基于 K-均值聚类算法对肥料登记数据挖掘分析

摘要

肥料是重要的农业生产资料，农业的持续稳定发展离不开肥料，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。肥料因其速效养分含量丰富，增产效果显著，在生产中被广泛应用。但目前也存在化肥过量施用、盲目施用等问题。为了能更好的解决这些问题，本文基于聚类算法对肥料登记数据进行数据挖掘分析。

关键词：K-均值聚类算法、层次聚类算法、正则表达式

目 录

一、 问题重述

1.1 问题背景

化肥是重要的农业生产资料，是粮食的“粮食”。化肥在促进粮食和农业生产发展中起了不可替代的作用，但目前也存在化肥过量施用、盲目施用等问题，带来了成本的增加和环境的污染，亟须改进施肥方式，提高肥料利用率，减少不合理投入，保障粮食等主要农产品有效供给，促进农业可持续发展。

1.2 问题简述

任务一：数据预处理

(1)按照四种肥料类别，处理“产品通用名称”不规范名称；

(2)计算氮、磷、钾养分百分比之和。

任务二：肥料产品的数据分析

(1)对复混肥料产品进行等距分组，并打标签。绘制产品登记数量直方图；

(2)按照总无机养分百分比和有机质百分比分别对有机肥料进行等距分组，并打标签。同时，按照规定要求，绘制有机肥料产品的分布热力图；

(3)对复混肥料中氮、磷、钾养分的百分比进行聚类分析，根据聚类标签绘制肥料产品的三维散点图和散点图矩阵，分析每个聚类的特征。

任务三：肥料产品的多维度对比分析

(1)提取发证日期中的年份，分析比较复混肥料中各组别不同年份产品登记数量的变化趋势；

(2)提取 2021 年 9 月 30 日仍有效的有机肥料产品，选出广西和湖北产品登记数量在前 5 的组别，分析其分布差异；

(3)提取产品登记数量大于 10 的肥料企业，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。

任务四：肥料产品的多维度对比分析

(1)设计算法或处理流程，从技术指标中提取出氮、磷、钾养分和有机质的百分比，以及肥料含氯的程度；

(2)设计算法或处理流程，从原料与百分比中提取各种原料的名称及其百分比。

1.3 任务目标

(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。

二、数据说明

3.1 数据基本情况

(1)数据来源:政府网站

(2)统计日期范围:2012 年 1 月-2021 年 3 月

(3)数据提供方:

3.2 数据量

(1)附件 1：2926 条

(2)附件 2：7620 条

(3)附件 3：551 条

(4)附件 4：201 条

一、 数据挖掘与分析

3.2 任务二 肥料产品的数据分析

3.2.1 复混肥料总无机养分百分比分析(任务 3.1)

●绘制产品登记数量的直方图。具体结果如图 3.2.4 所示：

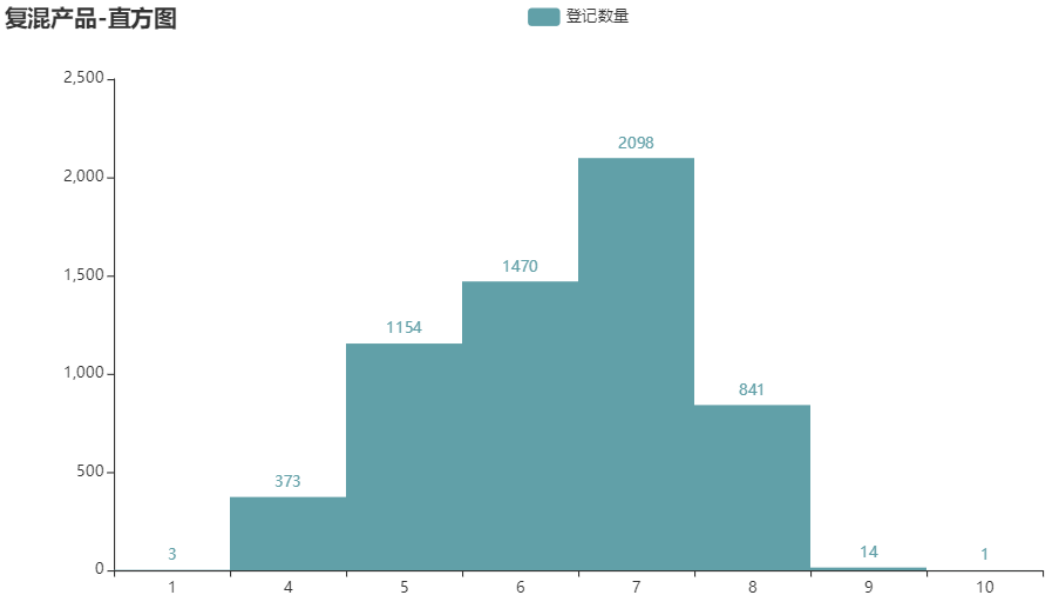

图 3.2.4 等宽切分直方图

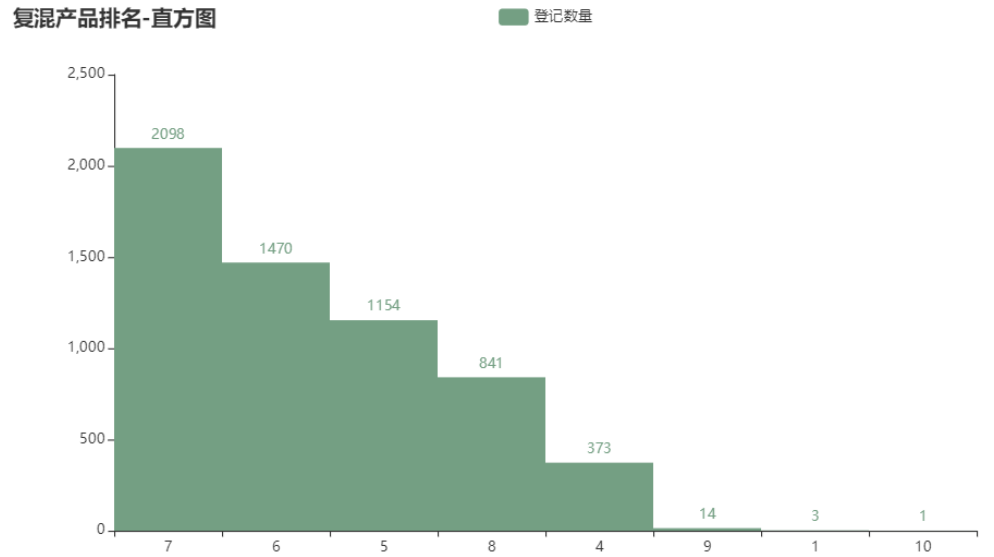

图 3.2.5 登记数量排名直方图

根据图 3.2.5 所示，可知登记数量排名前三分别是：7(2098)、6(1470)、5(1154)

排名	一	二	三
分组标签	7	6	5
产品登记数量	2098	1470	1154

3.2.2 有机肥料产品研究(任务 2.2)

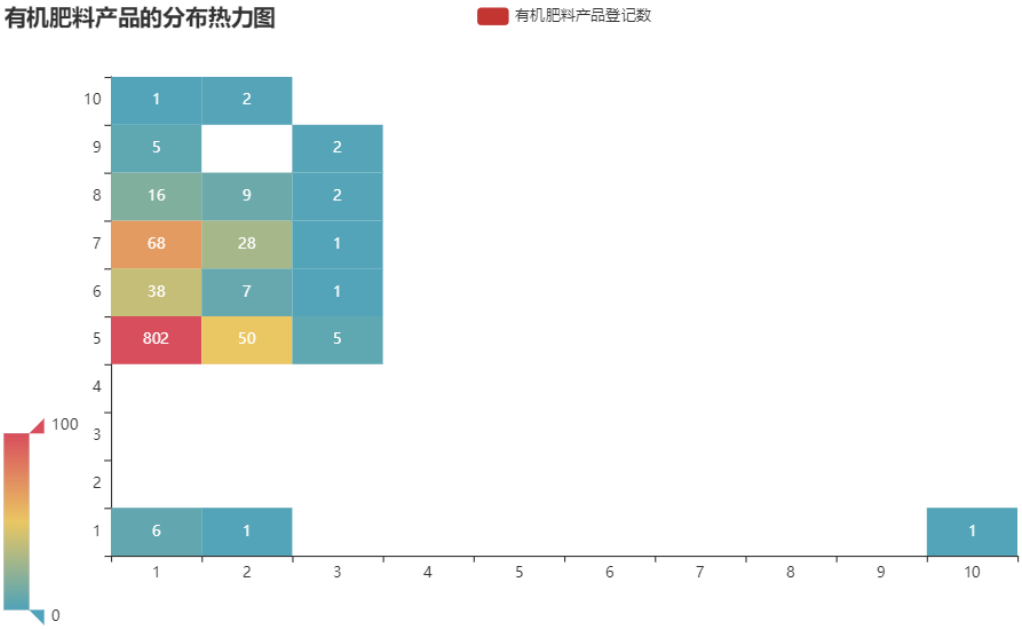

图 3.2.8 有机肥料产品的分布热力图

●**有机肥料产品分布特点分析：**按登记数量从大到小列出登记数量最大的前 3 个分组及相应的产品登记数量：

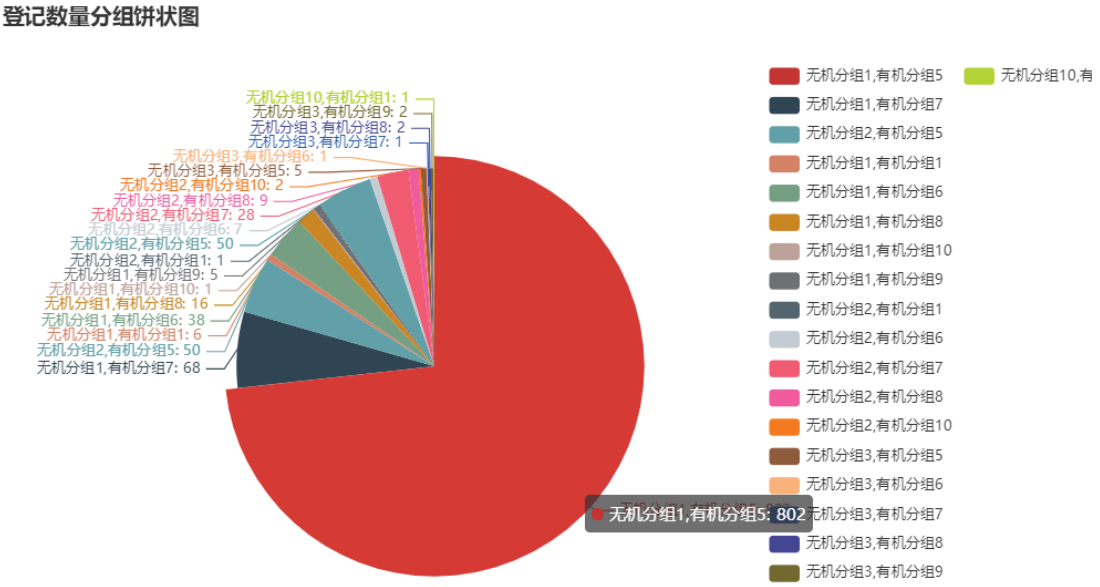

图 3.2.9 登记数量占比分布图

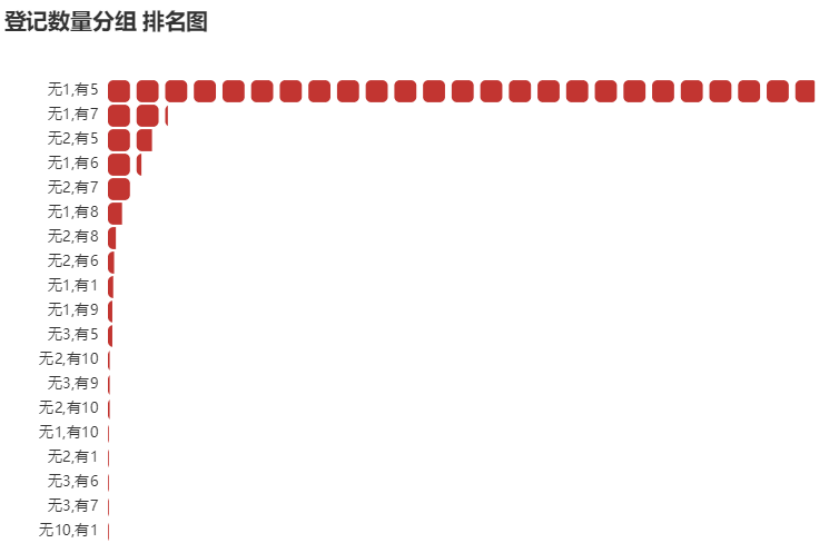

图 3.2.10 登记数量排名

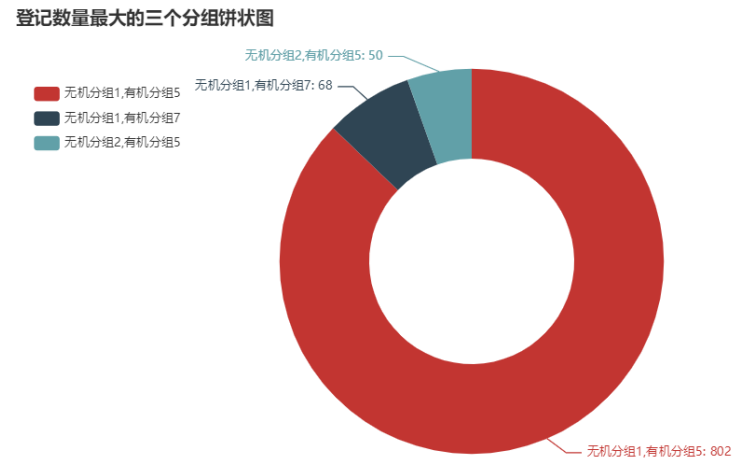

图 3.2.11 登记数量排名前三占比图

参考文献

[1] 中华人民共和国农业部，农业部关于印发《到 2020 年化肥使用量零增长行动方案》和《到 2020 年农药，使用量零增长行动方案》的通知，http://www.moa.gov.cn/nybgb/2015/san/-201711/t20171129_5923401.htm，2021.11.14

[2] Hugh Kim，数据分析(二)：数据清洗步骤，https://zhuanlan.zhihu.com/p/109413107，2021.8.6

[3] data learning，数据可视化之热力图&相关系数图，https://blog.csdn.net/weixin_45481473/article/details/112549366，2021.8.7

[4] Little_Rookie,K-means 算法原理,https://www.cnblogs.com/nxld/p/6376496-.html,2021.11.14

[5] 飞翔得蓝鲸，聚类分析常用算法原理：KMeans,DBSCAN,层次聚类，https://blog.csdn.net/leonliu1995/article/details/78944798，2021.11.14