基于数据挖掘的肥料登记数据分析

摘要

我国是农业大国，也是肥料生产和使用大国，但目前的监督管理模式为市场监督管理、农业农村等部门均共同监管，导致多头监管多头执法，存在监管漏洞，从而导致肥料市场监管执法不畅。为规范肥料市场秩序，促进农业发展，维护农民合法权益，要强化监管主体责任，建议国家制定并出台《肥料管理条例》，并完善配套法律法规，将肥料生产、经销及推广使用纳入法制轨道，使各个阶段的管理有法可依。

关键词:数据处理 文本聚类 肥料 数据可视化

目录

一. 技术介绍

1.1 数据预处理

在任务实施的时候，我们得到的数据会存在有缺失值、重复值等，这些未经过处理的数据是无法直接进行数据挖掘，或挖掘结果差强人意，所以在使用之前需要进行数据预处理。数据预处理没有标准的流程，通常针对不同的任务和数据集属性的不同而不同。为了提高数据挖掘的质量而产生了数据预处理技术，其中数据预处理的主要内容包括数据清洗、数据集成、数据变换和数据规约。这些数据处理技术在数据挖掘之前使用，大大提高了数据挖掘模式的质量，降低实际挖掘所需要的时间。

数据的预处理是指对所收集数据进行分类或分组前所做的审核、筛选、排序等必要的处理，数据预处理过程如图 1 所示。

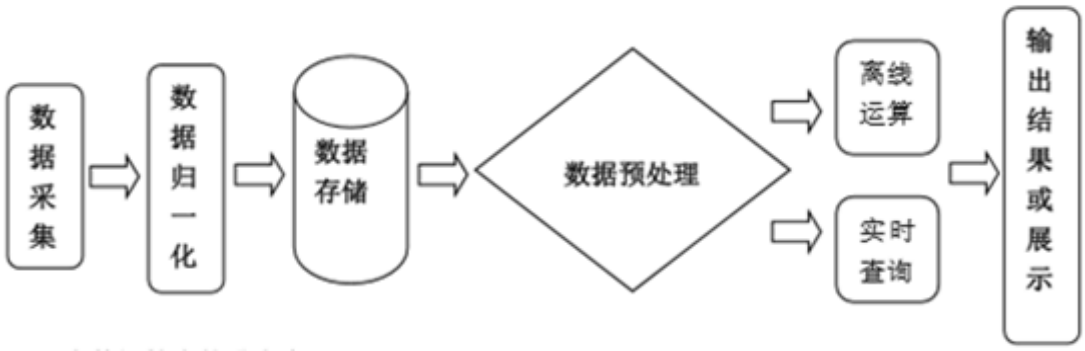

图 1 数据预处理过程

1.2 数据分析技术

1.2.1K-Means 聚类算法

K-Means 算法又名 k 均值算法，K-means 算法中的 k 表示的是聚类为 k 个簇，means 代表取每一个聚类中数据值的均值作为该簇的中心，或者称为质心，即用每一个的类的质心对该簇进行描述。

其算法思想大致为：先从样本集中随机选取 k 个样本作为簇中心，并计算所有样本与这 k 个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中，对于新的簇计算各个簇的新的“簇中心”。

K-Means 算法是最常用的聚类算法，主要思想是:在给定 K 值和 K 个初始类簇中心点的情况下，把每个点(亦即数据记录)分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的所有点重新计算该类簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数。

聚类属于非监督学习，K 均值聚类是最基础常用的聚类算法。它的基本思想是，通过迭代寻找 K 个簇（Cluster）的一种划分方案，使得聚类结果对应的损失函数最小。其中，损失函数可以定义为各个样本距离所属簇中心点的误差平方和如公式 1 所示。

1.2.2 杰卡德系数计算

杰卡德系数，英文叫做 Jaccardindex，又称为 Jaccard 相似系数，用于比较有限样本集之间的相似性与差异性。Jaccard 系数值越大，样本相似度越高。实际上它的计算方式非常简单，就是两个样本的交集除以并集得到的数值，当两个样本完全一致时，结果为 1，当两个样本完全不同时，结果为 0。针对杰卡德系数定义为，结合两个集合 A、B，杰卡德系数定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值，公式如 2 所示。

1.2.3 可视化工具

(1)图表秀，图表秀拥有丰富的可视化图表，提供数十种传统图表和高级可视化图表，涵盖地图、词云、弦图、散点图等。支持图表、文字、图片混排和任意拖拽布局，自定义主题、个性化定制数据分析报告。

支持设置多图表联动交互，可以 动态播放，展示图表间数据关系。并且可以导出图片、PPT、PDF、动态图册，还可以公开或私密的方式将图册便捷分享到微信、微博、QQ 等社交平台中。

(2)Chart.js，商业级数据图表，一个纯 Javascript 的图表库，可以流畅的运行在 PC 和移动设备上，兼容当前绝大部分浏览器（IE6/7/8/9/10/11，chrome，firefox，Safari 等），底层依赖轻量级的 Canvas 类库 ZRender，提供直观，生动，可交互，可高度个性化定制的数据可视化图表。创新的拖拽重计算、数据视图、值域漫游等特性大大增强了用户体验，赋予了用户对数据进行挖掘、整合的能力。

支持折线图（区域图）、柱状图（条状图）、散点图（气泡图）、K 线图、饼图（环形图）、雷达图（填充雷达图）、和弦图、力导向布局图、地图、仪表盘、漏斗图、事件河流图等 12 类图表，同时提供标题，详情气泡、图例、值域、数据区域、时间轴、工具箱等 7 个可交互组件，支持多图表、组件的联动和混搭展现。

二. 任务一

2.1 任务简述

截止任务实施前，根据题意得知，本任务需要大量的数据整合，利用 Python 中的 pandas 库，进行肥料产品的规范化分类处理。并且需要计算氮、磷、钾总无机养分，分别占比的营养养分。

2.3 任务实施

2.3.1 缺失值查询与去除

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2925 entries, 0 to 2924
Data columns (total 12 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   序号       2925 non-null   int64
 1   企业名称     2925 non-null   object
 2   产品通用名称   2925 non-null   object
 3   产品形态     2925 non-null   object
 4   总氮百分比    2925 non-null   float64
 5   P2O5百分比  2925 non-null   float64
 6   K2O百分比   2925 non-null   float64
 7   含氯情况     2925 non-null   object
 8   有机质百分比   2925 non-null   float64
 9   正式登记证号   2925 non-null   object
 10  发证日期     2925 non-null   object
 11  有效期      2925 non-null   object
dtypes: float64(4), int64(1), object(7)
memory usage: 274.3+ KB
```

图 2 附件 1 缺失值查询结果

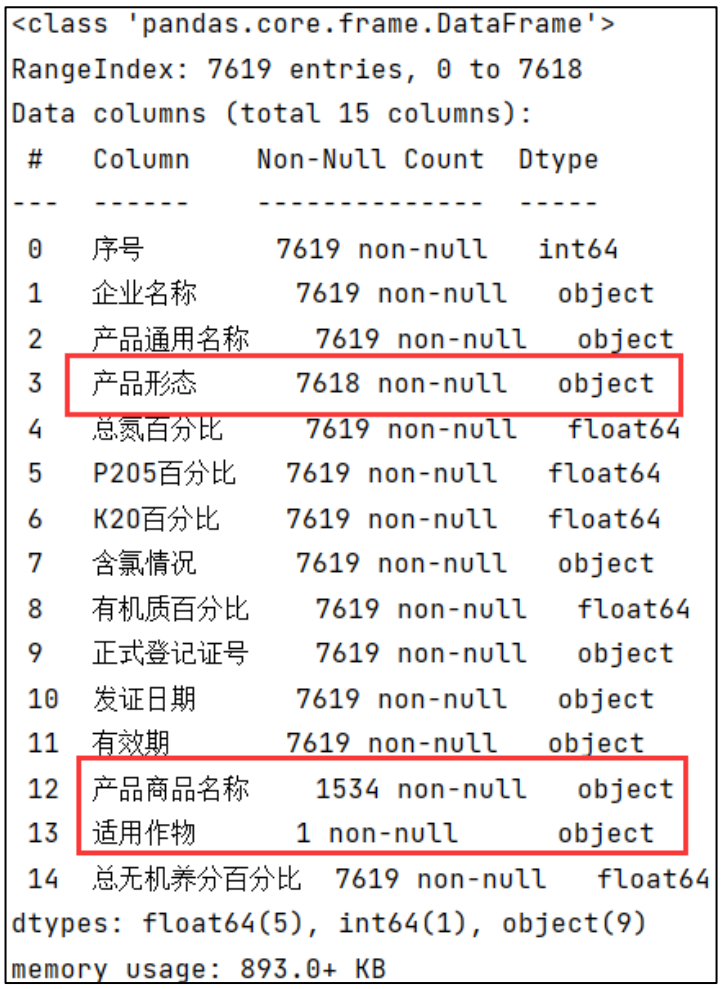

图 3 附件 2 缺失值查询

2.3.2 数据规范化处理

2.3.3 针对总无机养分百分比计算求和

三. 任务二

3.1 任务简述

此任务是对肥料产品的数据分析，需要从复混肥料、有机肥料角度去筛选取值。将所有复混肥料按照总无机养分百分比，进行等距取值，并且按照登记数量，从大到小列出登记数量值最大的前 3 个分组及相关产品登记数量。针对有机肥料产品，按照总无机养分百分比和有机质百分比分别等距分为 10 组，并

且进行热力图可视化呈现，重复上述登记要求。筛选复混肥料产品，按照无机养分百分比占比，进行聚类算法，绘制散点图和三维散点图进行描述。绘制雷达图进行聚类特征分析。

3.3 任务实施

3.3.1 复混肥料筛选

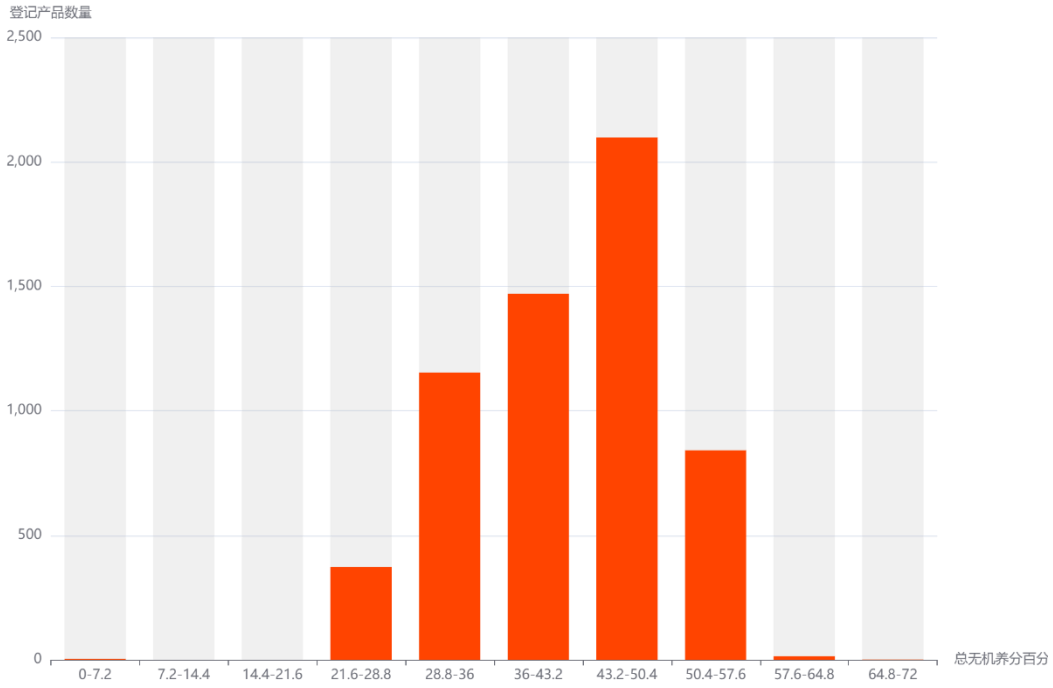

图 6 产品登记数量直方图

我们由上图可以了解到无机养分百分比在 28.8%-57.6%的登记产品数量最多，而其中前三个分组及相应登记产品数量如表 4 所示。

表 4 登记数量前三的分组

排名	一	二	三
分组标签	7	6	5
产品登记数量	2098	1407	1154

3.3.2 有机肥料筛选

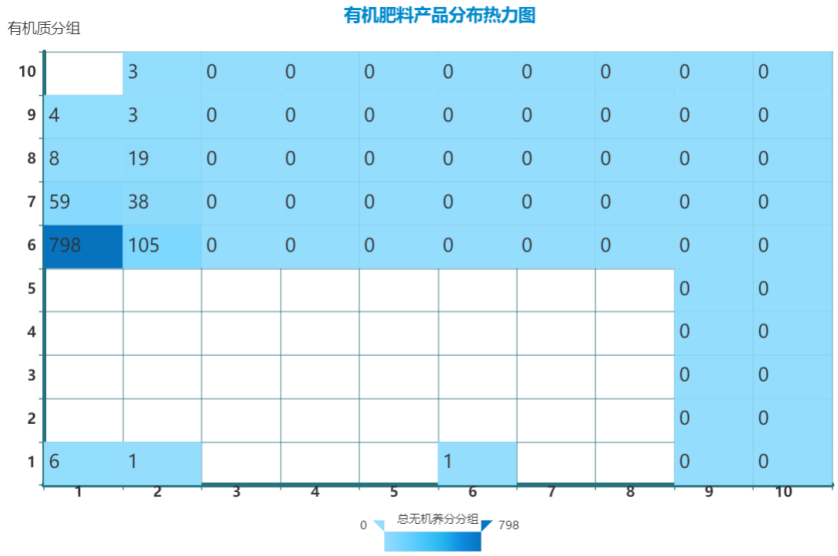

图 7 有机肥料产品分布热力图

3.3.3 复混肥料无机养分占比分析

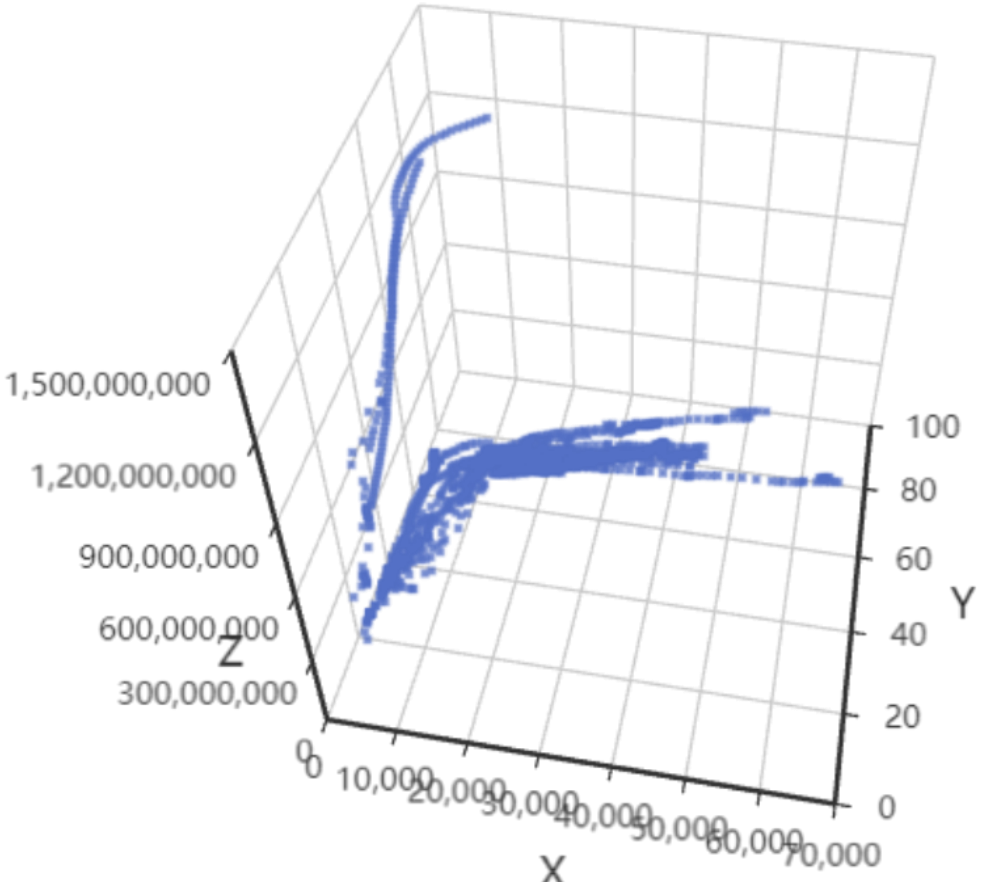

图 9 聚类后肥料产品三维散点图

图 10 氮磷钾三种养分散点图

散点图虽然可以直观的观察三种养分的占比情况，但是无法清晰的了解到此类养分的具体，聚类标签特征，针对这种现象我们有目的制造雷达图。去分析聚类后的标签特征。雷达图如图 11 所示。

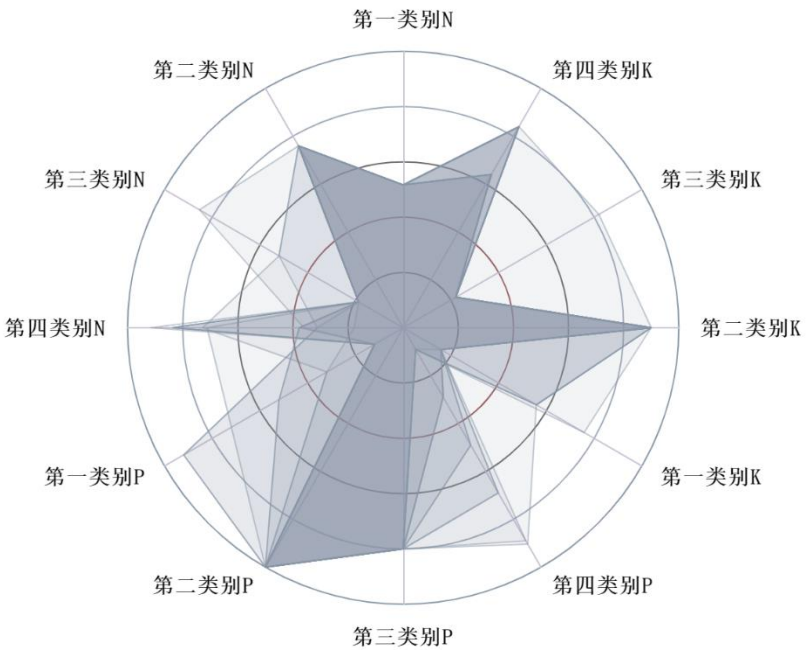

图 11 雷达图展示

四. 任务三

4.1 任务简述

此项任务是进行肥料产品的多维度对比分析。我们需要获取企业提取发证日期的年份。分析比较复混中各组别不同年份产品的登记数据变化趋势，将结果可视化呈现。筛选至 2021 年 9 月 30 日肥料产品依旧有效的有机肥料产品，并且在有效的有机肥料产品里面分别筛选广西和湖北产品登记数量在前 5 的组别，分析两个省份在组别中的差异。提取附件 3 中产品登记数量前 10 的肥料企业。以企业用到的原料为特征，计算企业间的杰卡德相似系数矩阵。

五. 任务四

5.1 任务简述

本项任务需要从附件 4 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。在提取前，我们需要设计算法，建立算法模型去进一步计算此类百分比。进而我们也需要从附件 4 原料与百分比中提取各种原料的名称及其百分比。同样需要设计算法模型去进行算术计算。

总结

我国属于农业大国，肥料对改变传统的农业生产方式有着十分重要的作用，可以改善土壤，为农作物提供更多营养，提升农业产量。我们针对此次肥料登记进行数据分析，从而达到想要了解的趋势以及不同地区登记差异等重要信息。

参考文献

[1]陶振水，周辉，王涛.化肥发展探讨[J].现代农业科技.2018/1/20

[2]陈宏坤，王志刚.我国复合肥产业发展现状、机遇与挑战[J].磷肥与复肥.2018/12

[3]张利军.土壤肥料在农业可持续发展中的地位和作用[J].新农业.2020(19)

[4]杨成梅.土壤肥料在农业持续发展中的地位和作用分析[J].农业与技术.2018(10)

[5]刘星毅，农国才.几种不同缺失值填充方法的比较[J].南宁师范高等专科学校学报，2007

[6]晔沙.数据缺失及其处理方法综述[J].网络与信息工程，2017

[7]xiaojimanman.文本聚类算法介绍[R].CSDN.2015-04-10