

《大数据挖掘》实验报告1

题 目 ：电影数据分析与预处理

专 业 数据科学与大数据技术

学 号

姓 名

日 期 2025年4月17日

**目 录**

[第1章 数据预处理 3](#_Toc195891646)

[1.1数据集概况 3](#_Toc195891647)

[1.2缺失值填充 4](#_Toc195891648)

[1.2.1网络中断处理 4](#_Toc195891649)

[1.3增加分类标签属性**profit\_level** 4](#_Toc195891650)

[1.4对 genres字段进行处理 4](#_Toc195891651)

[1.5 在处理后的genres属性基础上，添加新的特征（属性） 5](#_Toc195891652)

[1.6 新增属性“director” 5](#_Toc195891653)

[1.7 新增一个属性：导演执导电影总数(total\_directed) 5](#_Toc195891654)

[1.8 删除director、genres、crew、revenue、title后列出当前属性 5](#_Toc195891655)

[第2章 降维与聚类分析 6](#_Toc195891656)

[2.1特征编码与标准化 6](#_Toc195891657)

[2.1.1特征编码 6](#_Toc195891658)

[2.1.2标准化 6](#_Toc195891659)

[2.2降维与散点图 6](#_Toc195891660)

[2.2.1 PCA降维 6](#_Toc195891661)

[2.2.2 Isomap降维 7](#_Toc195891662)

[2.3 聚类及可视化 8](#_Toc195891663)

[2.3.1 PCA降维 + Kmeans聚类 8](#_Toc195891664)

[2.3.2 Isomap降维 + Kmeans聚类 9](#_Toc195891665)

[2.4 聚类内部**profit\_level属性值的分布** 9](#_Toc195891666)

[2.4.1 PCA+Kmeans聚类内部profit\_level属性值的分布 9](#_Toc195891667)

[第3章 实验结果分析 11](#_Toc195891668)

[3.1 结果分析 11](#_Toc195891669)

[3.2 改进思路 11](#_Toc195891670)

# 第1章 数据预处理

## 1.1数据集概况

数据集movies.csv记录了1916~2016年美国电影信息，数据集中包含23个属性，具体信息如表1所示。

表1 movie.csv属性信息

|  |  |  |
| --- | --- | --- |
| **编号** | **属性** | **描述** |
| **1** | budget | 电影预算 |
| **2** | genres | 电影类型 |
| **3** | homepage | 电影主页 |
| **4** | id | 电影ID |
| **5** | keywords | 电影关键词 |
| **6** | original\_language | 电影原始语言 |
| **7** | original\_title | 电影原始标题 |
| **8** | overview | 电影剧情摘要 |
| **9** | popularity | 电影流行度 |
| **10** | production\_companies | 电影出品公司 |
| **11** | production\_countries | 电影出品国家 |
| **12** | release\_date | 电影档期 |
| **13** | revenue | 电影票房 |
| **14** | runtime | 电影时长 |
| **15** | spoken\_languages | 电影语言 |
| **16** | status | 电影状态 |
| **17** | tagline | 电影标语 |
| **18** | title | 电影标题 |
| **19** | vote\_average | 电影平均评分 |
| **20** | vote\_count | 电影评分次数 |
| **21** | movie\_id | 编号 |
| **22** | cast | 演员阵容 |
| **23** | crew | 全体人员 |

需要删除的无关属性：'homepage', 'original\_title', 'overview', 'spoken\_languages', 'keywords', 'production\_companies', 'status', 'tagline', 'movie\_id', 'vote\_average', 'vote\_count', 'id', 'cast', 'release\_date', 'production\_countries', 'popularity'。它们对应的属性序号为2、3、4、6、7、8、9、10、11、14、15、16、18、19、20、21，因此只需要保留列0、1、5、12、13、17、22。调用pandas库中的iloc方法，通过索引选取DataFrame中的列，完成删除无关属性的任务。

## 1.2缺失值填充

填充budget、revenue、runtime、genres属性的缺失记录或者值为0的记录。这些属性的值需要从IMDB数据库中获取，这可以使用imdb库提供的API实现：对包含缺失值的样本row，首先调用search\_movie，通过电影名称row[‘title’]获得电影信息moive\_info，movie\_info是一个字典类型。然后调用movie\_info.get获取票房box office、电影时长runtimes、电影类型genres对应的值。

缺失值填充任务对应的代码文件是lab1\_fill.py。

### 1.2.1网络中断处理

网络延迟对缺失值填充的效率有较大影响，且网络中断会中断缺失值填充过程。本文采用try-catch捕捉抛出的网络中断异常。当捕捉到网络中断异常后会等待10秒尝试连接到IMDB数据库。这就能避免因为网络中断而导致缺失值填充失败的情况。

## 1.3增加分类标签属性**profit\_level**

profit\_ration的计算公式如式1所示。

|  |  |
| --- | --- |
|  | （1） |

根据profit\_ration的值设置profit\_level。

|  |  |
| --- | --- |
|  | （2） |

在实现过程中，对DataFrame的每一行计算profit\_ration，然后根据profit\_ration的值设置profit\_level，并将profit\_level加入到DataFrame中。

## 1.4对 genres字段进行处理

按照json格式处理genres字段，提取出电影的类型名称，然后将它们作为单独的列加入到DataFrame中。

遍历DataFrame的每一行，调用json.loads方法提取出genres字段中的电影类型名称name，将name加入到列表names中，将genres的属性值更新为names列表。处理后的genres属性值为电影类型名称的列表。

## 1.5 在处理后的genres属性基础上，添加新的特征（属性）

在处理后的genres属性基础上为每种电影类型创建一个新的列，并根据每部电影是否包含该类型来设置相应的值（0或1）。

遍历DataFrame的genres列，将所有的电影类型名称加入到集合all\_genres中。对于all\_genres中的每一个类型，在DataFrame中创建一个新列，其列名就是电影类型genre，然后使用apply方法遍历每一行。如果该电影的genres列表包含这个genre，则设置该列为1，表示该电影属于这个类型。如果该电影的genres列表不包含这个genre，则设置为0。

## 1.6 新增属性“director”

使用json格式解析crew字段，提取job == “Director”的name，这是导演的名字。在DataFrame中添加新属性director。

## 1.7 新增一个属性：导演执导电影总数(total\_directed)

（1）统计每个导演执导电影的总数。遍历DataFrame的director列，对于列表中的每个导演，以导演的名字为键，以导演执导电影的总数为值，更新字典director\_movie\_count。

（2）添加total\_directored属性。遍历DataFrame的每一行，对于每部电影的row[‘director’]列中的每个导演，从director\_movie\_count字典中查找该导演执导的电影次数，计算列表中所有导演执导次数的平均值，计算结果保存在total\_directored属性中，表示该电影所有导演的平均执导次数。

## 1.8 删除director、genres、crew、revenue、title后列出当前属性

调用pandas库的drop方法删除属性director、genres、crew、revenue、title，将结果保存在data/data\_result.csv。

最终的属性为：budget、original\_language、runtime、profit\_level、genres.names、total\_directed

（1）字符型：original\_language

（2）数值型：budget、runtime、profit\_level、total\_directed

（3）布尔型：只有0/1，genres.names

# 第2章 降维与聚类分析

## 2.1特征编码与标准化

### 2.1.1特征编码

对属性original\_language采用目标编码。以下是对各个编码方案的分析。

（1）One-Hot编码：original\_language 属性共有 37 种不同的值，其中 en 类别的样本数量为 4505，占总样本数 4804 的 93.7%。如果使用 One-Hot 编码，数据将变得高维且稀疏，而 PCA 处理的是连续数值型数据，这会使得捕捉数据的结构变得困难。

（2）Label编码、Ordinal编码：由于 original\_language 是无序的，因此不考虑使用 Label 编码或 Ordinal 编码。

（3）频数编码：使用频数编码时可能会引入偏见。en 类别在经过频数编码后会有一个非常高的值，这可能主导降维过程，导致降维后的数据在某些维度上过度集中，从而影响其他类别变量的分布和表现。

（4）目标编码：目标变量profit\_level与original\_language存在一定关系，目标编码能够保留这种关系，并生成数值型数据，有助于减少稀疏矩阵的出现，特别适用于像线性回归和树模型等。这使得目标编码在某些机器学习任务中非常有效。但是由于其依赖目标变量的统计信息，这可能会导致过拟合，特别是在类别较多且数据量较小的情况，因此还需要考虑增加平滑操作。

综上所述，对属性original\_language采用目标编码方法，同时要增加平滑操作。在具体实现上，可以使用CategoryEncoders的TargetEncoder类，在目标编码的同时使用平滑处理。

### 2.1.2标准化

对所有数值型属性进行标准化处理，这是因为PCA降维依赖于数据的方差，标准化能够确保每个特征在同一尺度上，从而避免某些特征因量纲差异而对主成分的计算产生过大影响。

调用sklearn库中的StandardScaler.fit\_transform方法对数值型属性进行标准化处理。

## 2.2降维与散点图

### 2.2.1 PCA降维

对于除profit\_level外的所有属性使用PCA算法降维到2维空间，降维可视化的结果如图1所示。

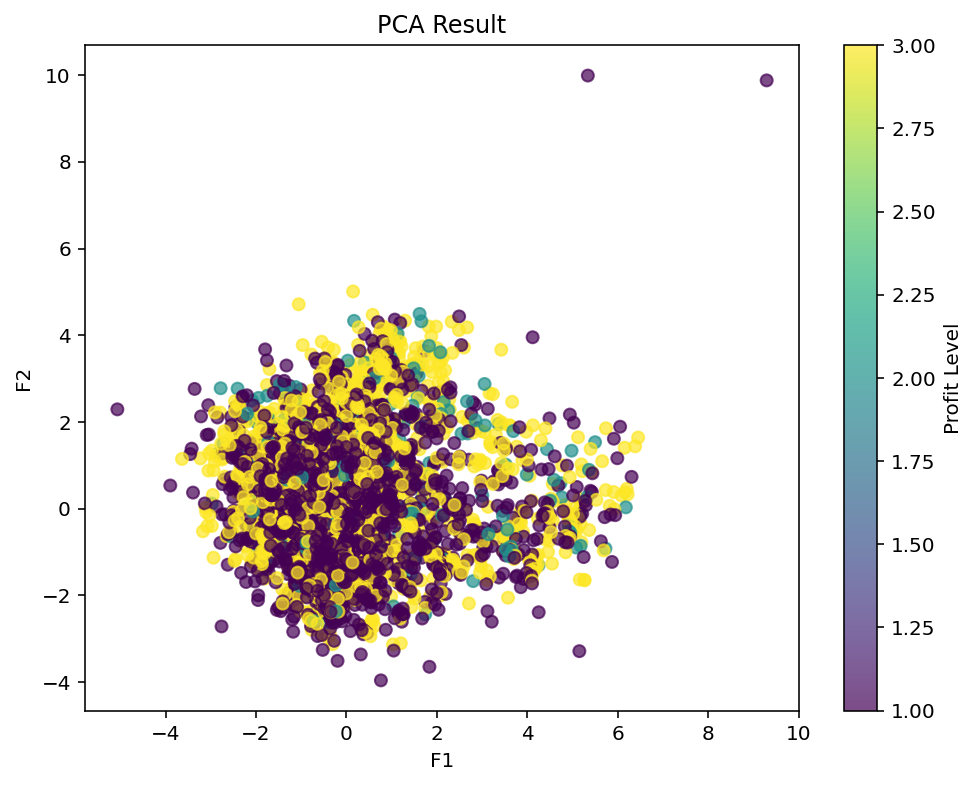


图1 PCA降维散点图

各个特征对主成分的贡献系数保存在feature\_contributions.csv文件中。

对于F1，Family和Animation特征的系数最大，budget和Adventure特征的系数较大，runtime和Crime特征的系数为负，因此推测F1可能表示家庭友好型、轻松娱乐型、较短时间的电影。

对于F2，Action和Science Fiction特征的系数最大，budget和runtime特征的系数为较大，Comedy和Romance特征的系数为负，因此推测F2可能表示更具视觉冲击力和搞预算的影片类型，如动作大片或科幻片。

### 2.2.2 Isomap降维

对于除profit\_level外的所有属性使用Isomap算法降维到2维空间，降维可视化的结果如图2所示。

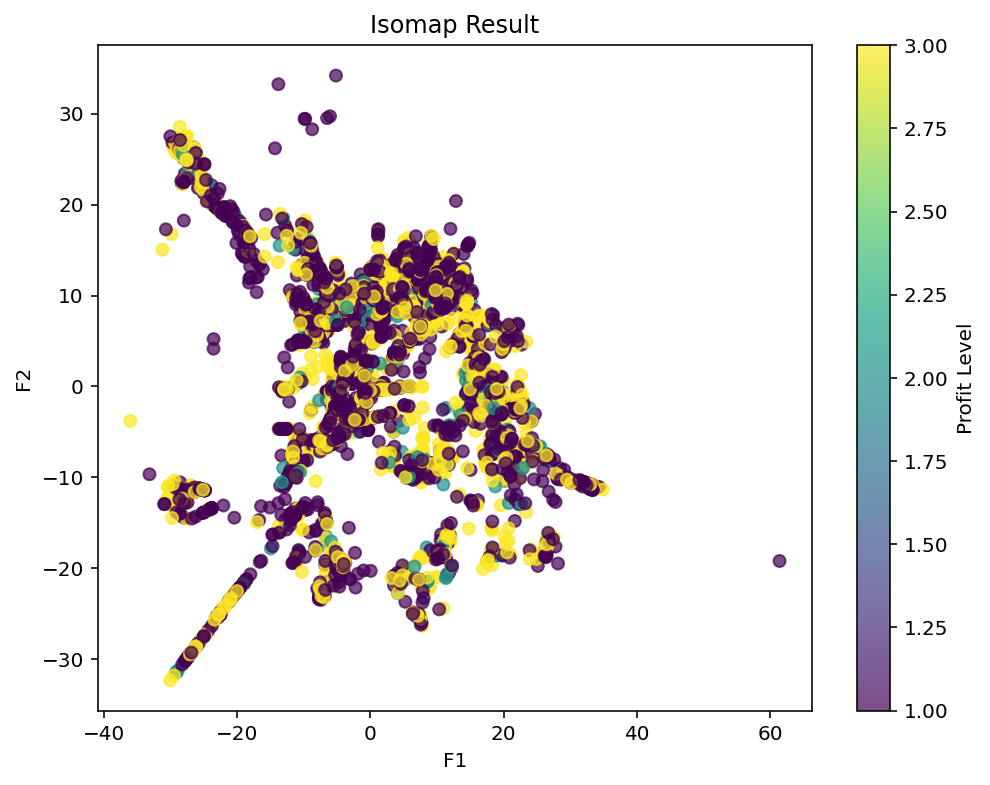


图2 Isomao降维散点图

## 2.3 聚类及可视化

在PCA和Isomap降维的基础上，使用Kmeans聚类算法，并绘制聚类散点图。

### 2.3.1 PCA降维 + Kmeans聚类

设置聚类数量为3，聚类结果如图3所示。

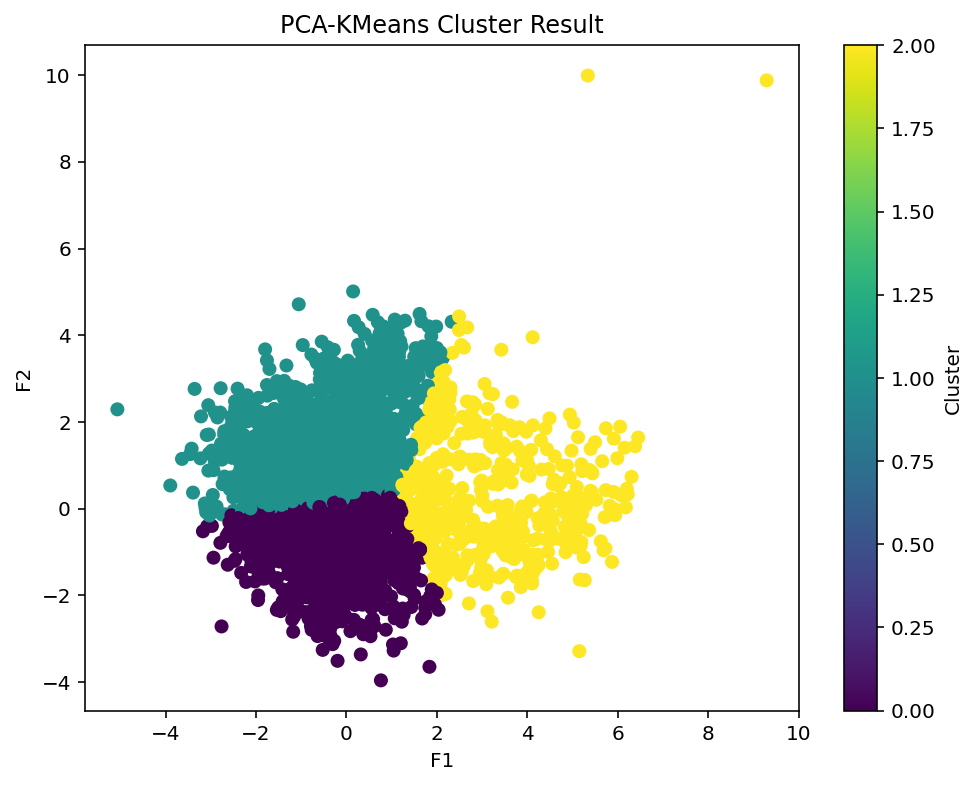


图3 PCA降维 + Kmeans聚类结果

不同颜色表示不同的类别。从图中可以看到，不同簇之间有明显的区分，尤其是黄色簇和紫色簇，它们的簇间距离较大，说明它们在特征空间中有明显差异。紫色簇和绿色簇内部很紧密，可能是因为这些数据点代表了相似类型的电影。

### 2.3.2 Isomap降维 + Kmeans聚类

设置聚类数量为3，聚类结果如图4所示。

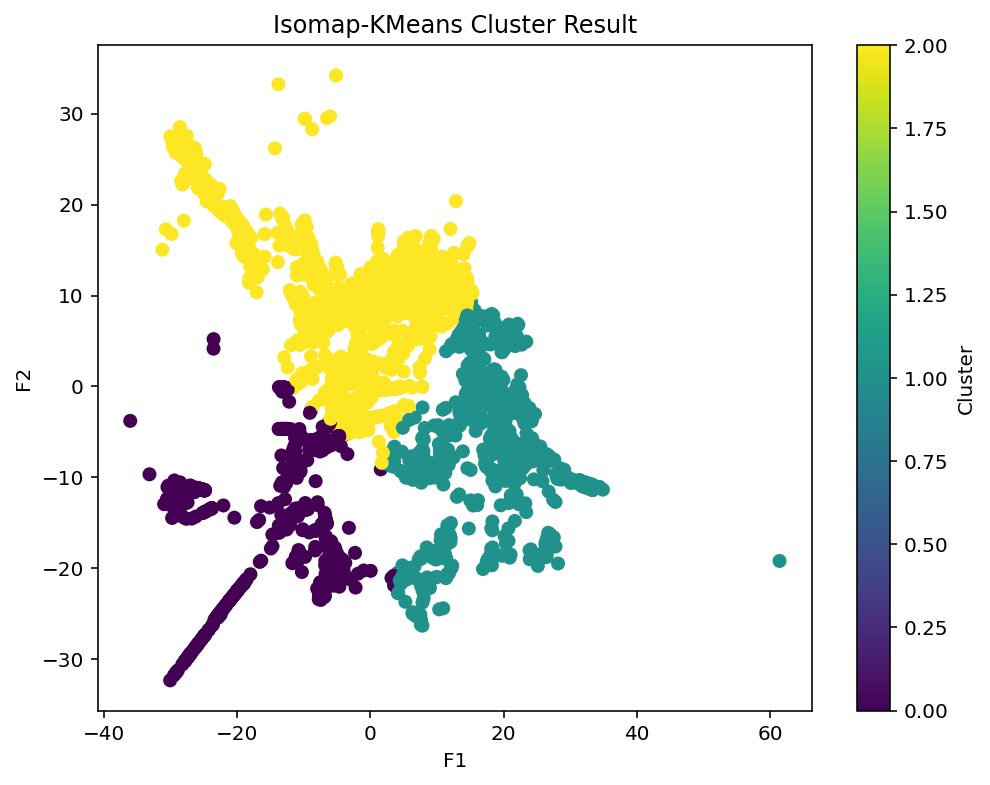
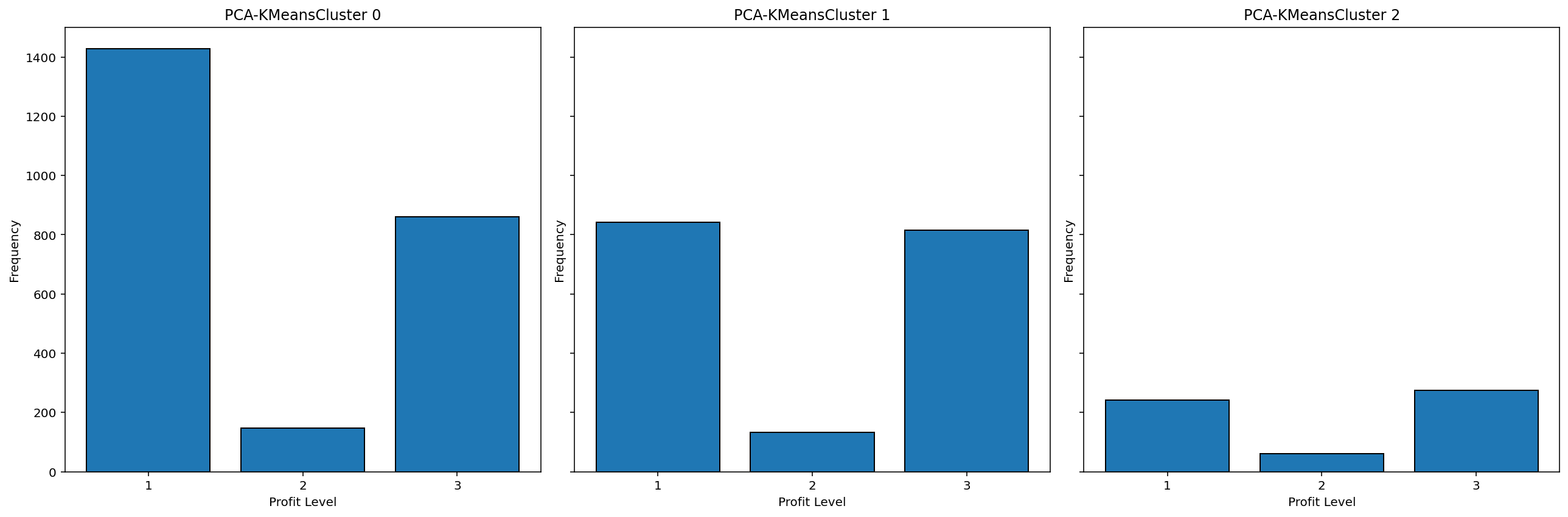


图4 Isomap降维 + Kmeans聚类

Isomap降维后的聚类结果中，簇间边界有重叠，簇内分布较为分散，而PCA降维后的聚类的效果显著更好，这表明数据可能具有线性结构，Isomap的聚类结果的复杂模式可能是由于算法对线性数据的过度拟合或参数设置不当导致的。

## 2.4 聚类内部**profit\_level属性值的分布**

### 2.4.1 PCA+Kmeans聚类内部profit\_level属性值的分布



类内对比，profit\_level=2的样本都是最少的，类2和类3的profit\_level=1和profit\_level=3的样本数量较为接近，类1的profit\_level=1的样本数是profit\_level=3样本数的一倍。类间对比，各个类profit\_level=2的样本数量相近，类3的profit\_level=1和profit\_level=3的样本数是最少的。

由上述分析可推测，类1表示的电影是收益较低的，可能是低利润电影。类2与类3中低收益与高收益的电影混合，可能是因为电影市场中存在一定不确定性或者多样性的类别。比如有些电影虽然有较高投入，但是结果好坏参半。而所有类中profit\_level=2的样本都是最少的，这很符合电影行业的“两级分化”的现象，即电影不是特别赚钱就是特别不赚钱。

# 第3章 实验结果分析

## 3.1 结果分析

（1）由聚类结果可知，Isomap降维后的聚类效果不如PCA降维后的聚类的效果，PCA降维后的聚类效果显著更好，这表明数据可能具有线性结构。但是也有可能是因为数据集中包含噪声，或者维度过低导致的信息损失过大而影响Isomap算法效果。

（2）PCA降维+聚类后的各个簇中，类别内部分析：各个profit\_level值的样本占簇样本总数的比例不同：类1中profit\_level=1的占比最大，表明该类电影可能具有低成本、低市场回报的特征。类2中profit\_level=1和profit\_level=3的占比近似相同，但是占比远比profit\_level=2的占比大，表示该类电影的收益呈现两极分化。类3中profit\_level=1和profit\_level=3的占比与profit\_level=2的相差不大，表明该类电影的收益分布较为均匀。

通过PCA降维 + KMeans聚类能够初步划分出不同收益水平的电影类别，尤其是低收益和高收益电影。然而，由于某些收益水平（如profit\_level=2）的电影在聚类中分布较为分散，表明该方法可能未能完全提取出影响电影收益的核心特征。因此，尽管该方法能够揭示部分结构信息，但可能不足以全面反映电影的最终收益水平。

## 3.2 改进思路

（1）特征构造上：可以增加更多有意义的特征，例如市场营销支出，演员阵容，上映时期等。这些特征或许能够更好地区分出收益不同的电影类别。

（2）使用更高维度的主成分：目前只使用了前两个主成分，但是收益水平可能存在更多潜在的影响因素，可以尝试使用更多主成分进行降维。