Московский Государственный Технический Университет имени Н.Э.Баумана

Факультет Информатика и системы управления

Кафедра ИУ-5

«Системы обработки информации и управления»

Отчёт по Рубежному Контролю No 2

Методы обработки данных

Выполнили стуленты группы ИУ5И 21М

Фань Лицзе

Москва 2024г.

Задача №23.

Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе правила трех сигм.

Используйте NumPy и pandas в Python, чтобы создать простой набор данных со случайно сгенерированными данными и намеренно добавить некоторые выбросы. Таким образом, мы можем гарантировать, что набор данных содержит некоторые очевидные выбросы для демонстрационных целей.

```
pip install numpy pandas
```

Далее давайте создадим набор данных с выбросами и применим правило трех сигм, чтобы удалить эти выбросы:

```
import numpy as np
import pandas as pd
# 设置随机种子以获得可重现结果
np.random.seed(0)
# 创建一个含有 100 个数据点的 DataFrame，数据正态分布，均值为 50，标准差为 10
data = pd.DataFrame({'Values': np.random.normal(50, 10, 100)})
# 故意添加异常值
outliers = pd.DataFrame({'Values': [150, -50, 200, -100]})
data = pd.concat([data, outliers], ignore_index=True)
```

```
print("原始数据集大小: ", data.shape)


# 计算均值和标准差
mean = data['Values'].mean()
std = data['Values'].std()


# 三西格玛规则的边界
lower_bound = mean - 3*std
upper_bound = mean + 3*std


# 筛选数据以去除异常值
filtered_data = data[(data['Values'] >= lower_bound) & (data['Values'] <=
upper_bound)]


print("过滤掉异常值后的数据集大小: ", filtered_data.shape)


# 展示被认定为异常的数据点
print("被认定为异常的数据点: ")
print(data[(data['Values'] < lower_bound) | (data['Values'] > upper_bound)])
```

Сначала был сгенерирован нормально распределенный набор данных со средним значением 50 и стандартным отклонением 10, а также было добавлено несколько очевидных выбросов (150, -50, 200, -100). Затем мы рассчитали среднее и стандартное отклонение набора данных и применили правило трех сигм, чтобы определить границы выбросов. Наконец, мы отфильтровываем выбросы из набора данных и показываем размер набора данных до и после фильтрации, а также точки данных, идентифицированные как аномалии.

```
原始数据集大小:  (104, 1)
过滤掉异常值后的数据集大小:  (100, 1)
被认定为异常的数据点:
     Values
100   150.0
101   -50.0
102   200.0
103  -100.0
```