATLA 绘图

2.2.2 Pandas 库

Pandas 是基于 Numpy 的一个 Python 的第三方数据分析库，它的作用是用于数据分析，纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据

集所需的工具。同时也提供了大量能使我们快速便捷地处理数据的函数和方法，这些函数和方法是使它成为强大而高效的数据分析环境的重要因素之一。

三、 任务一 数据预处理

3.1 数据的探查与名称的规范

3.2 统计总无机养分的百分比

四、 任务二 肥料产品的数据分析

4.1 分析复混肥料产品的分布特点

随后基于结果数据，我们利用 matplotlib 函数对数据进行可视化操作，如图 10 所示。

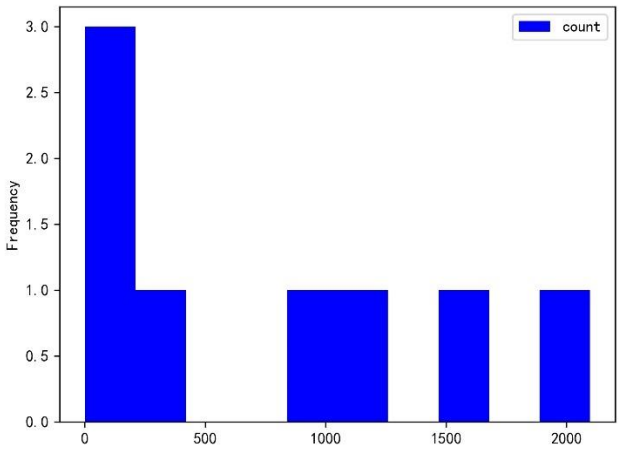

图 10 总无机养分分组直方图

排名	一	二	三
分组标签	7	6	5
产品登记数量	2012	1501	1038

4.2 分析有机肥料的产品产品的分布特点

基于图表，我们为了使数据更直观的表现，我们利用网页中图标秀对数据实现可视化操作，如图 13 所示。

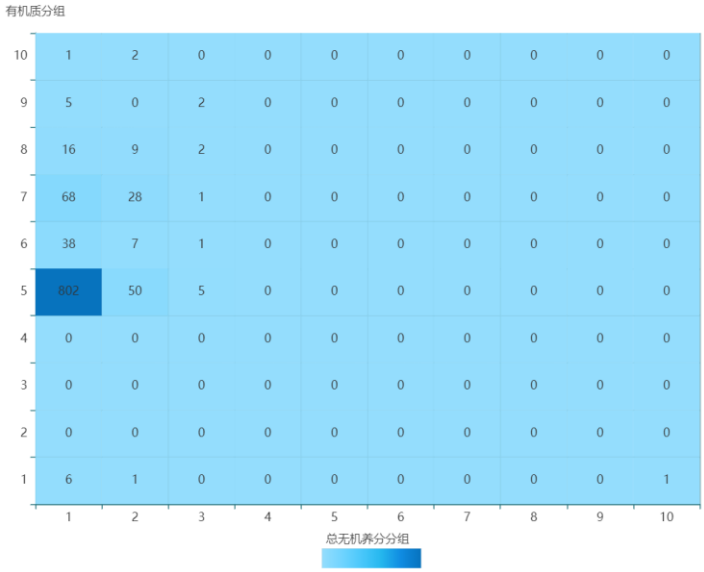

图 13 分布热力图

4.3 基于 K-Means 算法分析

针对任务 2.3，我们对附件中的数据进行初步的判断，发现使用 K-means 算法适合该任务。K-means 聚类分析算法是一种无监督的算法，也是数据挖掘算法的经典算法之一，其根据“物以类聚”的思想，对样本数据进行一个多元统计，基于样本间的相似度进行归类，使簇内的样本尽可能地大，不同的簇差异尽可能的大，从而达到分类的效果。

使用聚类方法，首先要确定一个数据的划分，然后采用迭代的算法，使得类中心不断进行修正于重定位，直到类中心在迭代地不在变化，则聚类完成。一个好的划分标准是：类内的样本相似度基本最大，类间样本相似度最小。

我们基于这四个聚类中心，进行归类，并在原始表中增加聚类结果标签，按照不同的类别提取出不同的数据，部分结果如图 15 所示。

	序号	企业名称	产品通用名称	产品形态	...	产品商品名称	适用作物	总无机养分百分比	聚类标签
4	5	湖北兴特尔化工有限公司	复混肥料	粒状	...	NaN	NaN	0.45	1
6	7	湖北兴特尔化工有限公司	复混肥料	粒状	...	NaN	NaN	0.45	1
10	11	嘉施利（应城）化肥有限公司	复混肥料	粒状	...	NaN	NaN	0.45	1
18	19	应城市新都化工有限责任公司	复混肥料	粒状	...	NaN	NaN	0.40	1
20	21	湖北兴特尔化工有限公司	复混肥料	粒状	...	NaN	NaN	0.40	1

图 15 聚类分析一部分结果

基于以上结果，我们利用 python 中的 matplotlib 函数绘制三维表，如图 16、17 所示。

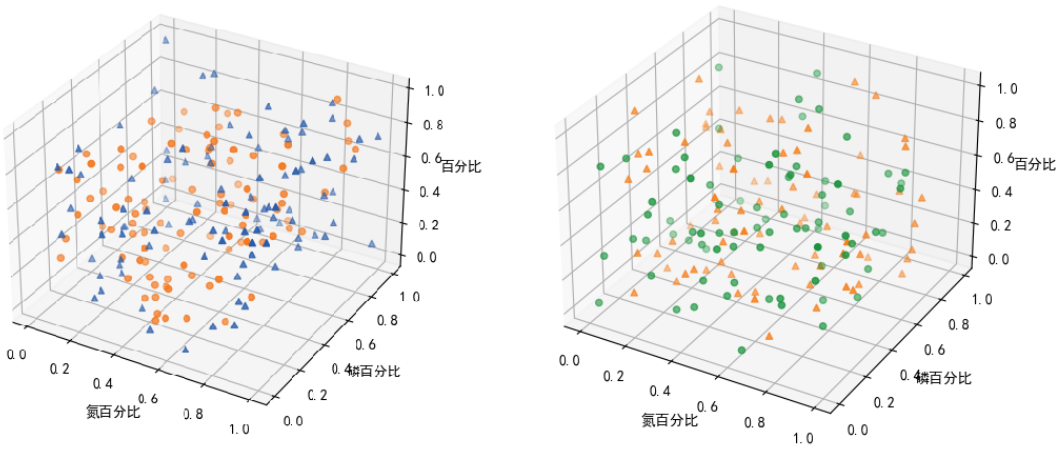

图 16 聚类分析三维散点图

五、 任务三 肥料产品的多维度对比分析

5.1 分析复婚化肥的变化趋势

基于以上数据，我们利用图标秀将数据进行可视化操作，如图 19 所示。

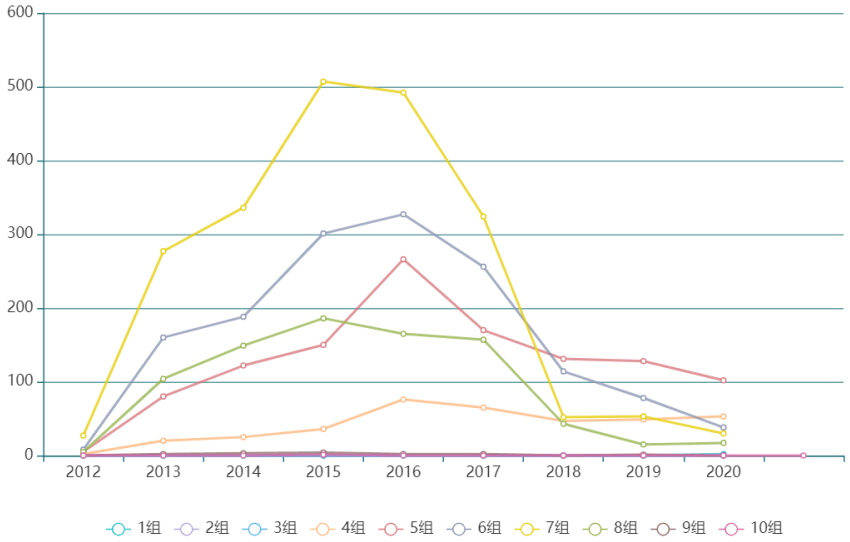

图 19 数据标签随年份的变化

5.3 分析广西和湖北产品分布差异

5.4 产品登记数量大于 10 的肥料企业

六、 任务四 肥料产品的多维度对比分析

6.1 基于算法提取出氮、磷、钾养分和有机质的百分比

6.2 各原料所占比

参考文献

[1]翁佳烽, 梁晓媛, 谭浩波,等. 基于 K-means 聚类分析法的肇庆市干季 PM_(2.5)污染天气分型研究[J]. 环境科学学报, 2020, 40(2):15.
[2]熊励, 王锟, 钟美芝. 大数据可视化分析在支撑智库研究中的应用与创新[J]. 2021(2018-4):15-24.
[3]刘婷婷, 龚敏琪, 陈泳琳,等. 可持续设计方法的多维分析及其可视化[J]. 包装工程, 2020, 41(4):9.
[4]王菲, 袁婷, 谷守宽,等. 有机无机缓释复合肥对不同土壤微生物群落结构的影响[J]. 环境科学, 2015, 36(4):7.
[5]张若愚. Python 科学计算[M]. 清华大学出版社, 2016.
[6]陈赫, 吕丽君. Matlab 在数字信号处理中的应用[J]. 长治学院学报, 2018, v.35;No.183(02):48-50.
[7]张恩平, 谭福雷, 王月,等. 氮磷钾与有机肥配施对番茄产量品质及土壤酶活性的影响[J]. 园艺学报, 2015, 42(10):2059-2058.