# 基于 KMeans 算法的作物聚类分析研究

## 摘要

随着农业数据的积累和数据挖掘技术的发展，使用聚类分析对作物的生产和市场信息进行有效分类，能够为种植决策和市场推广提供有力支持。本文采用 **KMeans 聚类算法**，对作物的**亩产量、种植成本、销售单价以及预期销售量**等指标进行聚类分析，旨在发现作物种类的潜在分组，进而为农户和农业研究人员提供参考依据。

## 关键词

KMeans聚类、聚类分析

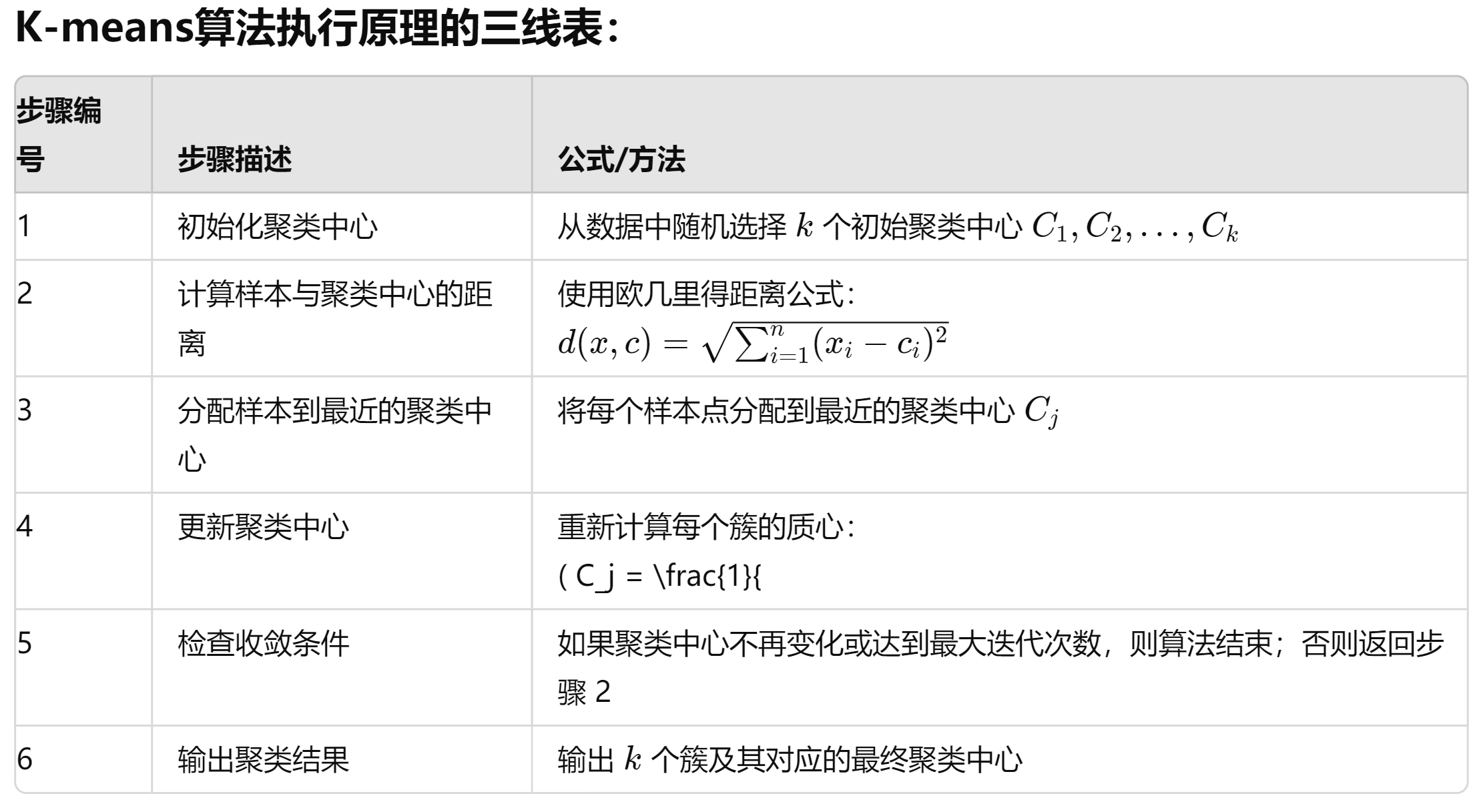
## 1. 引言

近年来，数据挖掘技术广泛应用于农业领域，用于分析农业生产、市场需求和销售趋势。作物的亩产量、种植成本、销售单价等特征数据能够反映其市场表现与种植价值。通过对这些特征的聚类分析，我们可以发现不同作物的潜在分组，从而帮助农户优化种植策略，并为作物的市场推广提供依据。

本文将使用 KMeans 聚类算法对多个作物的生产和市场数据进行分析，以期通过聚类算法发现作物种类之间的特性差异和相似性。

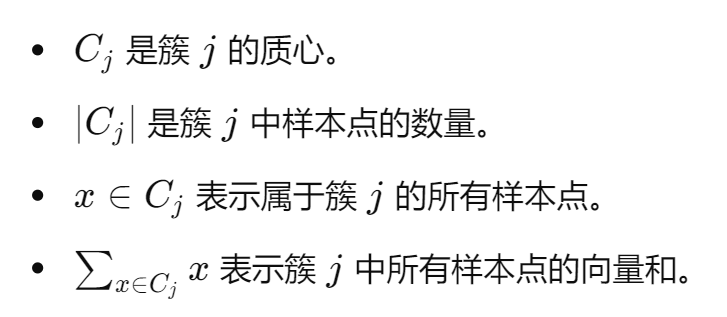
## 2. KMeans 聚类算法原理

KMeans 是一种无监督学习算法，广泛应用于数据聚类。其核心思想是将数据划分为 k 个簇，使得每个簇中的数据点尽可能相似，而不同簇之间的数据点差异尽可能大。该算法的工作流程如下：



（这里我就不转化了，我把里面的格式导出了，麻烦宇佳处理）

这个是每个簇的计算方式：



KMeans 的目标是最小化簇内样本点到簇中心的平方距离总和（惯性），其数学目标函数为：

其中，表示数据点，表示簇 i 的质心，为第 i 个簇中的所有数据点。算法停止的条件是质心的变化幅度小于设定的阈值，或达到预设的迭代次数。

## 3. 数据集与特征选择

本次研究的数据集包含多种作物的以下特征：

- \*\*亩产量/斤\*\*：单位面积的产出。

- \*\*种植成本/(元/亩)\*\*：每亩地的种植成本。

- \*\*销售单价中间值\*\*：作物的市场平均销售价格。

- \*\*预期销售量\*\*：每个作物的预期市场需求量。

这些特征反映了作物的经济效益和市场前景，是进行聚类分析的重要依据。

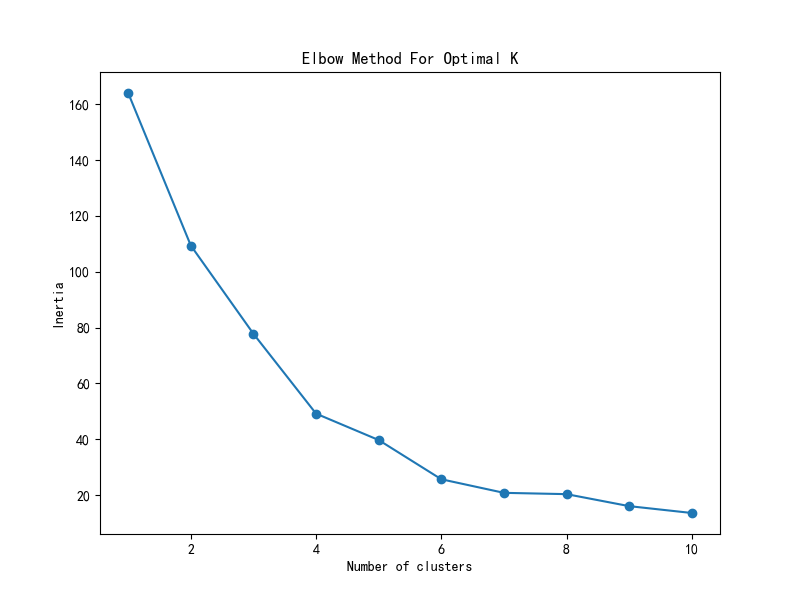
## 4. 聚类过程

### 4.1 数据预处理

在聚类之前，需要对数据进行标准化处理。由于不同特征的取值范围差异较大，例如亩产量和销售单价单位不同，如果不进行标准化，聚类结果可能会受到高值特征的影响。本文采用 \*\***StandardScaler**\*\* 对特征进行标准化处理，使每个特征的数据具有零均值和单位方差。

### 4.2 确定聚类数量

选择适当的聚类数量 k 对聚类效果有很大影响。为确定最佳的 k 值，本文采用了\*\***肘部法（Elbow Method）**\*\*。肘部法通过计算不同 k 值下的惯性（Inertia），绘制出 k 与惯性之间的关系曲线。惯性随着 k 增加而减少，但减少的幅度逐渐减小。当曲线的拐点出现时，即为合适的 k 值。通过分析肘部图，本文选择 k=4 作为最终的聚类数量。



### 4.3 聚类执行

在标准化处理数据后，使用 KMeans 算法进行聚类分析。本文设定聚类数k=4，并对作物数据进行聚类。通过分析聚类结果，可以将作物划分为三个不同的簇，每个簇代表具有相似经济特征的作物组。

## 5. 聚类结果分析

### 5.1 聚类分布

聚类结果将作物分为四类，每一类中的作物在亩产量、种植成本、销售单价和预期销售量等方面表现出相似的特征。通过对每个类的分析，我们可以总结以下结论：

**基于 KMeans 聚类算法的作物数据分析**

根据聚类结果，作物被划分为四类。以下是对每个类的分析和解释，结合数据中的作物特性来理解每个聚类的特点。

**1. Cluster 0: 大田作物与高产作物（例如黄豆、黑豆、玉米）**

在第 0 类中，作物包括黄豆、黑豆、红豆等传统的农作物，以及一些粮食作物，如谷子、黍子、高粱、大麦等。此类作物的特点主要体现在以下几点：

* **高亩产量**：这些作物通常亩产量较高，能够提供稳定的产出。
* **适中种植成本**：这些作物的种植成本处于中等水平，尤其是豆类作物，在较少的投入下可以获得不错的产量。
* **市场需求大**：豆类和谷物类作物属于日常粮食，市场需求广泛，适合大规模种植。

在 Cluster 0 中，作物大多是市场上需求量较大、较为常见的农作物，适合传统农业种植，能够为农户提供稳定的收入来源。

**2. Cluster 1: 高附加值和绿色蔬菜作物（如黄瓜、空心菜、白灵菇）**

第 1 类主要集中于蔬菜和菌类作物，包括黄瓜、空心菜、白灵菇。这类作物的特性如下：

* **高附加值**：白灵菇等高价值菌类属于附加值较高的产品，通常市场单价较高，但亩产量可能相对较低。
* **种植成本较高**：这些作物对种植条件要求较高，特别是白灵菇的种植，通常需要特殊的气候和环境，种植成本较大。
* **市场特定需求**：这类作物主要面向特定的消费群体或城市市场，适合有一定种植技术和经验的农户。

Cluster 1 的作物虽然亩产量和市场需求相对有限，但其较高的市场售价和附加值使得这些作物适合于有条件的高效农业项目。

**3. Cluster 2: 常见粮食作物与大宗作物（如小麦、玉米、大白菜）**

第 2 类包括小麦、玉米、大白菜、白萝卜等作物。这类作物通常是粮食作物或大宗蔬菜，具有以下特点：

* **低种植成本**：这些作物的种植成本相对较低，适合大面积耕种，尤其是玉米和小麦等农作物。
* **市场价格稳定**：此类作物的市场价格波动较小，需求也较为稳定，属于粮食安全保障的基础作物。
* **高亩产量**：玉米、小麦等作物具有较高的亩产量，是农民进行大规模种植的重要选择。

Cluster 2 的作物代表了传统的粮食作物和一些常见的大宗蔬菜，适合用于广泛种植，能够满足大量的市场需求。

**4. Cluster 3: 稀有高端作物（如羊肚菌）**

第 3 类仅包括一种作物，即羊肚菌。羊肚菌是非常高端的作物，具有以下特性：

* **高市场价值**：羊肚菌是一种稀有的菌类作物，市场价格非常高，属于高端消费品。
* **较高种植难度**：羊肚菌对生长环境要求较高，其种植技术较为复杂，且对气候条件敏感，导致其产量相对较低。
* **小规模种植**：由于羊肚菌的高价值和种植的特殊性，它通常不适合大面积耕种，更多用于高端农产品市场。

Cluster 3 代表了稀有高价值的作物，适合精细化管理和高效农业领域，这类作物能够带来较高的收益，但仅适合有足够种植条件和市场渠道的种植者。

**结论**

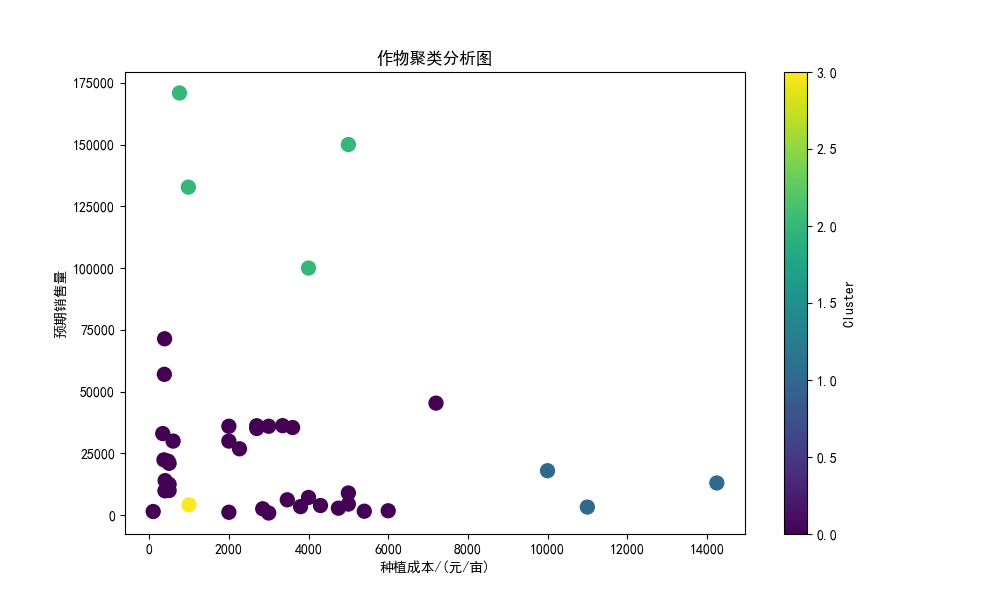
通过 KMeans 聚类分析，我们可以看到作物在不同簇中的特点各异：

* **Cluster 0** 包含了产量较高、成本适中、市场需求大的作物，适合大面积种植，能够提供稳定收入。
* **Cluster 1** 以高附加值蔬菜和菌类为主，适合有条件的高效农业和特定市场需求的种植项目。
* **Cluster 2** 包含了常见的粮食作物和大宗蔬菜，适合大规模种植，具有较低的种植成本和稳定的市场需求。
* **Cluster 3** 羊肚菌作为稀有的高端作物，具有高市场价值，但种植难度较大，适合小规模高效种植。

通过这些分析，农户可以根据自身条件和市场需求选择合适的作物进行种植，充分利用各类作物的特点以实现最佳的经济效益。

### 5.2 可视化结果

为了更直观地展示聚类效果，本文对作物的亩产量和预期销售量进行了可视化分析。聚类结果表明，各类作物在这两个维度上表现出显著的差异，为不同类型的种植决策提供了参考。



## 6. 结论

本文通过 KMeans 聚类算法，对作物的生产和市场特征进行了聚类分析，成功将不同作物划分为三类。聚类结果可以为农户提供种植建议，并为农业决策者提供依据。未来可以结合更多的作物特征，如气候适应性、病虫害抗性等，进一步优化聚类结果。