数据采样方法实践实验报告

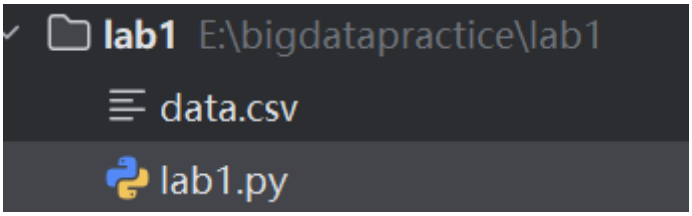
实验目的

本次实验主要学习如何使用Python对数据进行采样和过滤。从一个包含网络节点信息的数据集开始，通过删除空行、过滤特定条件的数据，然后使用三种不同的采样方法（加权采样、随机抽样和分层抽样）来获取样本数据。

实验步骤

1. 准备阶段

首先导入需要的工具库，下载数据集并放置在同一文件夹中。



2. 清理数据

删除所有包含空值的行，确保数据的完整性。

```
# 2. 删除空行
print("2. 删除所有包含空值的行...")
primitive_data_1 = primitive_data.dropna(how='any')
print("删除空行后的数据形状：", primitive_data_1.shape)
print("前5行和后5行如下：")
print("前5行：")
print(primitive_data_1.head())
print("\n后5行：")
print(primitive_data_1.tail())
print()
```

3. 数据过滤

筛选出流量不为0且来自"一般节点"的数据。

```
print("3. 过滤数据：traffic != 0 且 from_level == '一般节点'...")
data_after_filter_1 = primitive_data_1.loc[primitive_data_1["traffic"] != 0]
data_after_filter_2 = data_after_filter_1.loc[data_after_filter_1["from_level"] == '一般节点']
print("过滤后的数据形状：", data_after_filter_2.shape)
print("前5行和后5行如下：")
print("前5行：")
print(data_after_filter_2.head())
print("\n后5行：")
print(data_after_filter_2.tail())
print()
```

4. 三种采样方法

- **加权采样**：给"网络核心"节点5倍权重，"一般节点"1倍权重
- **随机抽样**：完全随机选择50个样本
- **分层抽样**：按节点类型分层，抽取17个"一般节点"和33个"网络核心"

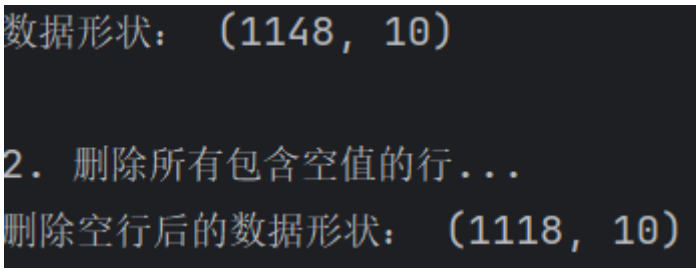
```
# 4. 加权采样：to_level为"一般节点"和"网络核心"的权重比为1:5
print("4. 进行加权采样（权重比：一般节点:网络核心 = 1:5）...")
data_before_sample = data_after_filter_2.copy()
data_before_sample['weight'] = data_before_sample['to_level'].apply(
    lambda x: 1 if x == '一般节点' else 5
)
weight_sample_finish = data_before_sample.sample(n=50, weights='weight',
random_state=42)
weight_sample_finish = weight_sample_finish.drop(columns=['weight'])  # 移除权重列
print("加权采样后的50个样本（前5行和后5行）：")
print("前5行：")
print(weight_sample_finish.head())
print("\n后5行：")
print(weight_sample_finish.tail())
print("采样数据形状：", weight_sample_finish.shape)
print()

# 5. 随机抽样
print("5. 进行随机抽样...")
random_sample_finish = data_before_sample.sample(n=50, random_state=42)
random_sample_finish = random_sample_finish.drop(columns=['weight'])  # 移除权重列
print("随机抽样后的50个样本（前5行和后5行）：")
print("前5行：")
print(random_sample_finish.head())
print("\n后5行：")
print(random_sample_finish.tail())
print("采样数据形状：", random_sample_finish.shape)
print()

# 6. 分层抽样：按 to_level 分层，一般节点抽17个，网络核心抽33个
print("6. 进行分层抽样：一般节点17个，网络核心33个...")
ybjd = data_before_sample[data_before_sample['to_level'] == '一般节点']
wlhx = data_before_sample[data_before_sample['to_level'] == '网络核心']
ybjd_sample = ybjd.sample(n=17, random_state=42)
wlhx_sample = wlhx.sample(n=33, random_state=42)
after_sample = pd.concat([ybjd_sample, wlhx_sample])
after_sample = after_sample.drop(columns=['weight'])  # 移除权重列
print("分层抽样后的50个样本（前5行和后5行）：")
print("前5行：")
print(after_sample.head())
```

```python
print("\n后5行：")
print(after_sample.tail())
print("采样数据形状：", after_sample.shape)
print("各类别数量：")
print(after_sample['to_level'].value_counts())
```

结果展示




```
4. 进行加权采样（权重比：一般节点:网络核心 = 1:5）...
加权采样后的50个样本（前5行和后5行）：
前5行：
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
292         63          6        通辽  ...     网络核心  51392218854  1.000000e+11
1005     36036        499        长春  ...     网络核心  49116324777  1.000000e+11
534         47        258        通辽  ...     网络核心  50060087433  1.000000e+11
411        591         19        绥化  ...     网络核心  50171685281  1.000000e+11
80         180        200      呼和浩特  ...     网络核心  51884294458  1.000000e+11

[5 rows x 10 columns]

后5行：
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
365        180        260      呼和浩特  ...     网络核心  48917626581  1.000000e+11
378        474        472       哈尔滨  ...     网络核心  50470657254  1.000000e+11
1053      2473        769        吉林  ...     网络核心  51047474759  1.000000e+11
558         96         99      呼和浩特  ...     网络核心  49166600948  1.000000e+11
942      36036         52        长春  ...     网络核心  49916177327  1.000000e+11
```

```
5. 进行随机抽样...
随机抽样后的50个样本（前5行和后5行）：
前5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
296        63         58        通辽  ...     网络核心  49092144382  1.000000e+11
79        180        192       呼和浩特  ...      一般节点  49504348509  1.000000e+11
830     36036         54        长春  ...     一般节点  49794381448  1.000000e+11
113       474        678       哈尔滨  ...      网络核心  49044545927  1.000000e+11
997     36036         52        长春  ...     一般节点  50822505842  1.000000e+11

[5 rows x 10 columns]

后5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
545        63         58        通辽  ...     网络核心  51132553467  1.000000e+11
994        63          6        通辽  ...     网络核心  50680536460  1.000000e+11
942     36036         52        长春  ...     网络核心  49916177327  1.000000e+11
382       474        614       哈尔滨  ...      网络核心  51241236810  1.000000e+11
962      4448        127        无锡  ...     一般节点  50961073987  1.000000e+11
```

```
5. 进行分层抽样：一般节点17个，网络核心33个...
分层抽样后的50个样本（前5行和后5行）：
前5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
622       180         20       呼和浩特  ...     一般节点  49636788433  1.000000e+11
913      2473        799        吉林  ...     一般节点  50993016382  1.000000e+11
966     36539       1146        杭州  ...     一般节点  49520418698  1.000000e+11
530        47        249        通辽  ...     一般节点  49803820036  1.000000e+11
498        47        314        通辽  ...     一般节点  50043006782  1.000000e+11

[5 rows x 10 columns]

后5行：
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
1107     36036         52        长春  ...     网络核心  49345226162  1.000000e+11
1093       591        586        绥化  ...     网络核心  47929885030  1.000000e+11
37          96        108       呼和浩特  ...      网络核心  48210462086  1.000000e+11
152        591        638        绥化  ...     网络核心  49178187887  1.000000e+11
60          96        399       呼和浩特  ...      网络核心  50243694923  1.000000e+11

[5 rows x 10 columns]
采样数据形状： (50, 10)
各类别数量：
to_level
网络核心    33
一般节点    17
```

实验结论

通过本次实验，我成功掌握了数据处理的三个重要步骤：

1. **数据清理**：学会了如何识别和删除空行，确保数据的质量
2. **数据过滤**：能够根据特定条件（如流量不为0、节点类型）筛选出需要的数据

3. **数据采样**：掌握了三种不同的采样方法：
    - 加权采样：给重要数据更高的被选中机会
    - 随机抽样：每个数据被选中的机会均等
    - 分层抽样：按类别比例抽取样本，保证各类别都有代表

实验结果显示，三种采样方法都成功抽取了50个样本，但样本组成有所不同。加权采样中网络核心节点更多，分层抽样严格按照设定比例，随机抽样则完全随机。这些方法在不同场景下各有优势，可以根据实际需求选择合适的采样方式。