# 山东大学计算机科学与技术学院

# 大数据分析与实践课程实验报告

| 学号：202320130203 | 姓名：李姿含 | 班级：数据 23 |
|---|---|---|
| 实验题目：数据采样方法实践 | | |
| 实验学时：2 | 实验日期：20250919 | |

实验目标：

本实验旨在实现多种数据采样和过滤方法，深入理解数据预处理过程中的数据清洗，数据过滤以及不同采样方式的原理与应用。

作品描述（实验背景、数据集来源、描述思路、实验背景）：

1、实验背景：

在大数据分析中，原始数据常存在空值、无效记录等问题，且直接分析全量数据易导致资源消耗大、效率低。数据清洗（如删除空行、条件过滤）是保障数据质量的关键步骤，而合理的采样方法能在减少数据量的同时，保留数据核心特征，为后续分析（如网络流量规律挖掘）提供高效支持。本实验围绕网络流量数据，开展数据清洗与多采样方法实践，为大数据预处理流程建立标准化认知。

2、数据集来源：

http://storage.ameshol land.xyz/data.csv

数据集包含 10 个字段，分别记录源设备（from_dev）、源端口（from_port）、源城市（from_city）、源节点级别（from_level）、目标设备（to_dev）、目标端口（to_port）、目标城市（to_city）、目标节点级别（to_level）、流量（traffic）、带宽（bandwidth）信息。

3、描述思路：

①数据预处理：导入 pandas，numpy 库，使用 pd.read_csv 读取数据集，查看原始数据结构；调用 dropna(how='any')删除含任意空值的行，解决空行问题；通过两次 loc[]条件过滤（traffic!=0 和 from_level=='一般节点'），筛选出有效分析数据。

```
import pandas as pd
fro  C:\Users\86134\PycharmProjects\knn\.venv\big_data_fenxi_lab1.py
import numpy as np
primitive_data=pd.read_csv("D:\\大三上学期\\大数据分析实践\\data.csv",encoding='gbk')
#print(primitive_data)

primitive_data_1=primitive_data.dropna(how='any')
#print(primitive_data_1)

data_before_filter=primitive_data_1
data_after_filter_1=data_before_filter.loc[data_before_filter["traffic"]!=0]
data_after_filter_2=data_after_filter_1.loc[data_after_filter_1["from_level"]=='一般节点']
#print(data_after_filter_2)
```

②采样方法：

加权采样：根据目标节点级别（to_level）分配权重，一般节点 权重为 1，网络核心 权重为 5，通过 sample(n=50,weight='weight)实现。

```python
data_before_sample=data_after_filter_2
columns=data_before_sample.columns
weight_sample=data_before_sample.copy()
weight_sample['weight']=0
for i in weight_sample.index:
    if weight_sample.at[i,'to_level']=='一般节点':
        weight=1
    else:
        weight=5
    weight_sample.at[i,'weight']=weight
weight_sample_finish=weight_sample.sample(n=50,weights='weight')
weight_sample_finish=weight_sample_finish[columns]
#print(weight_sample_finish)
```

随机采样：直接使用 sample(n=50)从过滤后的数据中随机抽取，保证每个样本被选中概率均等。

```python
#随机抽样
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
#print(random_sample_finish)
```

分层采样：按 to_level 分为一般节点（ybjd）和网络核心（wlhx）两层，分别抽取 17 和 33 个样本，再用 pd.concat()合并，确保样本分层分布与总体一致。

```python
#分层抽样
ybjd=data_before_sample.loc[data_before_sample['to_level']=='一般节点']
wlhx=data_before_sample.loc[data_before_sample['to_level']=='网络核心']
after_sample=pd.concat([ybjd.sample(17),wlhx.sample(33)])
print(after_sample)
```

结果图片：

1、原始数据读入：

```
big_data_fenxi_lab1
D:\anaconda3\python.exe C:\Users\86134\PycharmProjects\knn\.venv\big_data_fenxi_lab1.py
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
0           47         71        通辽  ...     网络核心  49636052613  1.000000e+11
1           47         74        通辽  ...     网络核心  50056871412  1.000000e+11
2           47        240        通辽  ...     网络核心  49453581081  1.000000e+11
3           47        241        通辽  ...     网络核心  49733361585  1.000000e+11
4           47        242        通辽  ...     一般节点  50492573662  1.000000e+11
...        ...        ...       ...  ...      ...          ...           ...
1113      1129        546        上海  ...     网络核心  48731433404  1.000000e+11
1114      1129        514        上海  ...     一般节点  50060666120  1.000000e+11
1115     36036        499        长春  ...     网络核心  50545082113  1.000000e+11
1116     36422        346        天津  ...     网络核心  50628787089  1.000000e+11
1117      2701        619        大连  ...     网络核心  48753971761  1.000000e+11

[1118 rows x 10 columns]
```

2、删除多余空行：

big_data_fenxi_lab1

```
D:\anaconda3\python.exe C:\Users\86134\PycharmProjects\knn\.venv\big_data_fenxi_lab1.py
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
0           47         71        通辽  ...     网络核心  49636052613  1.000000e+11
1           47         74        通辽  ...     网络核心  50056871412  1.000000e+11
2           47        240        通辽  ...     网络核心  49453581081  1.000000e+11
3           47        241        通辽  ...     网络核心  49733361585  1.000000e+11
4           47        242        通辽  ...     一般节点  50492573662  1.000000e+11
...        ...        ...       ...  ...      ...          ...           ...
1113      1129        546        上海  ...     网络核心  48731433404  1.000000e+11
1114      1129        514        上海  ...     一般节点  50060666120  1.000000e+11
1115     36036        499        长春  ...     网络核心  50545082113  1.000000e+11
1116     36422        346        天津  ...     网络核心  50628787089  1.000000e+11
1117      2701        619        大连  ...     网络核心  48753971761  1.000000e+11
```

3、进行过滤：

```
D:\anaconda3\python.exe C:\Users\86134\PycharmProjects\knn\.venv\big_data_fenxi_lab1.py
      from_dev  from_port from_city  ... to_level      traffic     bandwidth
0           47         71        通辽  ...     网络核心  49636052613  1.000000e+11
1           47         74        通辽  ...     网络核心  50056871412  1.000000e+11
2           47        240        通辽  ...     网络核心  49453581081  1.000000e+11
3           47        241        通辽  ...     网络核心  49733361585  1.000000e+11
4           47        242        通辽  ...     一般节点  50492573662  1.000000e+11
...        ...        ...       ...  ...      ...          ...           ...
1097      2473       1460        吉林  ...     一般节点  48409925693  1.000000e+11
1103     36036         18        长春  ...     网络核心  48663350759  1.000000e+11
1104        63          6        通辽  ...     一般节点  50355678076  1.000000e+11
1107     36036         52        长春  ...     网络核心  49345226162  1.000000e+11
1115     36036        499        长春  ...     网络核心  50545082113  1.000000e+11

[550 rows x 10 columns]
```

4、加权采样：

```
     from_dev  from_port from_city  ... to_level      traffic     bandwidth
363       180        254      呼和浩特  ...     网络核心  50252917820  1.000000e+11
52         96        157      呼和浩特  ...     网络核心  50096366926  1.000000e+11
98        474        417       哈尔滨  ...     网络核心  51874083489  1.000000e+11
300        63         70        通辽  ...     网络核心  50635697563  1.000000e+11
75        180         84      呼和浩特  ...     网络核心  49100967003  1.000000e+11
634      2473        769        吉林  ...     网络核心  49319842054  1.000000e+11
292        63          6        通辽  ...     网络核心  51392218854  1.000000e+11
18         63         10        通辽  ...     网络核心  52195591947  1.000000e+11
804       180        264      呼和浩特  ...     一般节点  49012460413  1.000000e+11
422       591        526        绥化  ...     网络核心  48492868383  1.000000e+11
942     36036         52        长春  ...     网络核心  49916177327  1.000000e+11
168       787         52        玉溪  ...     网络核心  50468642387  1.000000e+11
393       474       1238       哈尔滨  ...     网络核心  49693039378  1.000000e+11
100       474        422       哈尔滨  ...     一般节点  48084671443  1.000000e+11
144       591         98        绥化  ...     网络核心  50256295026  1.000000e+11
32         63        282        通辽  ...     网络核心  49455678350  1.000000e+11
546        63         60        通辽  ...     一般节点  47970715088  1.000000e+11
127       474       1399       哈尔滨  ...     一般节点  50372436809  1.000000e+11
490        47        243        通辽  ...     网络核心  50075073640  1.000000e+11
336        96        407      呼和浩特  ...     网络核心  50219393940  1.000000e+11
```

5、随机抽样：

| | from_dev | from_port | from_city | ... | to_level | traffic | bandwidth |
|---|---|---|---|---|---|---|---|
| 415 | 591 | 56 | 绥化 | ... | 网络核心 | 47741796615 | 1.000000e+11 |
| 535 | 47 | 259 | 通辽 | ... | 网络核心 | 51012708275 | 1.000000e+11 |
| 532 | 47 | 251 | 通辽 | ... | 网络核心 | 51158383342 | 1.000000e+11 |
| 440 | 591 | 1290 | 绥化 | ... | 网络核心 | 50090927530 | 1.000000e+11 |
| 68 | 180 | 30 | 呼和浩特 | ... | 网络核心 | 49596659754 | 1.000000e+11 |
| 116 | 474 | 1227 | 哈尔滨 | ... | 网络核心 | 48505909225 | 1.000000e+11 |
| 354 | 180 | 192 | 呼和浩特 | ... | 一般节点 | 51828297117 | 1.000000e+11 |
| 329 | 96 | 159 | 呼和浩特 | ... | 一般节点 | 51159730271 | 1.000000e+11 |
| 375 | 474 | 422 | 哈尔滨 | ... | 网络核心 | 50424883915 | 1.000000e+11 |
| 443 | 787 | 52 | 玉溪 | ... | 网络核心 | 49322809158 | 1.000000e+11 |
| 620 | 180 | 264 | 呼和浩特 | ... | 网络核心 | 50207994896 | 1.000000e+11 |
| 74 | 180 | 52 | 呼和浩特 | ... | 一般节点 | 49155371449 | 1.000000e+11 |
| 87 | 180 | 252 | 呼和浩特 | ... | 一般节点 | 49137975001 | 1.000000e+11 |
| 306 | 63 | 278 | 通辽 | ... | 网络核心 | 51091741717 | 1.000000e+11 |
| 993 | 36036 | 18 | 长春 | ... | 网络核心 | 49826827167 | 1.000000e+11 |
| 54 | 96 | 159 | 呼和浩特 | ... | 网络核心 | 51625089370 | 1.000000e+11 |
| 549 | 63 | 70 | 通辽 | ... | 一般节点 | 49551919218 | 1.000000e+11 |
| 609 | 96 | 391 | 呼和浩特 | ... | 网络核心 | 48978587445 | 1.000000e+11 |
| 487 | 47 | 240 | 通辽 | ... | 网络核心 | 49873119534 | 1.000000e+11 |
| 168 | 787 | 52 | 玉溪 | ... | 网络核心 | 50468642387 | 1.000000e+11 |

6、分层抽样：

| | from_dev | from_port | from_city | ... | to_level | traffic | bandwidth |
|---|---|---|---|---|---|---|---|
| 847 | 47 | 252 | 通辽 | ... | 一般节点 | 51065218921 | 1.000000e+11 |
| 151 | 591 | 586 | 绥化 | ... | 一般节点 | 49061517661 | 1.000000e+11 |
| 828 | 47 | 314 | 通辽 | ... | 一般节点 | 50910415109 | 1.000000e+11 |
| 924 | 2473 | 1043 | 吉林 | ... | 一般节点 | 50311375989 | 1.000000e+11 |
| 674 | 591 | 586 | 绥化 | ... | 一般节点 | 50565152517 | 1.000000e+11 |
| 498 | 47 | 314 | 通辽 | ... | 一般节点 | 50043006782 | 1.000000e+11 |
| 447 | 787 | 63 | 玉溪 | ... | 一般节点 | 49557001334 | 1.000000e+11 |
| 13 | 47 | 314 | 通辽 | ... | 一般节点 | 50161220081 | 1.000000e+11 |
| 451 | 787 | 324 | 玉溪 | ... | 一般节点 | 49843503409 | 1.000000e+11 |
| 423 | 591 | 558 | 绥化 | ... | 一般节点 | 48364223310 | 1.000000e+11 |
| 157 | 591 | 1106 | 绥化 | ... | 一般节点 | 50954337724 | 1.000000e+11 |
| 96 | 474 | 360 | 哈尔滨 | ... | 一般节点 | 51819320173 | 1.000000e+11 |
| 548 | 63 | 66 | 通辽 | ... | 一般节点 | 48141568100 | 1.000000e+11 |
| 780 | 96 | 391 | 呼和浩特 | ... | 一般节点 | 50103206178 | 1.000000e+11 |
| 556 | 63 | 282 | 通辽 | ... | 一般节点 | 49489299594 | 1.000000e+11 |
| 376 | 474 | 460 | 哈尔滨 | ... | 一般节点 | 48394911971 | 1.000000e+11 |
| 404 | 474 | 1410 | 哈尔滨 | ... | 一般节点 | 49488245045 | 1.000000e+11 |
| 344 | 180 | 34 | 呼和浩特 | ... | 网络核心 | 50352242512 | 1.000000e+11 |
| 131 | 474 | 1473 | 哈尔滨 | ... | 网络核心 | 53304989080 | 1.000000e+11 |
| 20 | 63 | 54 | 通辽 | ... | 网络核心 | 49256234165 | 1.000000e+11 |
| 363 | 180 | 254 | 呼和浩特 | ... | 网络核心 | 50252917820 | 1.000000e+11 |
| 57 | 96 | 379 | 呼和浩特 | ... | 网络核心 | 49400869697 | 1.000000e+11 |
| 372 | 474 | 416 | 哈尔滨 | ... | 网络核心 | 49544939922 | 1.000000e+11 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 372 | 474 | 416 | 哈尔滨 | ... | 网络核心 | 49544939922 | 1.000000e+11 |
| 320 | 96 | 134 | 呼和浩特 | ... | 网络核心 | 48498103572 | 1.000000e+11 |
| 442 | 787 | 51 | 玉溪 | ... | 网络核心 | 50594027588 | 1.000000e+11 |
| 492 | 47 | 250 | 通辽 | ... | 网络核心 | 49014089485 | 1.000000e+11 |
| 295 | 63 | 54 | 通辽 | ... | 网络核心 | 49566827928 | 1.000000e+11 |
| 171 | 787 | 61 | 玉溪 | ... | 网络核心 | 50063136706 | 1.000000e+11 |
| 889 | 63 | 12 | 通辽 | ... | 网络核心 | 49823274555 | 1.000000e+11 |
| 1107 | 36036 | 52 | 长春 | ... | 网络核心 | 49345226162 | 1.000000e+11 |
| 1059 | 47 | 252 | 通辽 | ... | 网络核心 | 50358481161 | 1.000000e+11 |
| 326 | 96 | 156 | 呼和浩特 | ... | 网络核心 | 50272713910 | 1.000000e+11 |
| 322 | 96 | 136 | 呼和浩特 | ... | 网络核心 | 50541979348 | 1.000000e+11 |
| 341 | 180 | 26 | 呼和浩特 | ... | 网络核心 | 48797633450 | 1.000000e+11 |
| 90 | 180 | 260 | 呼和浩特 | ... | 网络核心 | 48006842653 | 1.000000e+11 |
| 587 | 96 | 141 | 呼和浩特 | ... | 网络核心 | 47941844052 | 1.000000e+11 |
| 561 | 96 | 108 | 呼和浩特 | ... | 网络核心 | 49739592973 | 1.000000e+11 |
| 77 | 180 | 98 | 呼和浩特 | ... | 网络核心 | 50330801190 | 1.000000e+11 |
| 163 | 591 | 1284 | 绥化 | ... | 网络核心 | 50187660151 | 1.000000e+11 |
| 317 | 96 | 123 | 呼和浩特 | ... | 网络核心 | 50500915133 | 1.000000e+11 |
| 1041 | 180 | 20 | 呼和浩特 | ... | 网络核心 | 50353235399 | 1.000000e+11 |
| 373 | 474 | 417 | 哈尔滨 | ... | 网络核心 | 50339382092 | 1.000000e+11 |
| 105 | 474 | 474 | 哈尔滨 | ... | 网络核心 | 49764093255 | 1.000000e+11 |
| 359 | 180 | 214 | 呼和浩特 | ... | 网络核心 | 52204574667 | 1.000000e+11 |
| 563 | 96 | 114 | 呼和浩特 | ... | 网络核心 | 51329552752 | 1.000000e+11 |
| 85 | 180 | 218 | 呼和浩特 | ... | 网络核心 | 50106572586 | 1.000000e+11 |
| 1093 | 591 | 586 | 绥化 | ... | 网络核心 | 47929885030 | 1.000000e+11 |
| 938 | 36036 | 18 | 长春 | ... | 网络核心 | 47728327758 | 1.000000e+11 |
| 939 | 63 | 6 | 通辽 | ... | 网络核心 | 49208374165 | 1.000000e+11 |
| 102 | 474 | 467 | 哈尔滨 | ... | 网络核心 | 49987703744 | 1.000000e+11 |

[50 rows x 10 columns]