肥料登记数据分析

摘要

肥料是农业生产中的一种重要的生产资料，其生产销售必须遵守《肥料登记管理办法》，依法在农业行业行政管理部门进行登记。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肥料登记工作，相关数据可从政府网站上自由下载。通过数据分析技术对肥料登记进行研究可以有效的帮助政府工作人员进行肥料登记管理，所以具有一定的意义。

目录

1、问题分析

2、任务一

2.1 规范化处理

已知数据的肥料产品通用名称存在不规范的情况。我们按照复混肥料（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 种类别产品通用名称进行规范化处理。

3、任务二

3.1 等距分组和分布

3.1.2 分布分析

复混肥料总氮百分比大多分布在百分之十二到百分之三十之间，P205 百分比大多分布在百分之零到百分之二十之间，K20 百分比多分布在百分之零到百分之二十之间，有机质百分比全都为零，总无机养分百分比大多分布在百分之二十四到百分之五十四之间。

有机肥料产品的热力分布图如下图所示，通过以上步骤后分析得有机肥料总氮、P205 和 K20 百分比大都分布在百分之一点七到百分之四点七之间，有机质百分比集中在百分之四十五附近，有机肥料产品全都无氯，总无机养分百分比在百分之五到百分之二十之间。

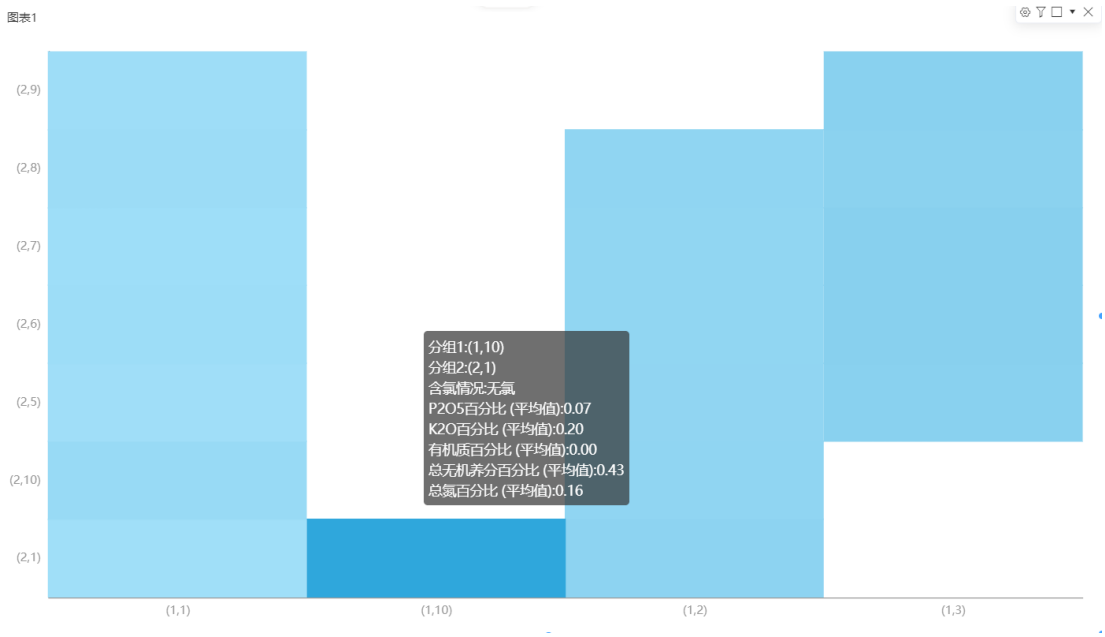

图 3.1 有机肥料产品的分布热力图

3.2 聚类算法

3.2.1 使用聚类算法分类

通过 Python 编写使用聚类算法将产品分类。在编写聚类算法时用到欧拉距

离公式与判别函数 Y=a1x1+a2x2+...anxn。通过欧拉公式计算距离。通过计算出新一轮的队列中心。

k 均值聚类算法是一种迭代求解的聚类分析算法，其步骤是，预将数据分为 K 组，则随机选取 K 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

3.2.2 聚类结果雷达图

通过 Python 绘制聚类结果雷达图。

3.2.3 分析聚类特征

每个聚类都有自己的特征。

4、任务三

4.3 杰卡德相似系数矩阵

杰卡德相似系数：

给定两个集合 A,B，Jaccard 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，定义如下：

当集合 A，B 都为空时，J(A,B)定义为 1。

与 Jaccard 系数相关的指标叫做 Jaccard 距离，用于描述集合之间的不相似度。Jaccard 距离越大，样本相似度越低。公式定义如下：

其中对参差（symmetric difference）

性质

6、参考文献

[1]Jaccardsimilarity coefficient. 相关系数之杰卡德相似系数[B]. csdn
[2]王秀文,郭明鑫,王宇韬. 超简单：用 Python 让 Excel 飞起来[B]. 机械工业出版社
[3]黑马程序员. 360° 解读机器学习经典算法——聚类算法. [B]机械工业出版社
[4]Jaccard.百度百科
[5]许培新,丁梁锋. 重视有机肥安全问题[N]. 光明日报,2005-03-17
[6]孙东辉. “贵族肥料”不再贵[N]. 人民日报