肥料登记数据分析

摘要

“民以食为天”，粮食是人类最基本的生存资料，而肥料作为农业生产中一种重要的生产资料，在农业生产中扮演着极其重要的角色，国家对其高度重视，规定其生产销售必须遵循《肥料登记管理办法》，依法在农业行政管理部门进行登记。

本文将围绕“肥料登记数据分析”这一论题，根据比赛方提供的数据展开描述，达成以下目标：1.对肥料登记数据进行预处理；2.根据养分的百分比对肥料产品进行细分；3.从省份、日期、生产商、肥料构成等维度对肥料登记数据进行对比分析；4.对非结构化数据进行结构化处理。

针对任务一，根据任务要求我们首先对数据进行了预处理，将附件 1 数据导入 pandas 库，按照复混肥料（掺混肥料归入这一类）、有机-无机复混肥料、有机肥料和床土调酸剂这 4 种类别对其进行规范化处理；并根据各肥料产品的氮、磷、钾养分百分比，计算出了总无机养分百分比。

针对任务二，根据任务要求我们对附件 2 中的产品进行了筛选。对有复混肥料和有机肥料的产品，对其进行了分组分类处理，并分别为其打上了分组标签，绘制出了复混肥料产品登记数量的直方图和有机肥料产品的分布热力图，分别列出两种产品登记数量最大的前 3 个分组及相应的产品登记数量，绘制出了复混肥料的产品三维散点图、散点图矩阵、雷达图，并根据聚类结果的雷达图分析每个聚类的特征。

针对任务三，在任务二的基础上，我们对肥料产品进行了多维度对比分析。分析比较了复混肥料中各组别不同年份产品登记数量的变化趋势，分别筛选出广西和湖北 2021 年 9 月 30 日仍有效的有机肥料产品登记数量在前 5 的组别，并具体阐述了两个省份上述组别的分布差异。从附件 3 中提取产品登记数量大于 10 的肥料企业，给出这些企业所用到的原料集合，并计算出了企业之间的杰卡德相似系数矩阵。

针对任务四，根据任务要求，我们对附件 4 中的肥料产品进行了多维度对比分析，提取出了氮、磷、钾养分和有机质的百分比、肥料含氯的程度以及各种原料的名称及其百分比。

关键字：肥料登记；数据预处理；产品登记数量；多维度对比分析

目录

一、背景介绍

“民以食为天”，粮食是人类最基本的生存资料，农业在国民经济中的基础地位，突出地表现在粮食的生产上。全国 14 亿多人口的粮食、肉类、蔬菜、水果等食物和纺织用纤维等，共 20 多万吨，除少数年份为调剂品种而有进口外，都来自本国农业。农业属于第一产业，在国民经济中是一个重要的产业部门，以土地资源为生产对象，支撑着国民经济的建设和发展。农业是支撑国民经济建设与发展的基础产品，农业是人们利用动植物体的生活机能，把自然界的物质和能转化为人类需要的产品的生产部门。

而肥料是农业生产中一种重要的生产资料，肥料的使用大幅度促进了农业生产产量，我们在网上查找发现截止 2021 年 4 月 26 日，2021 年全国肥料行业企业名录共有 13999 家，这个行业的发展关系着我们的民生发展，我国也对其进行了严格要求，其生产销售必须遵循《肥料登记管理办法》，依法在农业行政管理部门进行登记。

国家鼓励生产优质、高效、安全的肥料产品，支持肥料研究、科学施用以及地力培肥，相关企业也需要担负起社会责任，遵守相关规定要求，合规合法生产经营。各省、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登记工作，根据《中华人民共和国肥料管理条例》要求，自治区、直辖市人民政府农业行政主管部门负责本辖区生产的下列肥料产品的登记:(一)氮肥、磷肥、钾肥、复合肥料。(二)复混肥料、有机肥料。(三)床土调酸剂。国务院农业行政主管部门负责下列肥料产品登记:(一)含微量元素肥料、含中量元素肥料、微生物肥料。(二)土壤调理剂。(三)进口肥料。为便于相关管理部门更好的开展工作，我们有必要将肥料生产信息进行统计分类。

二、技术介绍

2.1 大数据

大数据（big data），或称巨量资料，指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。大数据

的 5V 特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。

2.2.1 大数据预处理

完成对已接收数据的辨析、抽取、清洗等操作。

（1）抽取：因获取的数据可能具有多种结构和类型，数据抽取过程可以帮助我们将这些复杂的数据转化为单一的或者便于处理的构型，以达到快速分析处理的目的。

（2）清洗：对于大数据，并不全是有价值的，有些数据并不是我们所关心的内容，而另一些数据则是完全错误的干扰项，因此要对数据通过过滤“去噪”从而提取出有效数据。

2.2.2 大数据分析

大数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，大数据分析可帮助人们做出判断，以便采取适当行动。

在本次论文中我们主要使用的是 Python 来实现具体的大数据分析流程，Python 可以利用各种 Python 库，如 NumPy、Pandas 以及 Matplotlib，高效的解决各种各样的大数据分析问题。

①Numpy 库

Numpy 是系统自带的一种开源的数值处理计算扩展，是一个用 Python 实现的科学计量包，它提供许多高级的数值编程工具，内置了并行运算功能，当系统有多个核心时，做某种计算时会自动做并行计算。

②Pandas 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。

③Matplotlib 库

Matplotlib 是一个 Python 的图形框架，是 Python 著名的绘图库，提供了一整套和 Matlab 相似的命令接口,十分适合交互式地进行制图，还可以方便的作为绘图控件，嵌入 GUI 应用程序中。

2.3 可视化分析

数据可视化起源于 1960 年计算机图形学，是利用图表呈现数据内容的一种方法。数据可视化的概念中，有一个关键信息——数据可视化研究的对象是数据可视化的表现形式。

数据可视化分析，是通过构建数据可视化图表，展示数据特征，从中发现数据信息的过程。它包括两个步骤：数据可视化呈现和数据分析洞察。数据可视化分析是利用数据可视化呈现能力，进行数据分析的一种方法，通过可视化呈现的图表，发现有用的信息，得出数据结论和辅助宏观决策。简单来说，就是把枯燥的数字变成各种各样的图表，更好地帮助你发现其中有价值的信息。

数据可视化分析是实现广义数据分析的一种模式，具有与狭义数据分析相同的体系结构，并且在某些方面，拓展了数据可视化的内容。数据可视化分析包括数据可视化呈现（制作可视化图表）和数据分析洞察（基于图表识别信息）两个过程。在实际的工作和业务场景中，通常用于发现业务运营过程中出现的问题，以及进行辅助决策。

2.4 K-Means 聚类算法

2.3.1 聚类概念

聚类分析又称群分析，它是研究（样品或指标）分类问题的一种统计分析方法，同时也是数据挖掘的一个重要算法。聚类（Cluster）分析是由若干模式（Pattern）组成的，通常，模式是一个度量（Measurement）的向量，或者是多 维空间中的一个点。聚类分析以相似性为基础，在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性。

三、任务实施

3.1 数据的预处理

3.2 肥料产品的数据分析

3.2.1 分析复混肥料产品的分布特点

直方图又被成为质量分布图可以直观地显示质量特性的分布状态，对于数据的分布的形状、中心位置和分散程度一目了然；关注数据和规格的关系，通过测定值与规格值比较，判断出不良是平均不良还是异常的不良，便于人们确定在何处进行质量改进；（在此对平均不良和异常不良作个解释：平均不良通常代表的是系统的问题，是整个过程的不良；异常不良却代表了个别的离散的不良，属于个别问题）。

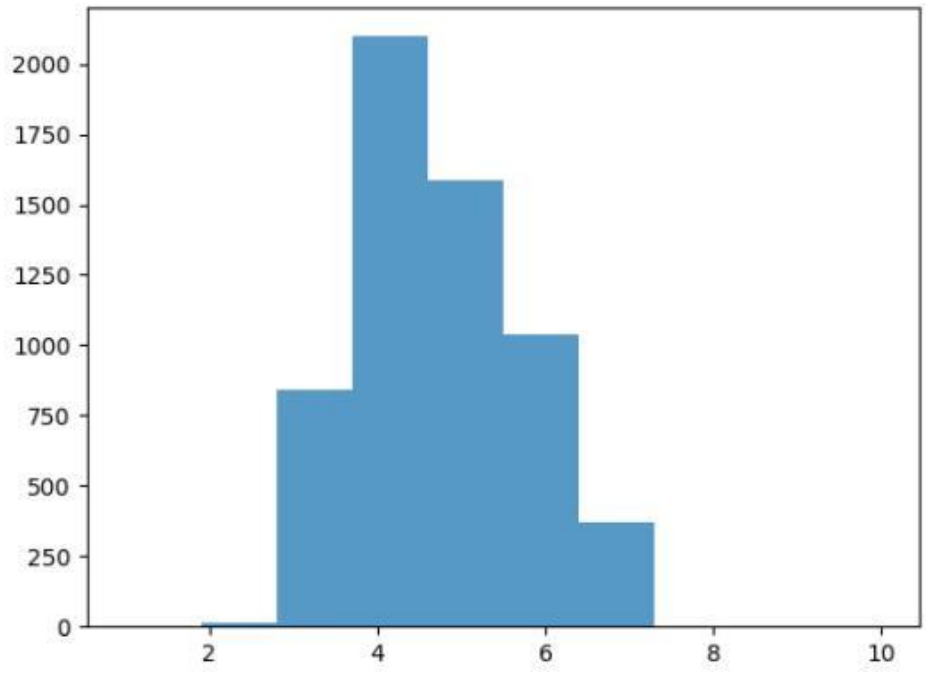

图 6 产品登记数量直方图

从图 6 中我们可以看到这是一个偏态形的直方图，第一组、第二组、第三组在直方图中的分布很少，说明无机养分百分比总和比较大的复混肥料产品登记数量也很少，同理，第 8 组、第 9 组、第 10 组中无机养分百分比总和大的复混肥料产品登记数量比较少。因此复混肥料产品主要分布在第 4 组、第 5 组和第 6

组。

利用已经整理得到的数据，我们将复混肥料产品登记数量排在前三的小组筛选出来，结果如表 2 所示。

表 1 复混肥料产品分布前三排行表

排名	一	二	三
分组标签	4	5	6
产品登记数量	2098	1584	1040

3.2.2 分析有机肥料产品

根据任务 2.1 相同的做法，首先我们需要从附件 2 中将有机肥料的筛选出来，其次将总无机养分百分比按照从大到小的顺序分为 10 组，之后利用有机质百分比最大值减最小值再除以 10 的方法将有机质百分比从大到小分为 10 组

3.2.3 用聚类算法将产品分类

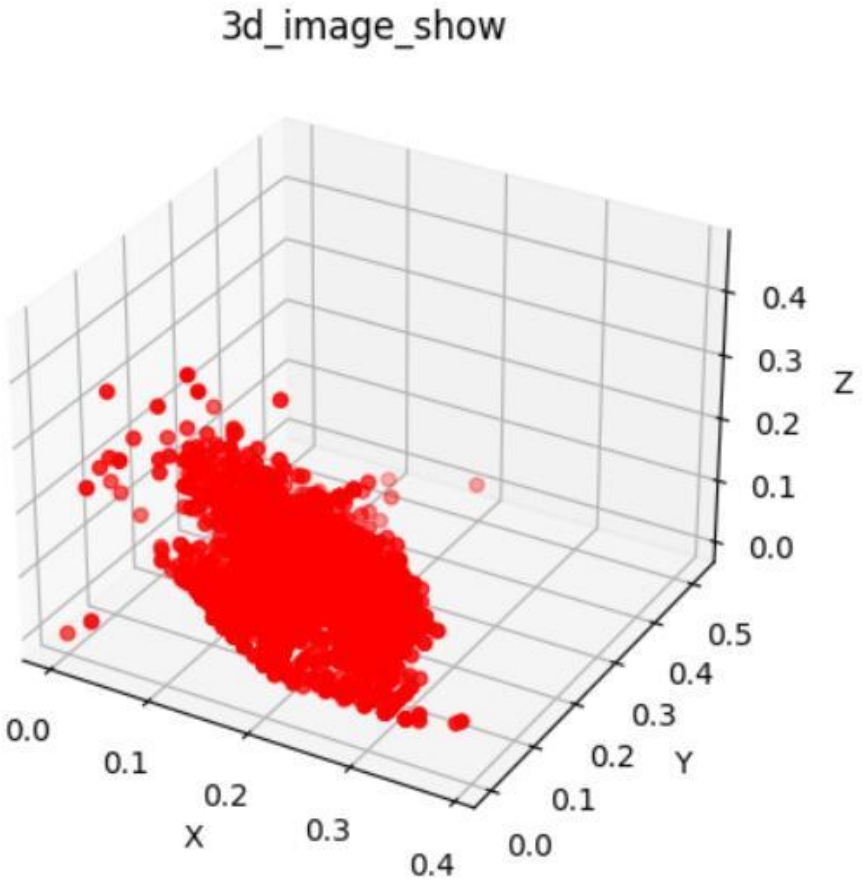

图 11 三维散点图

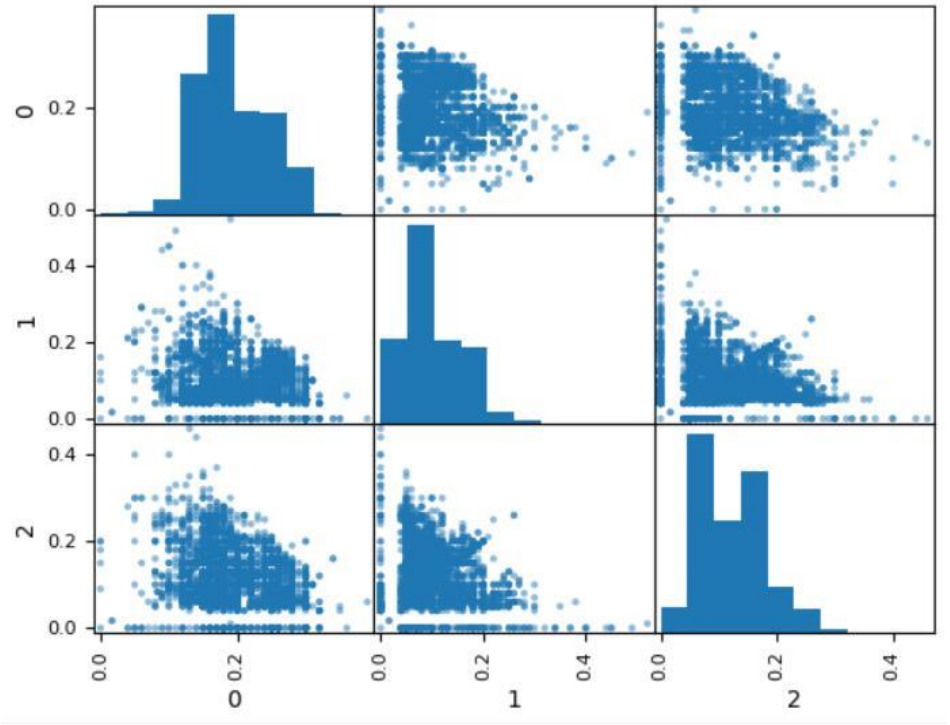

图 12 散点图矩阵

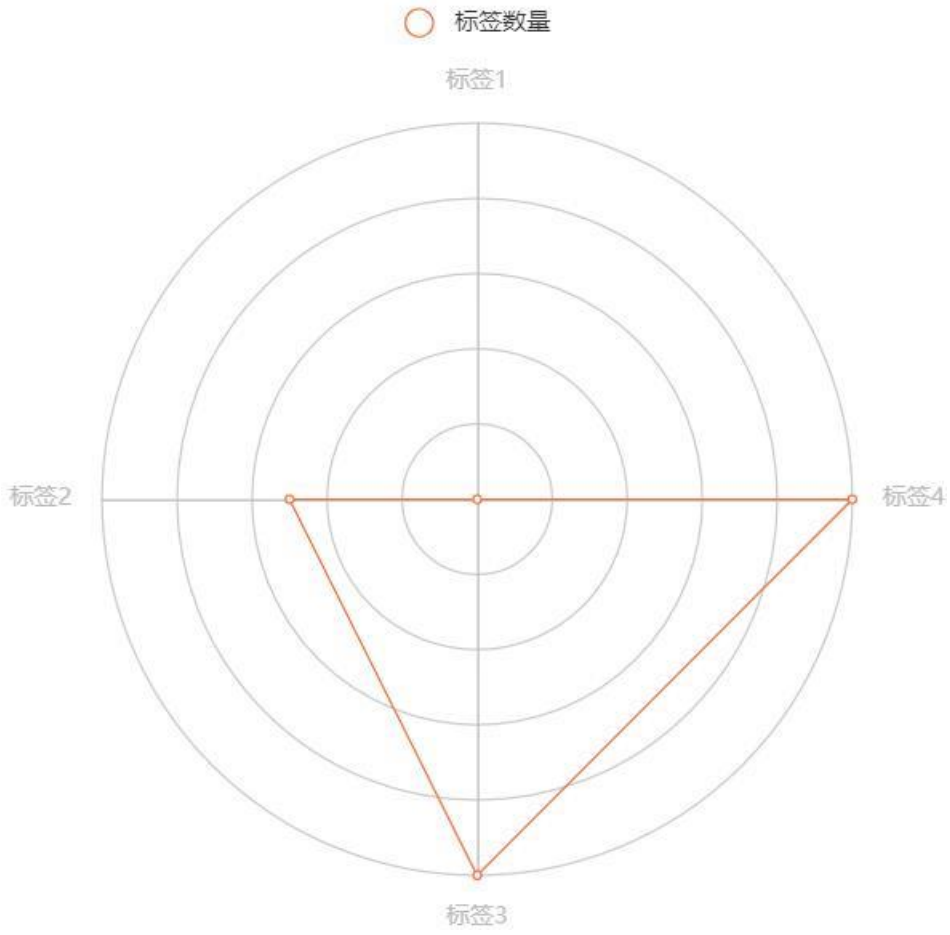

图 13 聚类结果雷达图

为了更好的展示我们的分析效果，根据任务要求，我们采用了雷达图来分析每个聚类的特征，从图中我们可以看到共四个标签，有一个点直接位于中心点，剩余三个点有两个直接位于标签 3 和 4 上面，另外一个点位于标签 2 和中心点之间，越靠近中心点的说明其簇类基本没有变化。

3.3 肥料产品的多维度对比分析

根据任务需求，从附件 3 中提取产品登记数量大于 10 的肥料企业，并以各企业用到的原料作为特征，计算企业之间的杰卡德相似系数矩阵。集合 A 与 B 的杰卡德相似系数定义的公式如（3.1）所示。

参考文献

[1] 李翠. Web 前端地理数据可视化技术研究与实践[D]. 2016.

[2] 郝爽, 李国良, 冯建华,等. 结构化数据清洗技术综述[J]. 清华大学学报(自然科 学版), 2018.

[3] 陈明. 大数据可视化分析[J]. 计算机教育, 2015(05):94-97.

[4] 龙虎 "大数据背景下的数据分析与可视化研究." 034.003(2016):98-102.

[5]杨东华, 李宁宁, 王宏志,等. 基于任务合并的并行大数据清洗过程优化[J]. 计算机学报, 2016, 000(001):97-108.