肥料登记数据分析

摘要

随着施肥技术的进步，农民普遍意识到“平衡施肥”和“测土施肥”的重要性。本文围绕肥料登记数据，对数据进行整理、分类、归纳和比较，利用EXCEL和PYTHON进行数据分析，求解问题。

关键词： 条件循环语句 K-Means聚类 矩阵 树

1 问题重述

1.1 问题的背景

近年来随着施肥技术的进步，农民普遍意识到“平衡施肥”和“测土施肥”的重要性。[1]因此对肥料登记数据进行整理、分类、归纳和比较，从肥料品种、养分含量及配比等情况分析肥料发展现状及存在的问题，可以为肥料的生产和使用提供一些有益的借鉴。

1.2 问题的重述

根据肥料登记数据，用适当的统计、分析方法对数据进行分析、汇总和理解，并解决以下问题：

（1） 预处理肥料数据，包括按照肥料类别规范化处理和计算总养分百分比；

（2）根据养分百分比的取值对复混肥料产品和有机肥料产品进行细分；

（3）对肥料产品进行多维度对比分析，维度包含省份、日期、企业等；

（4）处理文本数据，设计算法在指数指标和各种原料的名称及百分比中提取出所需的结构化数据。

2 问题的分析

针对问题一，需要设计算法读取出所有名称，按照复混肥料、有机—无机复混肥料、有机肥料和床土调酸剂4种类别将所有名称进行粗略分类，

针对问题二，根据研究对象的不同分为对复合肥料产品进行数据分析和对有机肥料进行数据分析。

针对问题三，将日期数据进行预处理后，对复混肥料中不同组别不同年份绘制直方图或是折线图进而观察变化趋势，采用同样的数据分析方法，分别分析广西、湖北有机肥料有效产品中排序得到的前五个标签中分布情况。

针对问题四，将非结构化数据转化成结构化数据。

3 模型的假设

（1）假设所收集的数据真实，且能够反映具体的情况。

（2）假设文中所引用的文献和结论均正确且可靠。

5 问题一的求解

5.1 名称规范化处理

第一步：读取出附件 1 中的所有名称。

通过结果分析发现名称存在以下不规范的地方：

（1）肥料名称错误。如“掺混肥料”、“稻苗床土调酸剂”；

（2）名称存在不应该存在的空格。如“ 有机化肥”、“有机　有机复混肥料”等；

（3）存在全角字符和半角字符的错误，如“有机—无机复混肥料”等；

（4）输入格式不规范，会存在如“/n 有机肥料/n”、“有机肥料/n”的错误。

5.2 规范化处理结果

5.3 求和计算总无机养分百分比

6 问题二的求解

6.1 对复合肥料产品进行数据分析

6.1.1 数据预处理

6.1.2 数据预处理结果

其中登记数量最大的前三个分组及对应的产品登记数量为：第 7 组 2098 个，第 6 组 1470 个，第 5 组 1154 个。

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

市场上有高、中、低浓度系列复混肥料，国家标准规定低浓度总无机养分含量一般在 25%至 30%之间，中浓度在 30%至 40%之间，高浓度在 40%以上。要因地域、土壤、作物不同，选择使用经济、高效的复合肥。从直观图可以看出绝大部分的企业会选择生产总无机养分百分比区间在[28.8%,50.4%]的复混肥料，是因为不同的土壤、不同的作物需要的营养元素的比例是多样的，常见的水稻专用肥要求总无机养分百分比≥25%，毛竹专用肥要求总无机养分百分比≥25%，无公害的蔬菜肥要求总无机养分百分比≥25%[2]。可见大部分的作物需要的肥料总无机养分区间在 28.8%至 50.4%之间，需求大，生产的厂家也较多。

6.1.3 K-Means 聚类

聚类的方法主要有：划分方法、分层类方法、基于密度方法、基于网格方法和基于模型方法[3]。其中基于划分方法的 K-均值聚类（K-Means）因其具有算法复杂程度较低、效率高、应用领域广、 具有一定的可扩展性等优点，本问题的求解在按照氮、磷、钾养分的百分比对产品进行分类时，利用了 K-Means 算法的思想。

K -means 算法步骤如下：

（1）输入：聚类个数k，以及包含n个数据对象的数据库；

（2）从n个数据对象任意选择k个对象作为初始聚类中心；

（3）循环（4）到（5）直到每个聚类不再发生变化为止；

（4）根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离； 并根据最小距离重新对相应对象进行划分；

（5）重新计算每个（有变化）聚类的均值（中心对象）；

（6）输出：满足方差最小标准的 k 个聚类。

具体的计算过程如图 6 所示：

6.1.4 K-Means 聚类结果

聚类将产品分为了 4 类，聚类结果如图 7 所示，并将用数字 1-4 将产品进行标签，并保存至附件“result2_3”。

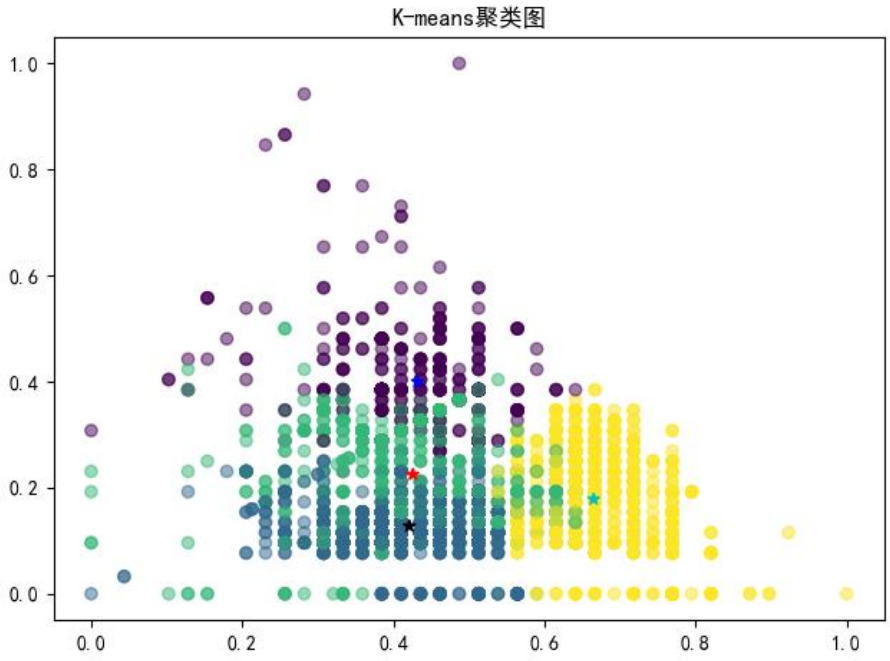

图 1 聚类结果

注：蓝色是类别 1，黑色是类别 2，红色是类别 3，青色是类别

根据聚类结果，将其映射到原数据的 3 维空间中，可得三维散点图如图 8 所示。

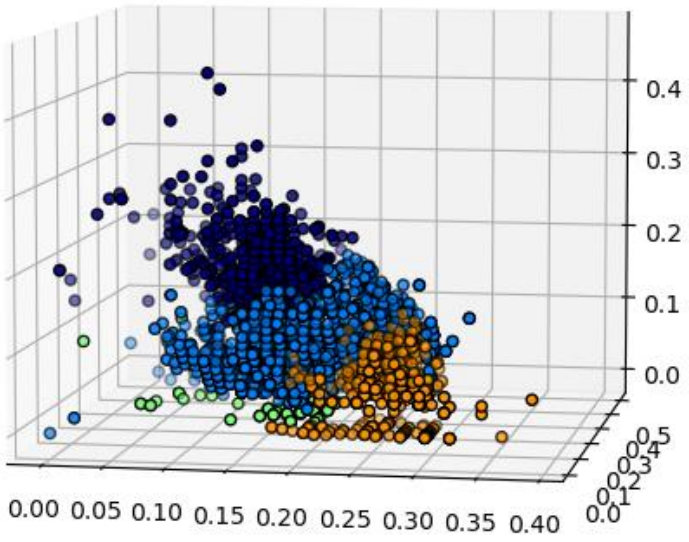

图 2 3D 三维散点图

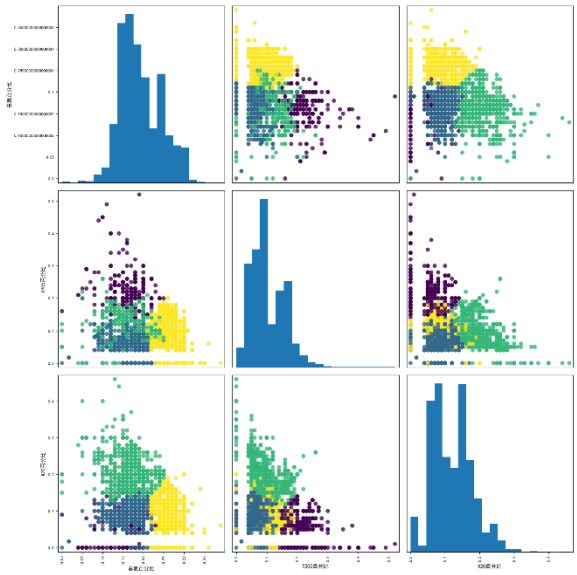

图 3 散点矩阵

6.1.5 雷达图

根据聚类结果绘制雷达图如图 10 所示。

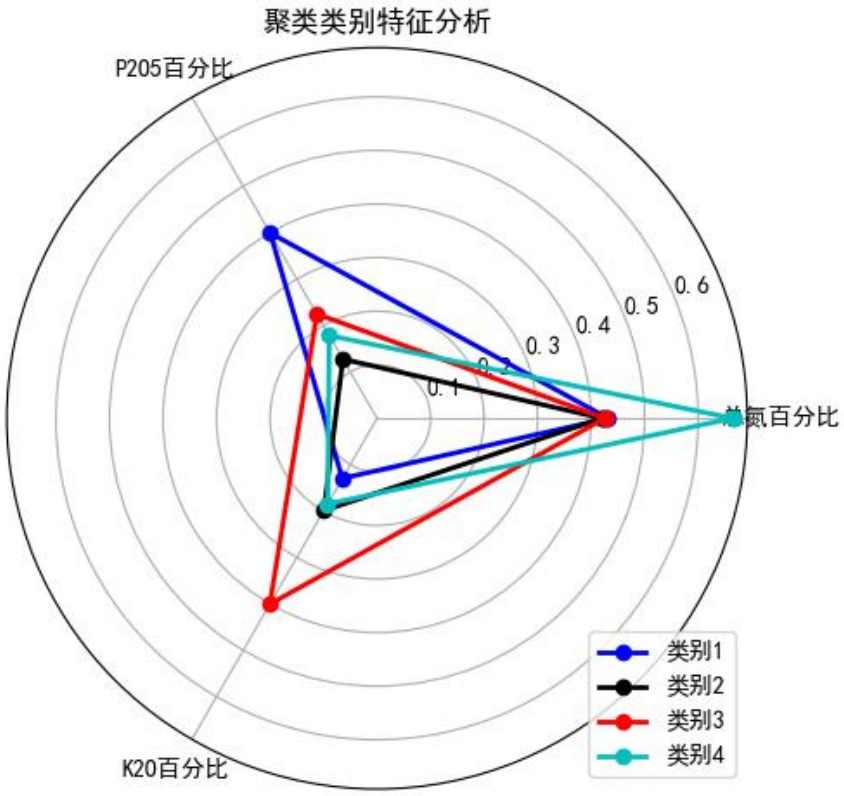

图 4 雷达图

7.问题三的求解

7.3 计算企业间的杰卡德相似系数矩阵

7.3.1 提取登记数量大于 10 的肥料企业

7.3.2 划分各企业的原料集合

7.3.3 杰卡德相似系数矩阵

1、杰卡德相似系数（Jaccardsimilarity coefficient）

两个集合 A 和 B 交集元素的个数在 A、B 并集中所占的比例，称为这两个集合的杰卡德系数，用符号 J(A,B) 表示。杰卡德相似系数是衡量两个集合相似度的一种指标，我们利用这个指标探究每两个企业之间化肥原材料的相似性。

假设 IDi是企业 i 的名称，A_i是企业原材料集合，

9 模型的评价与推广

在非结构化数据转化成结构化数据时，采用传统方法——树，这样的方法需要人工进行特征提取，操作繁琐且需要耗费大量人力进行数据标签。因为时间不充裕，所以未能从技术指标中提取出需要的结构化数据，从而未能完成任务 4.1 。时间充裕的话，可以学习和掌握新型利器—深度学习，进行数据清洗、异构数据和语义理解，从而解决传统方法中的耗费大量人力和时间进行数据标签的问题。

10 参考文献

[1]张福锁.测土配方施肥技术 [M] .北京: 中国农业大学出版社，2010.

[2]王连军,卢栋冬.国家标准《复合肥料》(GB/T 15063—2020)解析[J].肥料与健康,2021,48(05):65-69.

[3] Khan S S, Ahmad A. Cluster center initialization algorithm for K-means clustering [J]. Pattern Recognition Letters, 2004, 25(11): 1293-1302.