第四届“泰迪杯”数据分析技能赛

优秀报告

作品名称：肥料登记数据分析

荣获奖项：一等奖并获泰迪杯

作品单位：华南师范大学

作品成员：林苗芳 姚星河 田颖盈

目录

1 数据预处理

1.1产品通用名称规范化

附件 1 中肥料的产品通用名称存在不规范的情况，需要对这部分数据进行规范化处理。本文先对这部分通用名称不规范的数据进行探查，可以观察到肥料的产品通用名称出现名称的命名没有按照规范命名、名称前后有大量空格和换行符、链接符的半角和全角没有进行规范、用不同数量的空格做连接符等情况，肥料的产品通用名称情况如表 1 所示。

表 1 肥料的产品通用名称情况

产品通用名称
有机肥料
有机肥料
掺混肥料
床土调酸剂
稻苗床土调酸剂
复混肥料
有机肥料
有机无机　复混肥料
有机无机　复混肥料
有机-无机　复混肥料
有机-无机复混肥料
有机－无机复混肥料

本文根据要求的命名规范并按照以下步骤进行规范化。

1) 将“掺混肥料”改名为“复混肥料”。

2) 将“稻苗床土调酸剂”改名为“床土调酸剂”。

3) 将“有机无机　复混肥料”、“有机无机　复混肥料”、“有机-无机　复混肥料”和“有机－无机复混肥料”四种链接符不规范的情况更改为“有机-无机复混肥料”。

4) 去除掉名称中的空格和换行符。

经过命名规范化后得到的名称情况如表 2 所示。

表 2 规范化后得到的名称情况

产品通用名称
床土调酸剂
复混肥料
有机肥料
有机-无机复混肥料

1.2 计算总无机养分百分比

总无机养分百分比为肥料产品的氮、磷、钾养分百分比之和，其中磷、钾的含量分别用五氧化二磷（P_2O_5）的质量和氧化钾（K_2O）的质量来计算。因此，总无机养分百分比的公式为：

$$\text{总无机养分百分比} = \text{总氮百分比} + P_2O_5\text{百分比} + K_2O\text{百分比}$$

本文将总氮百分比、P_2O_5百分比、K_2O百分比相加得到总无机养分百分比，并将计算结果保留三位小数。计算得到的部分结果如表 3 所示。

表 3 总无机养分百分比部分结果

序号	正式登记证号	总无机养分百分比
1	皖农肥(2016)准字 4255 号	50.0%
2	皖农肥(2016)准字 4256 号	50.0%
3	皖农肥(2016)准字 4257 号	51.0%
4	皖农肥(2016)准字 4258 号	51.0%
5	皖农肥(2016)准字 4259 号	51.0%
6	皖农肥(2016)准字 4260 号	51.0%
7	皖农肥(2016)准字 4261 号	52.0%
8	皖农肥(2016)准字 4262 号	52.0%
9	皖农肥(2016)准字 4263 号	54.0%
10	皖农肥(2016)准字 4264 号	54.0%

2 肥料产品数据分析

2.1 筛选复混肥料并分组

观察附件 2 的数据，产品的通用名称是规范的，分为以下四类肥料产品：

表 4 附件 2 产品分类

名称	复混肥料	有机肥料	有机-无机复混肥料	床土调酸剂
计数	5954	1045	611	9

接着，我们筛选出复混肥料的产品，将其保存在变量 mix 中。然后获取总

无机养分百分比的取值范围 mix_min、mix_max 和组距 delta，其中取值范围为 0.0~0.72，以此将数据分为 10 组，组距为 0.072。

表 5 复混肥料产品的无机养分百分比取值范围及组距

mix_max	mix_min	delta
0	0.72	0.072

在变量表 mix 中创建空的标签列，按照 mix 中总无机养分百分比的分组范围为数据打上标签，最后保存在名为“result2_1.xlsx”的文件中。

本文统计了十个分组中各组复混肥料的产品登记数量，绘制了产品登记数量的直方图，并且展示登记数量最大的前 3 个分组及其对应的产品登记数。结果分别如图 1 和表 6 所示。

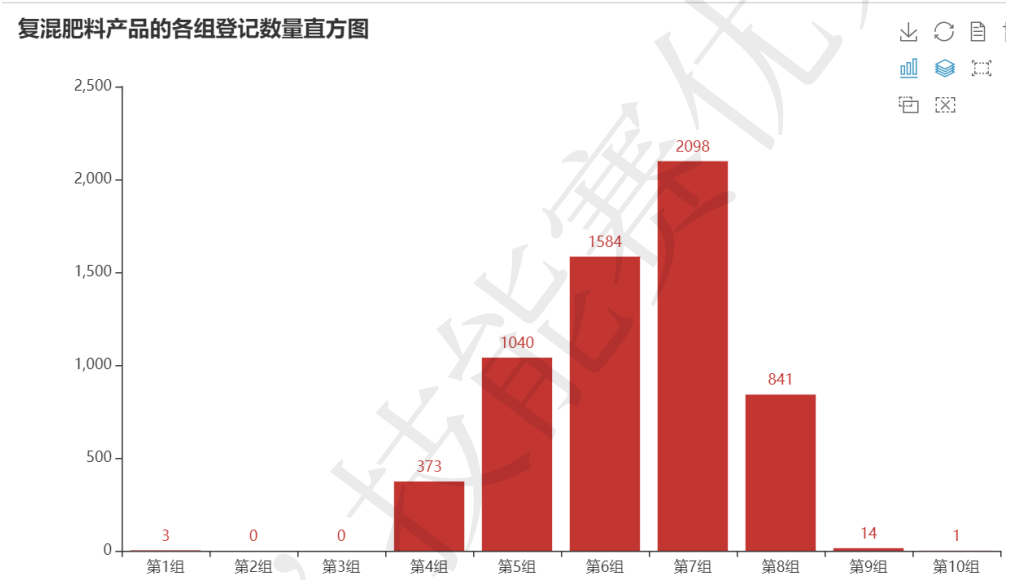

图 1 复混肥料产品各组登记数量直方图

表 6 肥料登记数量前 3 名的组别

分组标签	登记数量
7	2098
6	1584
5	1040

2.2 筛选有机肥料并分组

2.2.1 制定产品分组标签

对附件 2 的数据，首先筛选出有机肥料的产品，将其保存在变量 organic 中。

接着，进一步获取和计算总无机养分百分比的取值范围 organic_min1、organic_max1 和组距 delta1，有机质百分比的取值范围 organic_min2、

organic_max2 和组距 delta2。 结果如表 7 所示。

表 7 有机肥料产品的属性取值范围及组距

organic_min1	0.0501	**organic_min2**	0.0
organic_max1	0.43	**organic_max2**	0.9
delta1	0.03799	**delta2**	0.09

在变量表 organic 中创建空的标签列：标签 1、标签 2、标签。其中，按照 organic 中总无机养分百分比的分组范围为标签 1 打上标签，按照 organic 中有机质百分比的分组范围为标签 2 打上标签。再将标签 1 和标签 2 的结果构造成“(标签 1,标签 2)”的格式，赋值给标签列。最后将结果保存在名为“result2_2.xlsx”的文件中。

2.2.2 有机肥料产品的分布热力图

根据有机肥料的分组情况，本文以总无机养分的分组标签为横轴，以有机质的分组标签为横轴，绘制有机肥料产品的分布热力图。结果如图 2 所示。

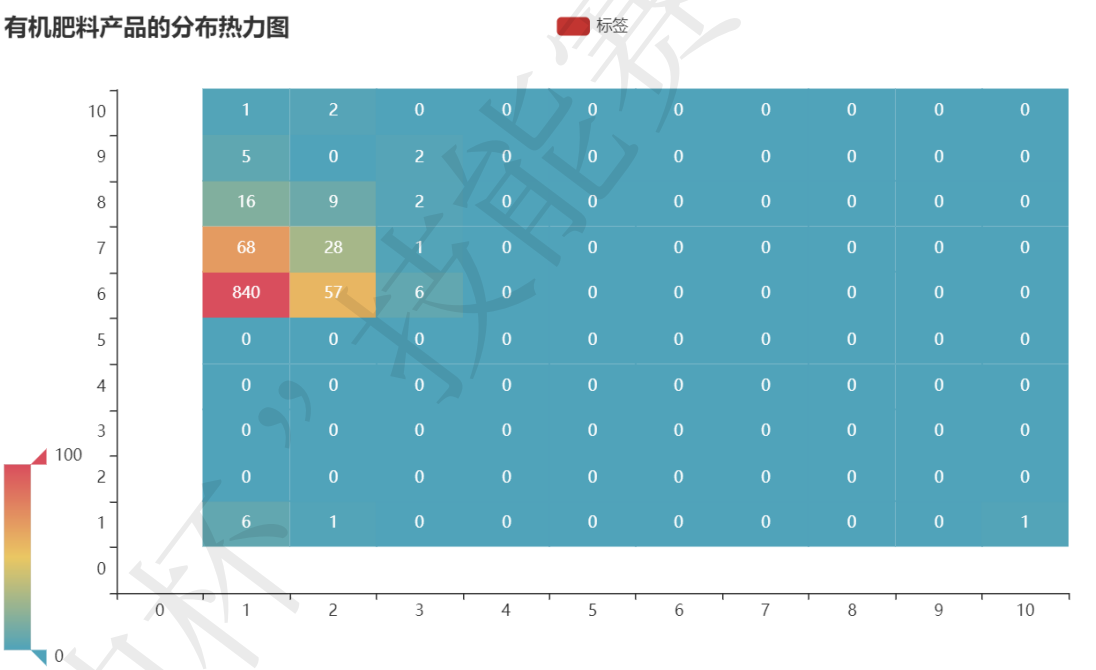

图 2 有机肥料产品的分布热力图

观察图 2，可以发现，有机肥料产品在总无机养分百分比的分组大部分都为 1、2 组，即有机肥料产品的总无机养分百分比大部分都处于 0.0501~0.12608 之间。而有机肥料产品在有机质百分比的分组大部分都为 6、7、8 组，即有机肥料产品的有机质百分比大部分都处于 0.45~0.72 之间。

本文对登记数量最大的前 3 个分组及相应的产品登记数量进行展示，结果如表 8 所示。

表 8 有机肥料登记数量前 3 名的属性表

总标签	总无机养分标签	有机质标签	登记数量
(1,6)	1	6	840
(1,7)	1	7	68
(2,6)	2	6	57

2.3 复混肥料聚类分析

2.3.1 K-Means 聚类

聚类通过对无标记的训练样本进行分析，从而得到数据间的内在联系和规律，进而将数据集划分成若干个互不相交的子集。本文采用 K-Means 聚类算法对复混肥料以氮、磷、钾养分的百分比作为特征集进行聚类。接下来，本文简要介绍 K-Means 聚类算法的原理。

K-Means 聚类算法是基于划分的聚类算法中的一种。基于划分的聚类算法是最简单的一种聚类算法，其思想是：在数据集中选择指定数量的样本点作为“质心”，计算所有样本点与“质心”的距离，将样本点归纳到最小距离的质心所在的聚簇中。之后，根据聚簇中的样本点更新“质心”并且重新划分数据集中的样本点，反复迭代优化直至“质心”不再发生变化。

K-Means 聚类算法表述如下所示。

算法 1 K-Means 聚类算法流程

Algorithm 1 K-Means Clustering Algorithm

输入：数据集$D=\{X_1,X_2,\cdots,X_n\}$; Number of clusters K

输出：A collection of partition-based clusters $C=\{C_1,C_2,\cdots,C_K\}$

从D中随机选择K个样本点$A=\{A_1,A_2,\cdots,A_K\}$作为初始的聚类中心；

重复以下步骤：

  令$C_i=\phi\ (1\leqslant i\leqslant K)$

  for $j=1,2,\cdots n$ **do**

    计算数据集中的每个样本点X_j与每个簇的中心点A_i之间的欧几里得距离$d(X_j,A_i)$

    根据最近距离原理，将样本点X_j划分为相应的类簇

  end for

  for $i=1,2,\cdots K$ **do**

    计算簇的新聚类中心点C_i：$A_{i-new}=\frac{1}{|C_i|}\sum_{X\in C_i}X$

    if $A_{i-new}\neq A_i$

      更新聚类中心:$A_i=A_{i-new}$

  end for

Until 聚类中心集A不再发生任何变化

2.3.2 复混肥料聚类

基于 K-Means 聚类算法原理，本文利用该算法对复混肥料进行聚类。

首先提取复混肥料产品中的氮、磷、钾养分的百分比并且将其作为聚类的输入变量。本文设定聚类数目为 4 类，并将聚类结果放入 mix 变量表的“聚类标签”中。其中，K-Means 聚类的聚类中心如表 9 所示。

表 9 聚类中心向量

聚类中心	总氮百分比	P_2O_5百分比	K_2O百分比
类型 1	0.25918936	0.09505569	0.08105198
类型 2	0.16483179	0.06633649	0.08903241
类型 3	0.16157704	0.17772833	0.12993139
类型 4	0.1751335	0.07570388	0.19924757

最后，将完整的结果保存在名为“result2_3.xlsx”的文件中。

2.3.3 聚类结果分析

首先，根据复混肥料产品中的氮、磷、钾养分的百分比数据绘制肥料产品的三维散点图(如图 3 所示)，其中，点都颜色由聚类标签决定，一共分为四种，分别为红色、蓝色、绿色、黄色。

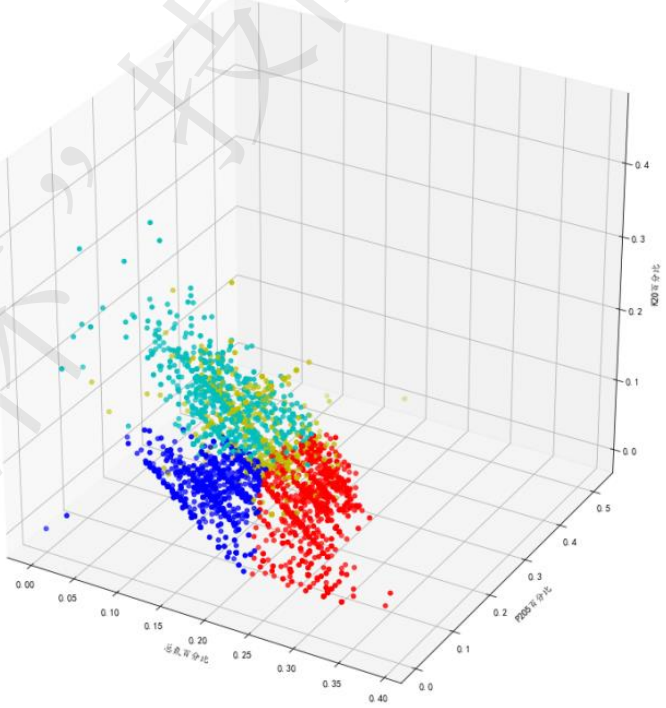

图 3 复混肥料聚类特征三维散点图

观察图 4，可以知道 K-Means 聚类的聚类效果很好，很大程度上将三维数据集分成了四类，并且分界线较为明显。

其次，再根据复混肥料产品中的氮、磷、钾养分的百分比数据绘制散点图矩阵(如图 5 所示)。其中，以聚类标签区分点的颜色，蓝色为第一类，黄色为第二类，绿色为第三类，红色为第四类。

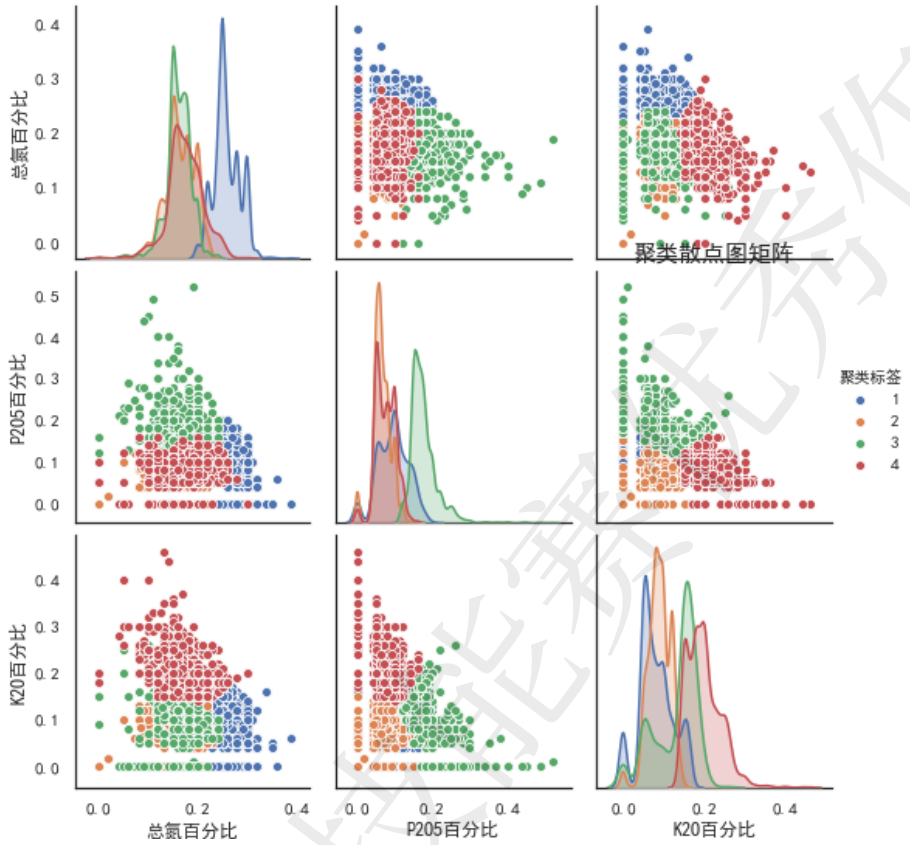

图 5 复混肥料特征散点图矩阵

观察散点图矩阵，可以清楚的看到复混肥料产品中氮、磷、钾养分百分比的两两对比结果。对于每一个二维特征的比较，K-Means 聚类几乎都是将数据点集分为三部分，一部分数据点是两特征的百分比都较低，另外两部分是其中一个特征的百分比高，另一个特征的百分比较低。

最后，本文绘制了聚类结果的雷达图来对每个聚类进行进一步的特征分析。聚类结果雷达图如图 6 所示。

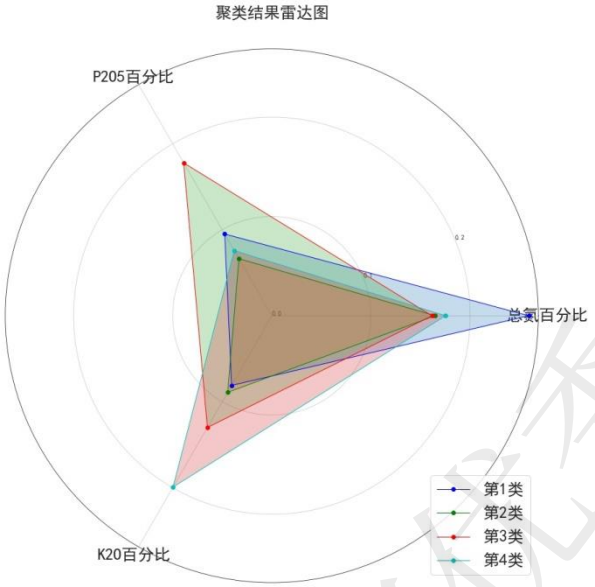

图 6 聚类结果雷达图

分析聚类结果雷达图，可以得到，聚类结果为“类型 1”(蓝色)的复混肥料的总氮百分比较高；聚类结果为“类型 2”(绿色)的复混肥料的P_2O_5百分比较高；聚类结果为“类型 3”(粉色)的复混肥料的K_2O百分比较高；而聚类结果为“类型 4”(蓝绿色)的复混肥料的三种属性与其它类型的复混肥料相比均偏低。

3 肥料产品多维度对比分析

3.1 复混肥料产品登记数量变化趋势

基于 2.1 得到的“result2_1.xlsx”文件，本文提取肥料产品发证日期中的年份，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。

观察“result2_1.xlsx”文件中的发证日期，发现发证日期的数据类型为字符型。因此，本文通过字符串提取技术，提取发证日期中的年份，并储存在为“发证年份”的标签列中。

接着，本文统计了复混肥料中各组别不同年份产品登记数量，并以时间为横轴，登记数量为纵轴，绘制折线图，如图 6 所示。

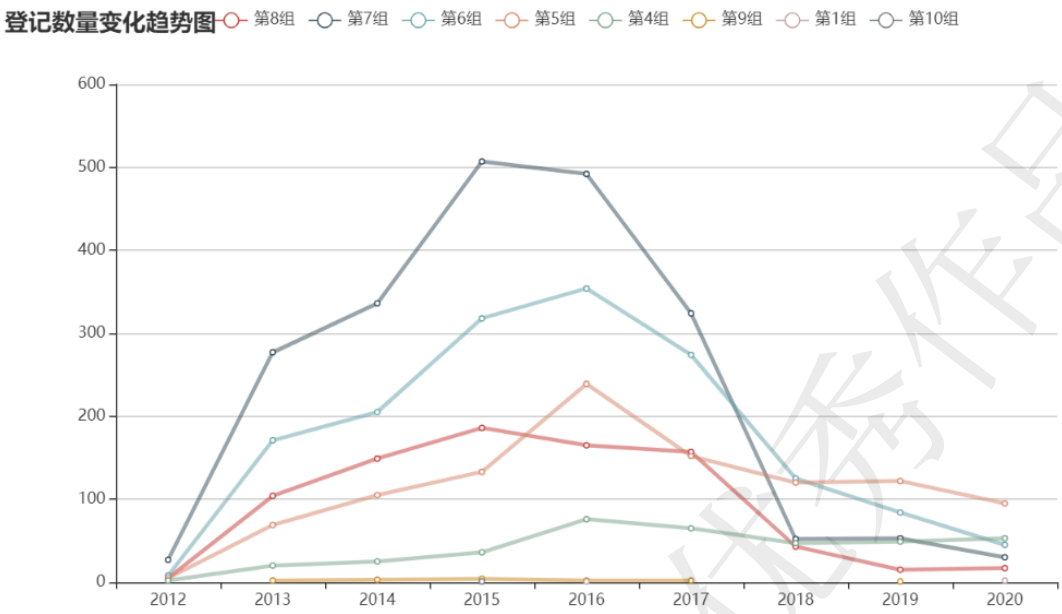

图 7 登记数量变化趋势图

由登记数量变化趋势图(图 6)，可以看到：复混肥料近年来的登记数量总体呈现先上升后下降的趋势。除此以外，历年的登记数量排序基本为“第 7 组>第 6 组>第 5 组>第 8 组>第 4 组”，其中第 9、1、10 组历年的登记数量基本上为 0。第 7 组与第 6 组的复混肥料登记数量在 2018 年前的下降幅度较大，而第 8 组的复混肥料登记数量在 2018 年后的下降幅度较大；相比之下，第 4 组、第 5 组的登记数量变化平缓，较为稳定。

因此，本文可以得出以下结论：

1、市面上对复混肥料的总无机养分百分比需求一般处 0.216~0.576 之间，过高或过低都会难以被市场接受，从而导致其登记数量基本为 0。

2、市场对于总无机养分处于 0.36~0.504 之间的复混肥料的需求较之以往下降许多，导致其登记数量在 2018 年之前下降许多，但在 2018 年后已趋于平稳。

3、市场对于高总无机养分(0.504~0.576)的复混肥料的需求在近几年来开始降低。

4、市场对总无机养分为中浓度(0.216~0.36)的复混肥料的需求一直都较稳定。

3.2 有效有机肥料产品分析

基于 2.2 得到的“result2_1.xlsx”文件，本文从中提取出 2021 年 9 月 30 日仍有效的有机肥料产品。

观察“result2_1.xlsx”文件中的有效期字段，发现发证日期的数据类型为字符类型。因此，本文通过字符串筛选技术，将有效期大于等于'2021-10'的数

据取出并储存在文件“result3_2.xlsx”中

其次，本文基于正式登记证号，从有效产品中分别筛选出广西省和湖北省产品登记数量排名前 5 的组别及其对应的和产品登记数，结果分别所下表示。

表 10 广西省产品登记数排名前 5 的组别及其产品登记数

省	标签	产品登记数量
桂	(1,6)	346.0
桂	(1,7)	55.0
桂	(2,6)	48.0
桂	(2,7)	23.0
桂	(1,8)	9.0

表 11 湖北省产品登记数排名前 5 的组别及其产品登记数

省	标签	产品登记数量
鄂	(1,6)	387.0
鄂	(1,7)	11.0
鄂	(1,8)	5.0
鄂	(2,6)	3.0
鄂	(2,7)	2.0

观察两个省的登记数量前 5 的组别，可以发现广西省的产品登记量主要集中在总无机养分分组的第一组，但第二组也有 71 条数据；有机质分组的第六组，但第七组仍有 78 条数据。而湖北省的产品登记量主要集中在总无机养分分组的第一组，有机质分组的第六组，其余的各小组分类都不足 20 条数据。由此可以发现，湖北省和广西省的肥料的总无机养分百分比绝大部分处于 0.0501~0.08809，有机质百分比处于 0.45~0.54，但湖北省的肥料在这两个区间的集中趋势比广西省的要高，即肥料的元素含量更稳定。

3.3 企业相似系数

3.3.1 数据预处理

附件 3 给出了某省登记肥料的产品配方，观察数据发现数据表是一个高阶的稀疏矩阵，为了方便后续的数据处理及分析，我们首先对附件 3 中的空值进行处理。本文默认空值代表的含义为：肥料的配方中不含有该原料。由此，我们对空值以零值进行填充。

其次，观察肥料的产品配方原料表，共有 294 种原料。我们发现原料中有名为“16”、“无”和“种”的化学原料。经过相关文献查找和资料查询，我们

发现没有命名为“16”或“十六”以及“无”和“种”的与肥料相关的化学用品或原料。观察数据发现，仅有一个企业的一种肥料使用了原料“16”，因此，我们将配方表中的“16”和“无”作为异常值进行，直接删去。

除此以外，我们还发现有名为“尿素尿素”的原料，与化学原料“尿素”具有高度类似性。观察数据发现，“尿素尿素”在企业 ID01 的两款肥料产品中具有较大比重，且这个两款肥料产品的配方表中“尿素”的含量均为 0。根据相关文献资料的查询，我们得知，尿素是肥料产品中重要的原材料。因此，我们认为“尿素尿素”等同于“尿素”。我们将二者取并集，合成新的数据集

与“尿素尿素”问题相似的还有原料“高龄土”和“高岭土”，我们对其作一样的数据处理。

最终，我们得到预处理后的数据，共 550 个肥料产品的配方表，包含 289 个原料信息。

3.3.2 筛选产品登记数大于 10 的企业

接着，我们提取出产品登记数大于 10 的企业名称，结果如表 12 所示。

表 12 产品登记数大于 10 的企业名称及其登记数

企业名称	ID1	ID10	ID12	ID2	ID3
产品登记数	40	11	11	26	19
企业名称	ID4	ID5	ID6	ID7	ID9
产品登记数	18	18	14	13	11

根据 10 家企业的名称，我们从附件 3 中提取这 10 家企业的肥料配方数据。

3.3.3 提取原料集合

进一步地，我们提取这些企业所用到的原料集合。核心方法是：首先找到这些企业的肥料配方中大于 0 的所有原料，再利用去重函数 unique()获得原料集合。

经过统计，我们最终得到的原料集合（除去发酵菌剂）如表 13 所示。

表 13 肥料原料集合(除发酵菌剂)

七水硫酸锌	氯化铵	磷酸二铵	酵菌剂
七水硫酸镁	泥炭土	磷酸氢钙	钙镁磷
亮蓝	滤泥	粉状磷酸一铵	钙镁磷肥
农用氯化铵	烟粉	粘土	钾灰
填料	白云石粉	糖蜜	颗粒尿素
大颗粒尿素	白钾	糖蜜酒精废液	颗粒氯化钾
大颗粒红钾	硅钙粉	肥料级磷酸氢钙	颗粒氯化铵
小颗粒尿素	硝铵磷	腐殖酸	颗粒磷酸二铵
尿素	硫酸钾	草炭	颗粒过磷酸钙
桐枯	硫酸铵	蔗髓	高岭土
桐麸	硫酸锌	虑泥	黄腐酸钾

母粒	硼砂	车马硼砂	黑色防结块剂
氧化镁	碳铵	过磷酸钙	
氯化钾	磷酸一铵	酒精废液	

3.3.4 企业相似系数矩阵

基于提取的 10 家企业和原料集合，我们以各企业用到的原料为特征，计算企业之间的杰卡德相似系数矩阵。

首先，我们给出杰卡德相似系数矩阵的定义，如下所示：

集合A与集合B的杰卡德相似系数定义为$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$，其中$|S|$表示集合$S$中元素的个数。

其次，我们设计算法来计算两个企业间的相似系数。

Step 1 生成企业的原料使用 0-1 矩阵

生成一个新矩阵$D_{10\times 54}$来记录企业使用原料的情况。基于 3.3.3 中所得到的原料集合M，对于M中的原料m_j，若企业X_i使用了原料m_j，则记$D_{ij}=1$，否则$D_{ij}=0$。

Step 2 计算任两个企业原料使用的交集与并集

读取矩阵D中的数据，计算企业间原料使用的交集和并集的元素个数，分别储存在矩阵T和H中。

对于两个企业X_i和X_j的原料使用矩阵D_i和D_j，对D_i与D_j分别进行取交集和并集运算，在 Matlab 中，利用与运算和或运算进行实现。计算得到矩阵K^{ij}和Q^{ij}。

因此，我们只需要记录K^{ij}中元素值为 1 的个数，并储存在T_{ij}中，即可得到企业X_i和X_j的原料使用交集的元素个数。而记录Q^{ij}中元素值为 1 的个数，并储存在H_{ij}中，即可得到企业X_i和X_j的原料使用并集。

Step 3 计算企业的相似系数矩阵

基于矩阵T和H，我们可以求解得到企业的相似系数矩阵J的计算公式如下：

$$J(i,j)=\frac{T(i,j)}{H(i,j)}$$

基于上述算法流程，我们得到企业的相似系数矩阵J如下所示。

表 14 企业相似系数矩阵

1	0.25	0.2632	0.3043	0.3333	0.3684	0.2308	0.3333	0.375	0.1667
0.25	1	0.3125	0.2857	0.3125	0.2778	0.2083	0.4	0.3571	0.1905
0.2632	0.3125	1	0.3	0.4286	0.2941	0.2174	0.4286	0.5	0.2632
0.3043	0.2857	0.3	1	0.3684	0.3333	0.2143	0.3684	0.4118	0.2
0.3333	0.3125	0.4286	0.3684	1	0.4667	0.3333	0.5385	0.6364	0.3333
0.3684	0.2778	0.2941	0.3333	0.4667	1	0.3043	0.375	0.4286	0.2381

0.2308	0.2083	0.2174	0.2143	0.3333	0.3043	1	0.2727	0.3	0.28
0.3333	0.4	0.4286	0.3684	0.5385	0.375	0.2727	1	0.8	0.2632
0.375	0.3571	0.5	0.4118	0.6364	0.4286	0.3	0.8	1	0.2941
0.1667	0.1905	0.2632	0.2	0.3333	0.2381	0.28	0.2632	0.2941	1

其中，相似系数矩阵J的横轴和纵轴所代表的企业为：

ID1	ID10	ID12	ID2	ID3	ID4	ID5	ID6	ID7	ID9

我们将相似系数矩阵J保存在“result3_3.xlsx”文件中。

3.3.5 结果分析

经过企业相似系数的计算，我们认为：若两家企业相似系数大于 0.5，则这两家企业显著地相似。分析数据得到，企业 ID3 和 ID6、ID3 和 ID7、ID6 和 ID7 显著地相似。

4 肥料产品属性提取

4.1 氮、磷、钾、有机质百分比及含氯程度的提取

本文从附件 4 的技术指标中，提取氮、磷、钾养分和有机质的百分比，以及肥料含氯的程度。观察数据发现，附件 4 的技术指标是文本数据。由于我么们需要提取出氮、磷、钾养分和有机质的百分比，而氮、磷、钾养分的百分比又需要总养分的百分比，因此可以考虑用正则表达式进行数据的提取。但使用正则表达式需要数据具有相同的格式，因此先对技术指标中的数据进行规范化处理。本文按照以下流程提取出氮、磷、钾养分和有机质的百分比。

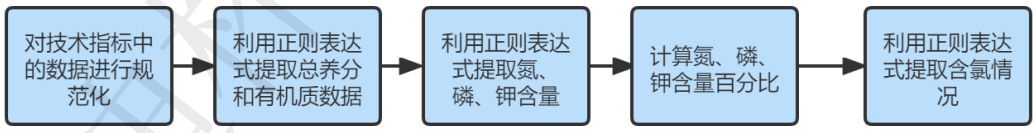

图 8 提取出氮、磷、钾养分和有机质的百分比流程

从技术指标中提取出氮、磷、钾养分和有机质的百分比流程中的具体步骤如下。

Step 1 对技术指标中的数据进行规范化

（1） 整理错误、错位、空的数据。

（2） 将全角符号改为半角符号，包括但不限于“%”、“（”、“）”、“-”。

（3） 去除所有空格、顿号、冒号和中文句号。

（4） 将不同格式的大于等于号全部更改为等号“=”。

（5） 将表示总养分和有机质的数据格式全部更改为包含“总养分=XX”，“有机质=XX”。将表示氮、磷、钾养分配比的数据格式更改为包含“XX-XX-XX”形式。（XX 为一定位数的数字）

Step 2 利用正则表达式提取总养分和有机质数据

（1） 利用正则表达式提取格式为“总养分=”的字符串后的数字部分作为总养分的百分比。

（2） 利用正则表达式提取格式为“有机质=”的字符串后的数字部分作为有机质的百分比。

（3） 将总养分的百分比和有机质的百分比中结果为空的值设置为 0。

（4） 将产品通用名称为“复混肥料”和“掺混肥料”的肥料的有机质百分比设为 0。

Step 3 利用正则表达式提取氮、磷、钾含量

（1） 利用正则表达式提取格式为“XX-XX-XX”的字符串后的数字部分作为氮、磷、钾的配比。

（2） 将氮、磷、钾的配比为空的值设置为 1-1-1。

（3） 将第一个破折号前的数字提取出作为氮的含量。

（4） 将两个折号中间的数字提取出作为磷的含量。

（5） 将最后一个破折号前的数字提取出作为钾的含量。

Step 4 计算氮、磷、钾含量百分比

（1） 计算氮、磷、钾含量的比例。

（2） 补充单独给出氮、磷、钾含量的数据。

Step 5 利用正则表达式提取含氯情况

（1） 利用正则表达式提取满足格式为“Y 氯”的字符串列表，Y 为一个中文字符。

（2） 对第一步中得到的列表进一步进行处理。若列表为空，则说明肥料产品的技术指标中没有给出含氯情况，将含氯情况的值设为“无氯”。若该列表中只含有一个元素且值为“含氯”，则将含氯情况的值设为“低氯”。若该列表中只含有一个元素，则该值必为“无氯”、“低氯”、“中氯”和“高氯”4 种中的一个，并将其值作为含氯情况的值。若列表中的值多于两个，提取列表中属于“无氯”、“低氯”、“中氯”和“高氯”的值作为含氯情况。

基于上述步骤，我们提取得到肥料产品的氮、磷、钾养分和有机质的百分

比，以及肥料含氯的程度，保存在名为“result4_1.xlsx”的文件中。本文展示部分提取结果，如表 15 提取得到的部分结果所示。

表 15 提取得到的部分结果

序号	产品通用名称	总氮百分比	P_2O_5 百分比	K_2O 百分比	有机质百分比	含氯情况
1	复混肥料	14.0%	8.0%	13.0%	0.0%	无氯
2	复混肥料	15.0%	6.0%	9.0%	0.0%	中氯
3	有机肥料	1.7%	1.7%	1.7%	45.0%	无氯
4	复混肥料	10.0%	18.0%	15.0%	0.0%	低氯
5	有机肥料	1.7%	1.7%	1.7%	45.0%	无氯
6	有机-无机复混肥料	10.0%	4.0%	6.0%	20.0%	低氯
7	有机肥料	1.7%	1.7%	1.7%	45.0%	无氯
8	有机肥料	2.0%	2.0%	2.0%	50.0%	无氯
9	有机肥料	2.3%	2.3%	2.3%	45.0%	无氯
10	有机肥料	2.3%	2.3%	2.3%	55.0%	无氯

4.2 原料名称及其百分比的提取

观察原料与占比的数据，可以发现大部分数据都满足“成分（占 XX%）”的格式，且每种成分之间以逗号分隔，因此可以考虑以逗号作为分隔符将各种成分分隔开并用正则表达式的方式进行提取。为了提取的简便程度和准确率更高，需要在提取前对数据进行预处理和规范化。本文按照以下流程提取各种原料的名称及其百分比。

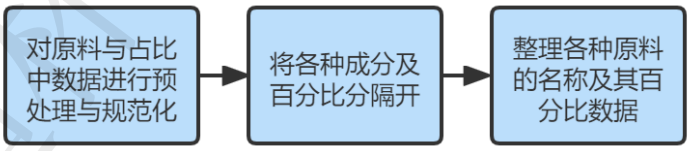

图 9 提取各种原料的名称及其百分比流程

从原料与占比数据中提取各项原料及其百分比的具体步骤如下。

Step 1 对原料与占比中数据进行预处理与规范化

（1） 整理错误、错位、空的数据。

（2） 将全角符号改为半角符号，包括但不限于“%”、“（”、“）”、“，”。

（3） 去除所有空格、括号和大于等于号。

Step 2 将各种成分及百分比分隔开

（1） 以逗号作为分隔符将形式为“成分占 XX%”的各种成分数据分隔开，并将得到的每个并将拆分得到的数据扩展成每个成分占一行。

（2） 将形式为“成分占 XX%”的成分和百分比的数据拆分为两列。

Step 3 整理各种原料的名称及其百分比数据

（1） 去除成分为空的数据。

（2） 去除无法正确分隔的多余的字符。

基于上述步骤，我们提取得到各种原料的名称及其百分比数据，保存在名为“result4_2.xlsx”的文件中。本文展示部分提取结果，如表 16 所示。

表 16 各种原料的名称及其百分比数据

序号	原料名称	百分比（%）
001	尿素	15.0%
001	高岭土	15.5%
001	硫酸铵	28.2%
001	磷酸一铵	16.3%
001	硫酸钾	25.0%
002	尿素	15.0%
002	高岭土	30.2%
002	氯化铵	28.0%
002	磷酸一铵	12.3%
002	氯化钾	14.5%
003	木薯渣干基	84.9%
003	菌种	0.1%
003	黄豆渣	15.0%
004	尿素	15.0%
004	高岭土	20.0%
004	粉状磷酸一铵	40.0%
004	氯化钾	25.0%