机器学习策划方案

小组成员：李克西（1班）
杨巧杰（2班）
钱伟（2班）
谢尚广（2班）
刘兴（2班）
夏杰（3班）

目标

- 居民用户：
  - 分析水用量的规律（规律用水用户、间断用水用户、偶尔用水用户、不用水用户）；
  - 对用户进行分类；
  - 找出其中不合理用水用户（规律用水->不规律用水、 用水量偏大或偏小）；
  - 可根据规律预测用户未来一段时间的用水量。

目标

- 工厂用户：
  - 分析出大用户的水用量规律；
  - 对用户进行分类；
  - 当规律不匹配时进行提醒；
  - 根据规律预测用户未来一段时间的用水量。

用户用水量情况分析

分析水用量的规律

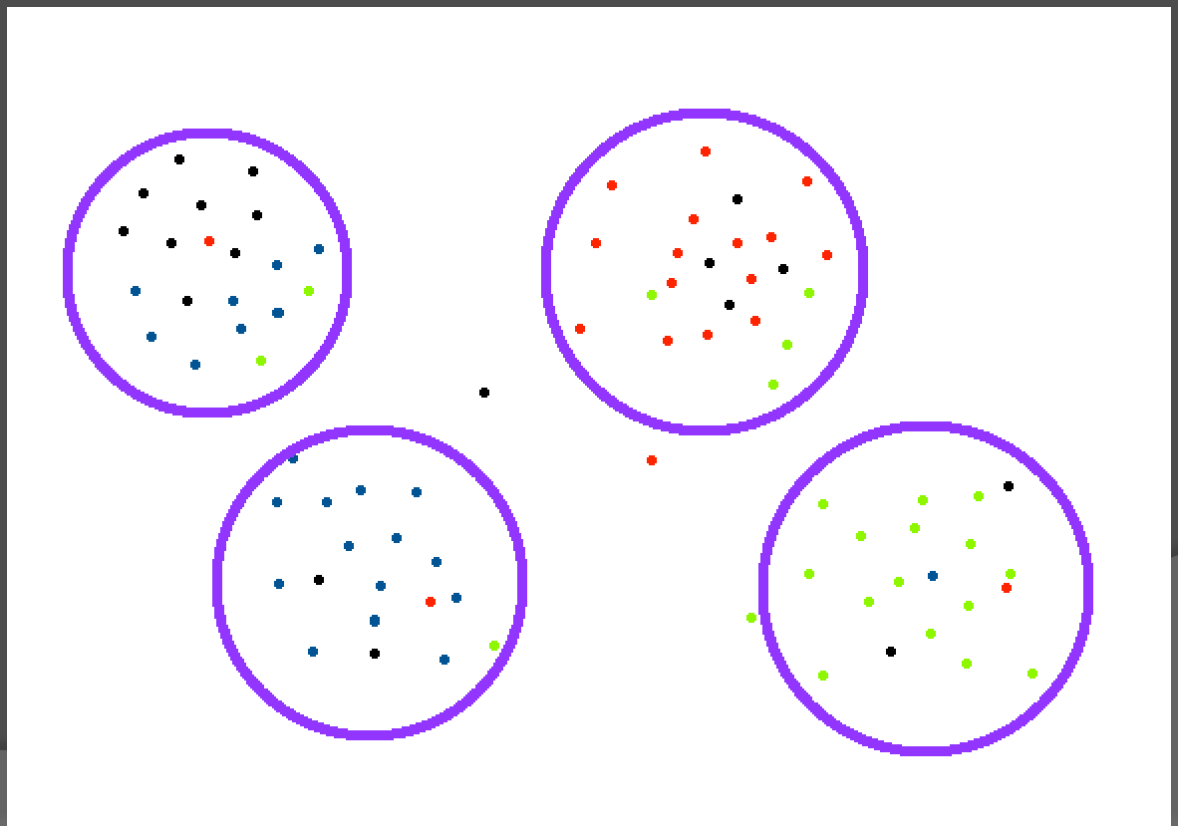
- 从时间角度来对用户用水行为进行分析
  - 一个时间段内用户用水量作为一个向量V
  - 例如以周为单位的向量，那么

    $V = \{ d1, d2, d3, d4, d5, d6, d7 \}$

    $V \Longleftrightarrow id\_date$
  - 采用聚类算法对向量簇进行聚类，输入可以表示成$[V1, V2, V3, V4, \ldots Vn]^T$，聚类结果如图所示。

聚类结果

- 如图，不同颜色的点表示不同用户，因为对用户的很多时间段进行聚类，因此同一颜色的点会有多个。
- 一个用户可能分布在多个聚类簇中，分布规律描述的这个用户的用水特征。



二次聚类

- 将第一的聚类结果分为m个类别（I1, I2, I3, …, Im），那么每个用户的特征向量:

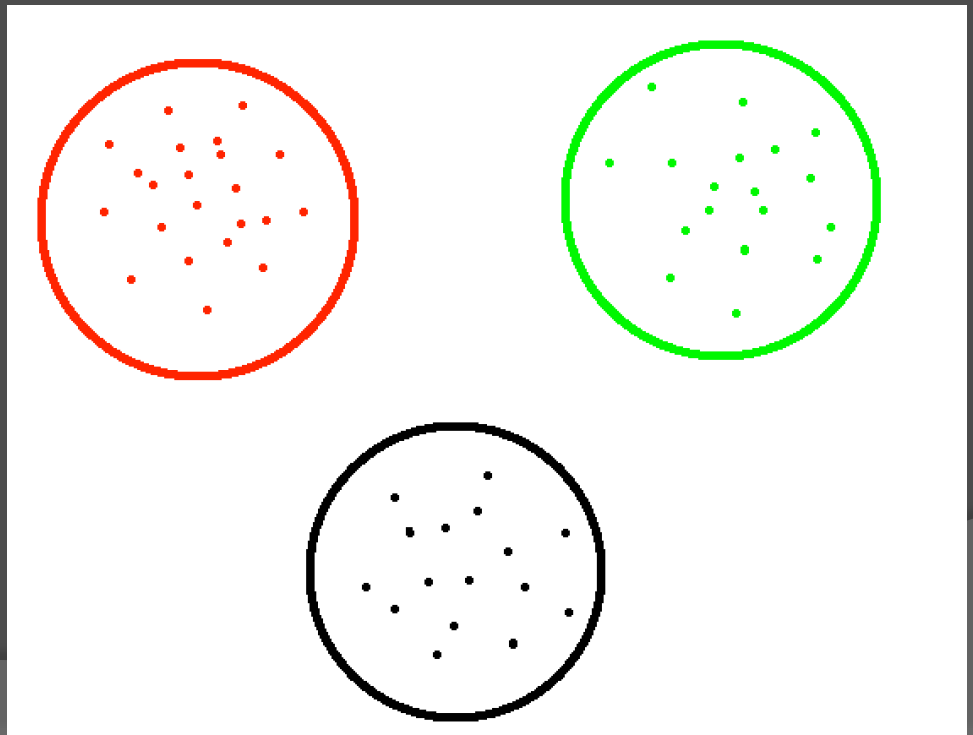
  Vu = {I1, I2, I3, … Im}

- Vu的每一个维度值是在分类Ii中出现的次数。
- 对向量进行归一化处理，每个维度除以该用户出现的总次数total，得：

$$Vu = \{\frac{l1}{total}, \frac{l2}{total}, \frac{l3}{total}, \cdots, \frac{l4}{total}\}$$

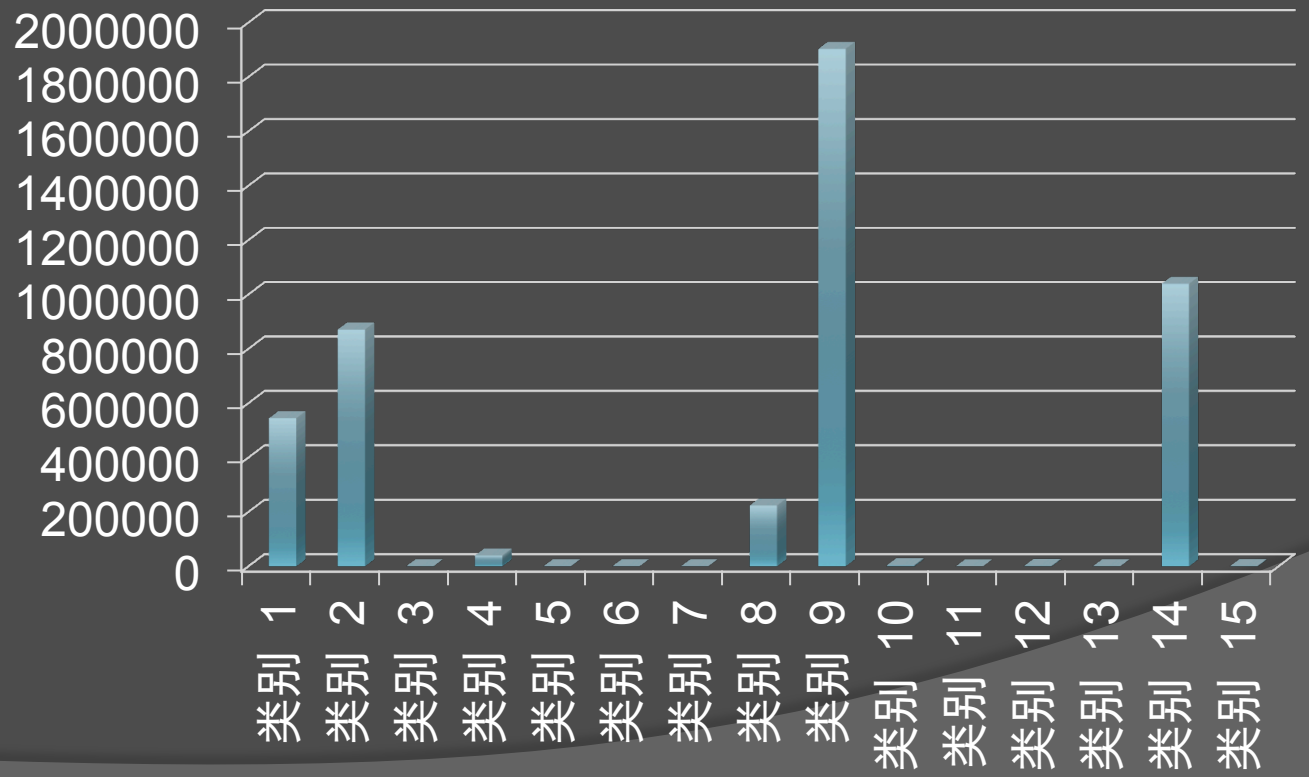
二次聚类结果

- 对Vu进行聚类，可以将相似行为的用户聚在一个类簇中。
- 图中相同颜色的点表示相似行为的用户。



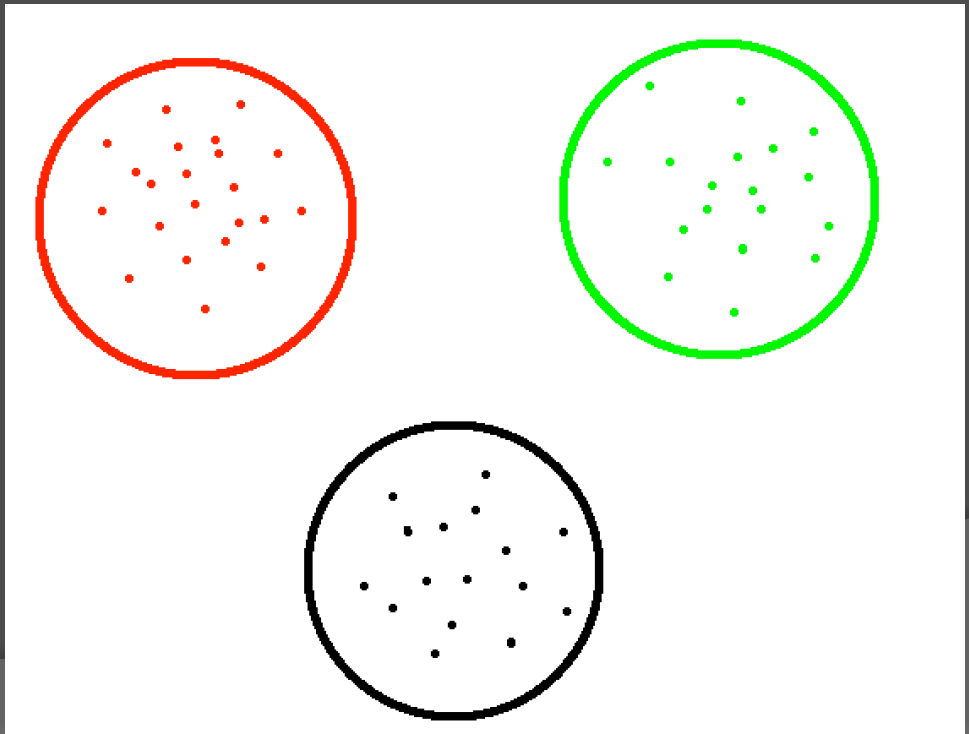
需求一：用水规律分析

- 对二次聚类结果的每一个类簇进行统计分析，可以找出不同类别的用户的用水情况。



需求二：用户分类

- 二次聚类的结果本身就是对用户进行的划分的结果；
- 当一个新用户的数据到来时，可以通过两次聚类模型得到最终用户归属于哪一类别；



需求三：不合理用水用户

- 对于形成一定用水规律的用户，如果以后的某段时间的用水数据在第一次聚类结果中与以往的分布显著不同，则可以将其判断为不规律用水用户；
- 对于用水量过大或过小的用户，可以通过统计分析得到。

需求四：用户用水量预测

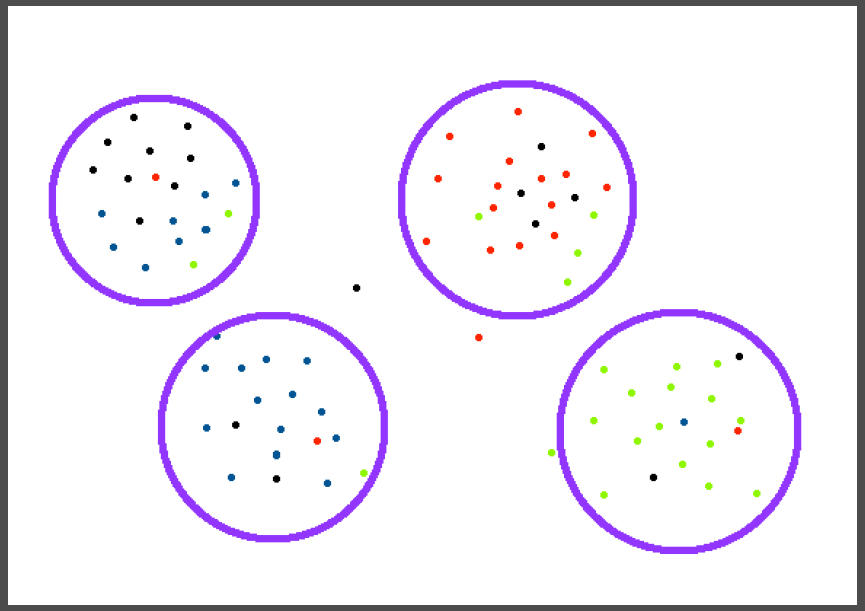
- 这里只考虑短时间内的用户用水量预测，如一天、一周、一月。
  - 建立神经网络回归模型：
    - 引入往年同期用水量这个指标对用户未来时间点用水量进行预测，这样的话输入样本向量为：

$$V = [a_{lastyear}, a_{-4}, a_{-3}, a_{-2}, a_{-1}, a_0]$$

Kmeans聚类过程

第一次聚类的输入和输出

```
0.0,0.0,0.0,0.12,0.13,0.12,0.0
0.05,0.0,0.02,0.0,0.0,0.0,0.0
0.0,0.0,0.0,0.0,0.0,0.04,0.05,0.03
0.05,0.01,0.01,0.0,0.0,0.0,0.02,0.0
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
0.06,0.0,0.0,0.0,0.0,0.0,0.0,0.0
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
0.0,0.0,0.0,0.0,0.0,0.03,0.04,0.19
0.0,0.33,0.39,0.1,0.08,0.06,0.3
0.09,0.31,0.07,0.17,0.07,0.22,0.15
0.1,1.79,0.37,0.22,0.21,0.1,0.11
0.24,0.08,0.11,0.1,0.26,0.23,0.28
0.15,0.24,0.05,0.08,0.08,0.11,0.29
```

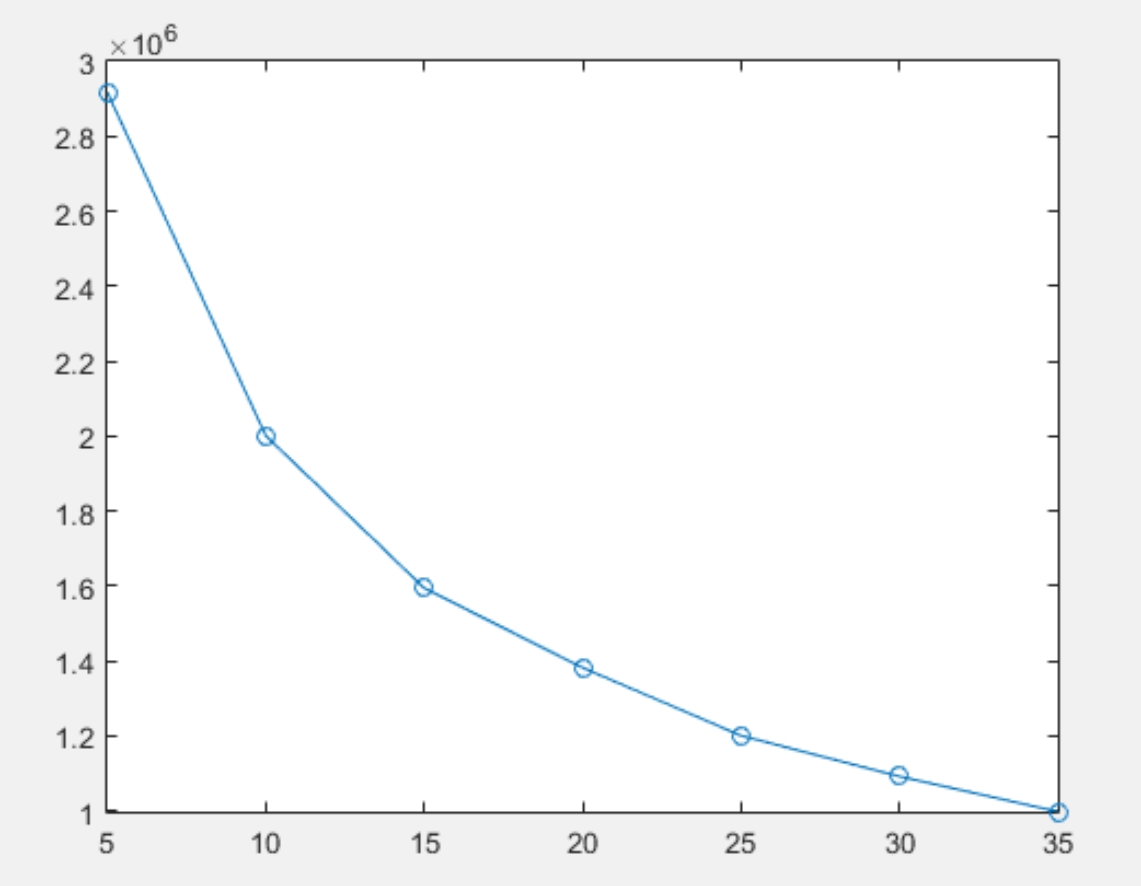


输入文件（每行表示一个用户一周的用水情况）

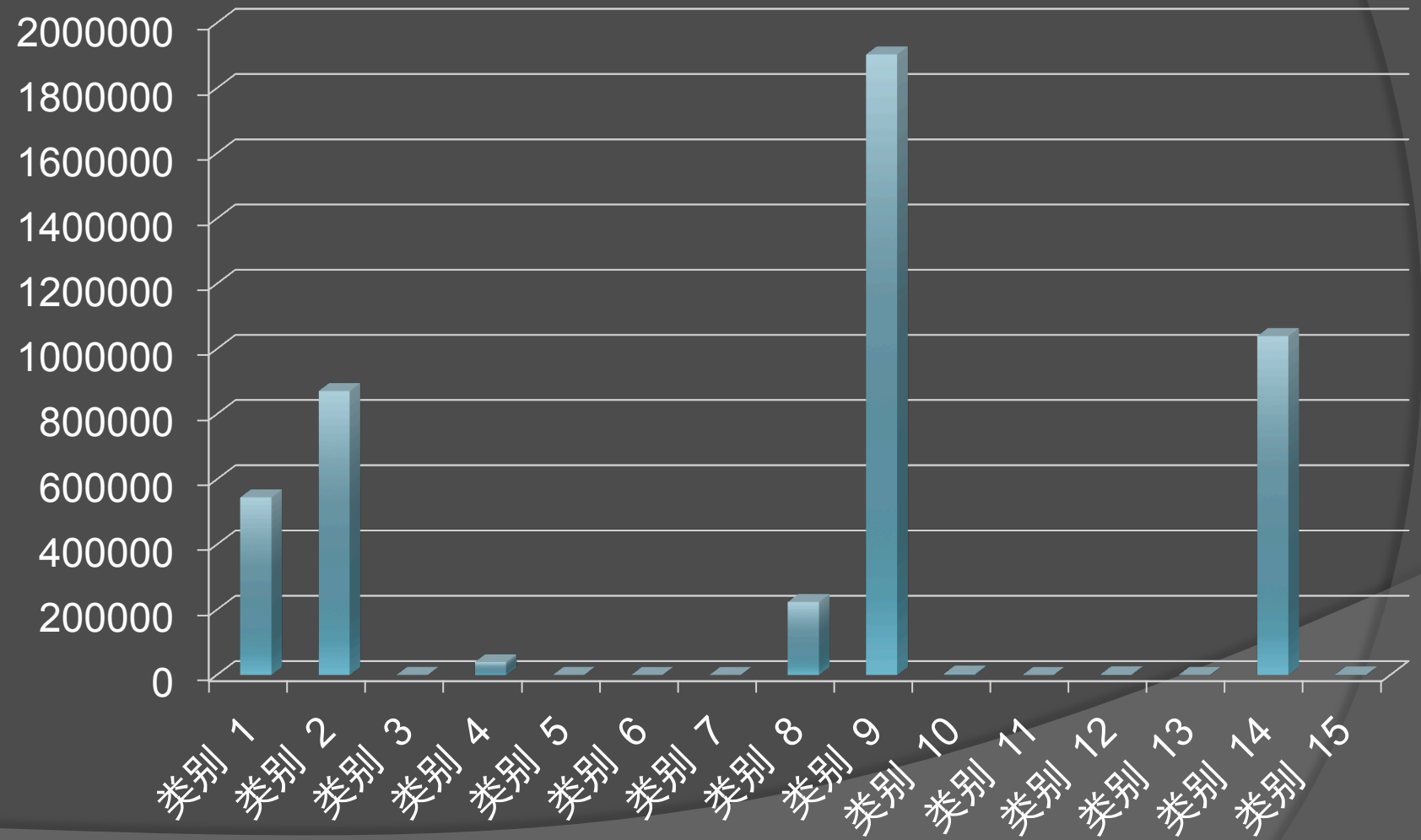
将相似的周向量聚到了一个类簇中

第一次聚类——K值选择

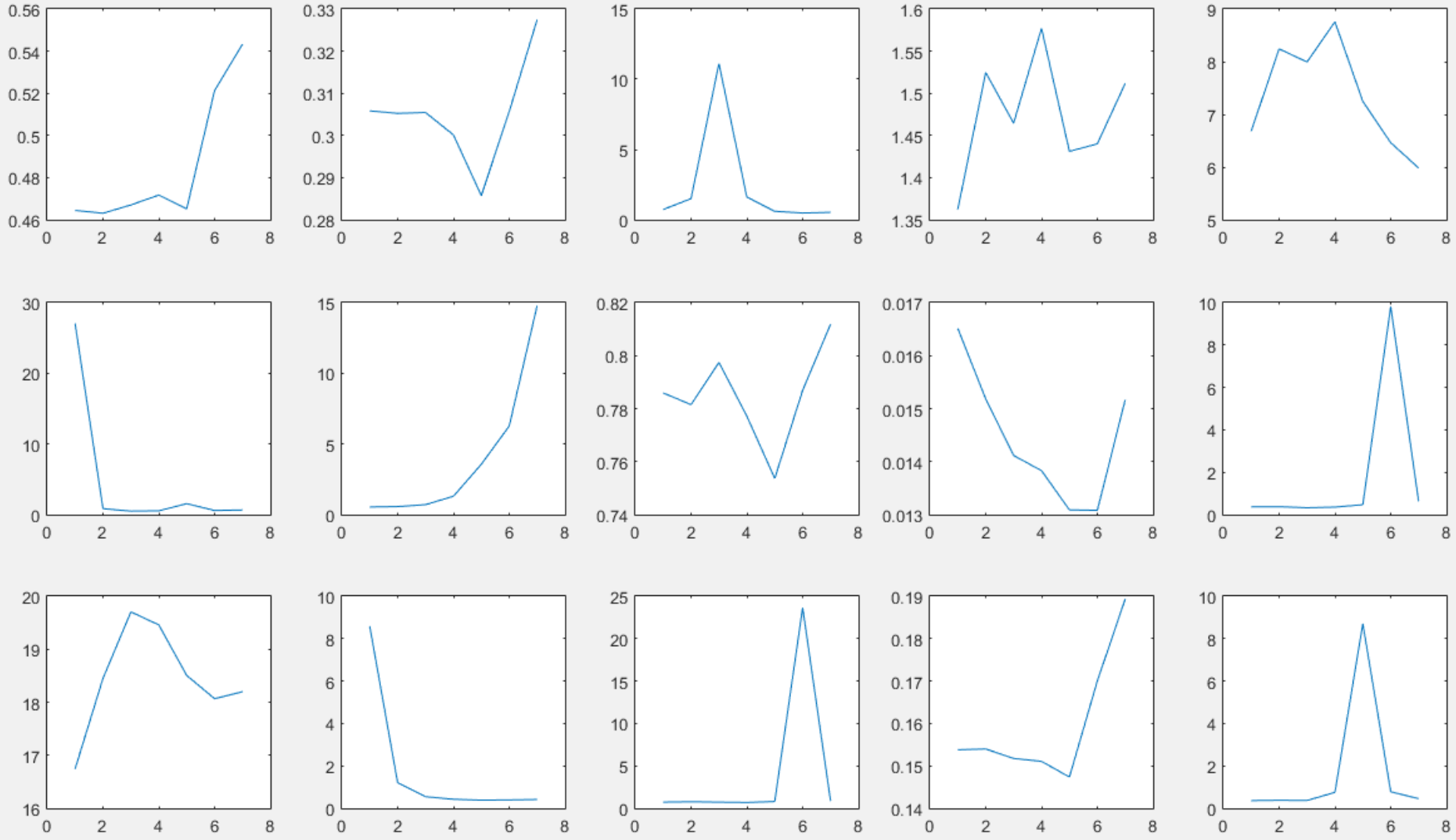
- 图像肘点 k=15， 选取这个k作为聚类参数



15种类型周样本分布情况

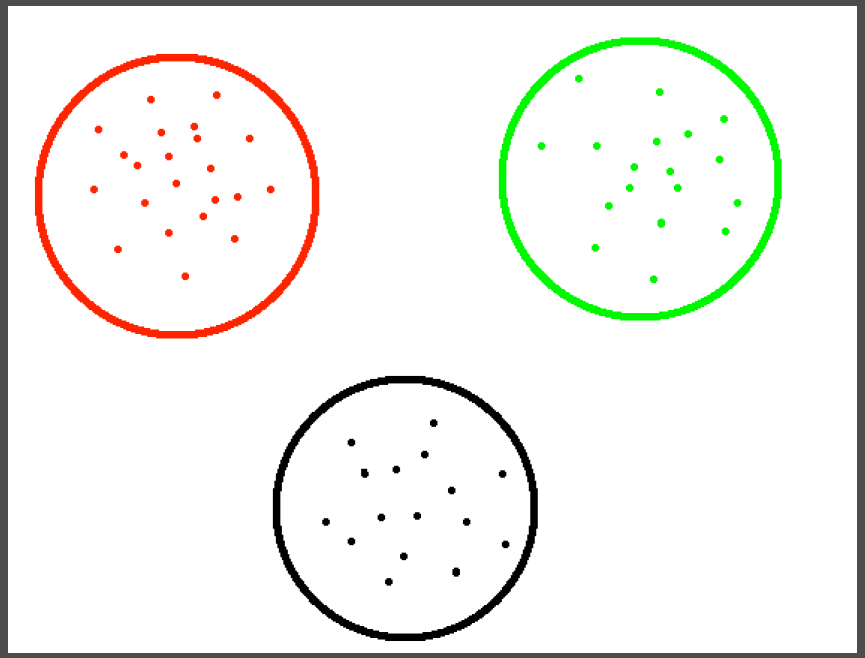


15种类别的周用水图像



第二次聚类——输入和输出

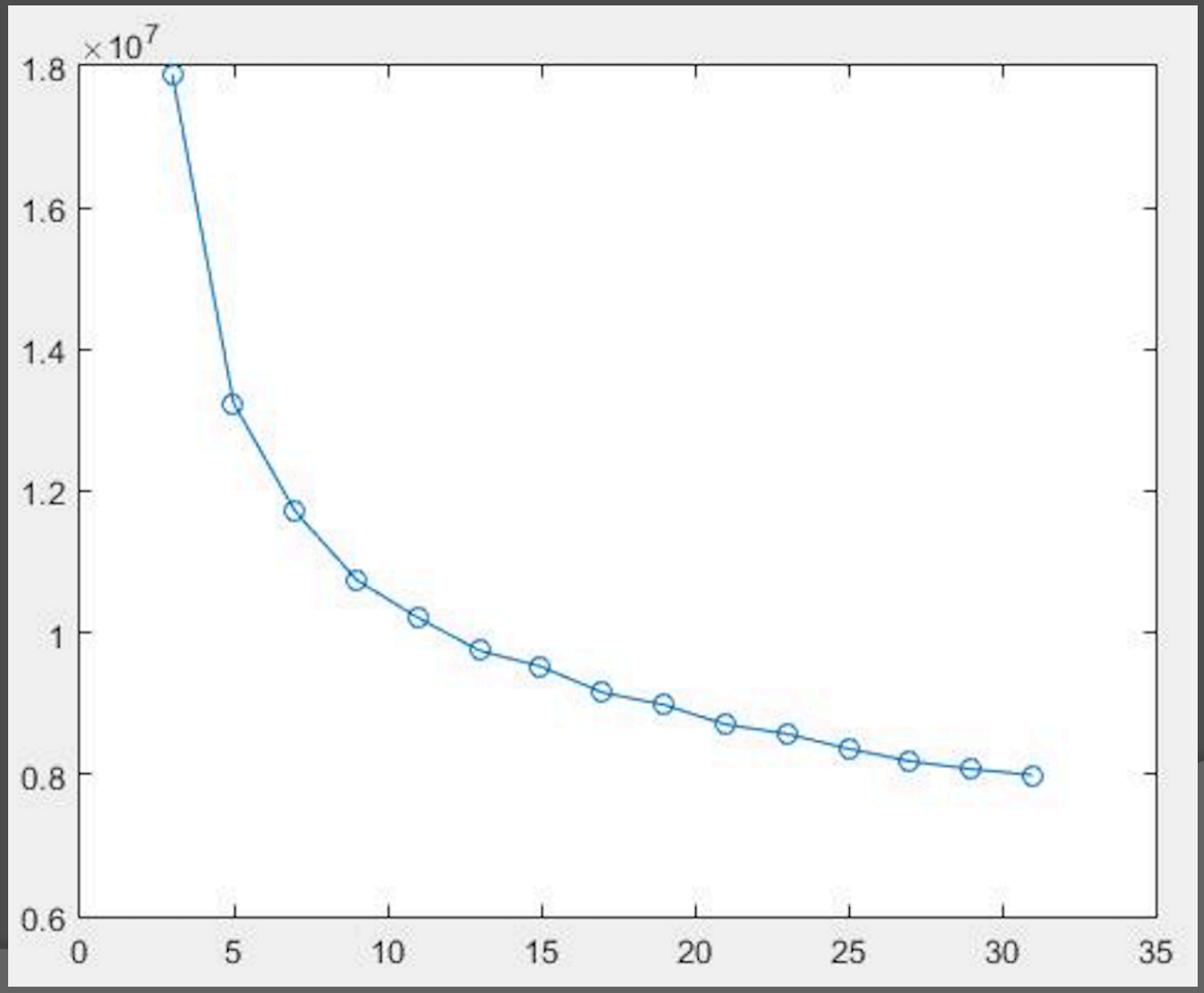
0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
0.0, 0.031746034, 0.0, 0.12698413, 0.84126985, 0.0, 0.0, 0.0, 0.0, 0.0
0.0, 0.0, 0.09497207, 0.0, 0.0, 0.08938547, 0.0, 0.0, 0.8156425
0.03125, 0.0, 0.24375, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.725
0.02631579, 0.0, 0.47368422, 0.0, 0.0, 0.0, 0.0, 0.13157895, 0.0, 0.36842105
0.0, 0.0, 0.63839287, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.36160713
0.0, 0.0, 0.027173912, 0.0, 0.0, 0.875, 0.0, 0.0, 0.097826086
0.109375, 0.0, 0.03125, 0.0, 0.0, 0.0, 0.4375, 0.015625, 0.40625, 0.0
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.06428572, 0.0, 0.0, 0.9357143
0.005076142, 0.0, 0.13197969, 0.0, 0.0, 0.0, 0.80203044, 0.02538071, 0.005076
0.03846154, 0.0, 0.23076923, 0.0, 0.0, 0.0, 0.65384614, 0.0, 0.0, 0.0, 0.07692308
0.044444446, 0.0, 0.8333333, 0.0, 0.0, 0.0, 0.0, 0.12222222, 0.0, 0.0
0.016666668, 0.0, 0.0055555557, 0.16666667, 0.0, 0.0, 0.0, 0.33888888, 0.4666
0.0, 0.0, 0.045918368, 0.0, 0.0, 0.0, 0.75510204, 0.010204081, 0.0, 0.18877551

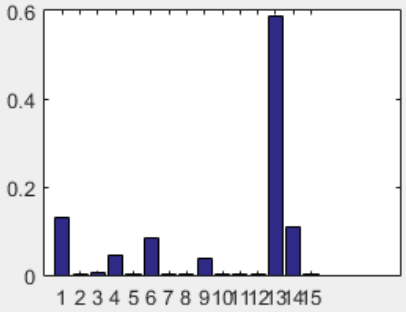


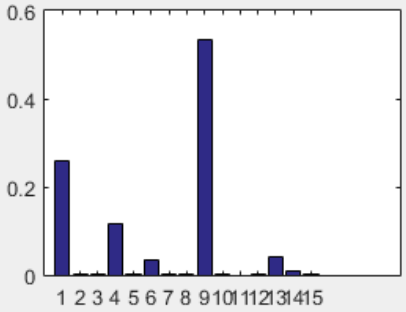
每行输入是归一化后的每个用户的用水类型分布情况

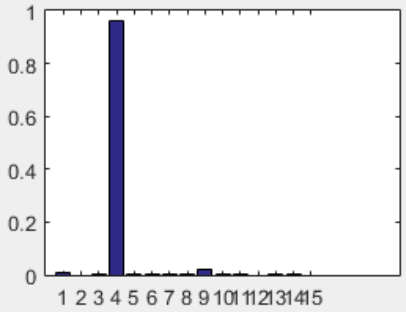
每个类簇表示用水行为相似的用户群

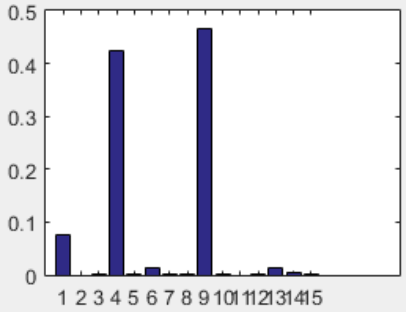
用户用水的第二次聚类——WSSSE曲线

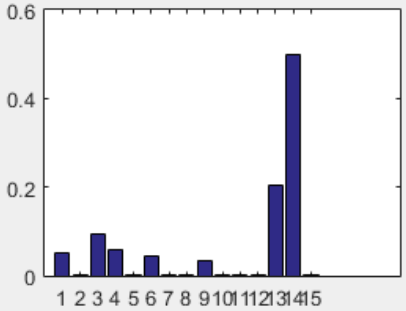


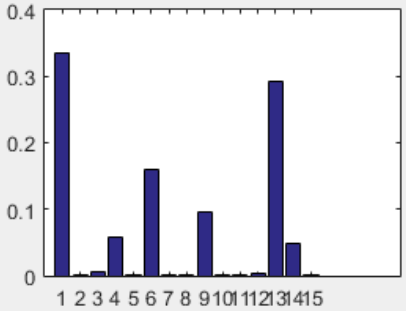


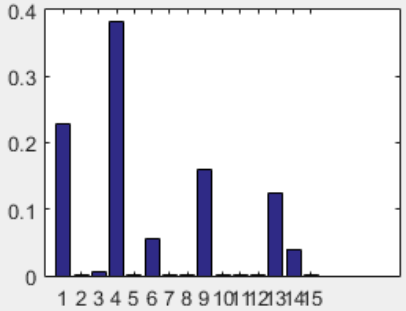


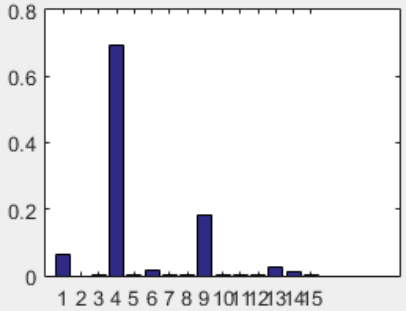


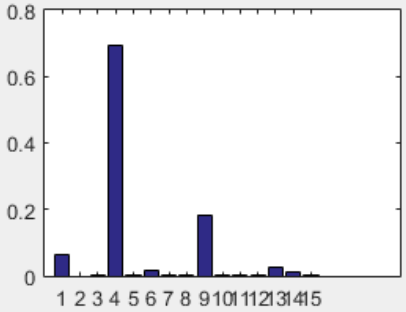


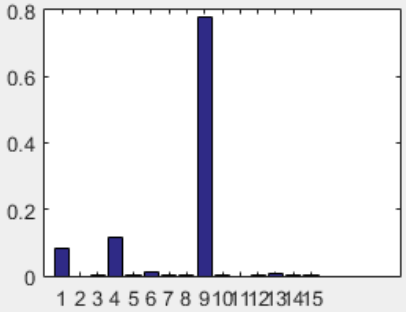


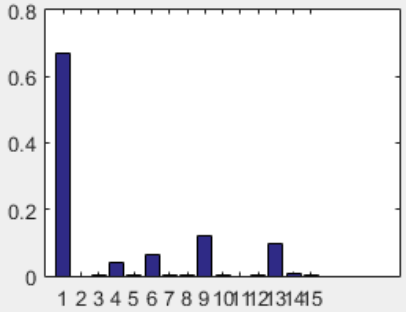








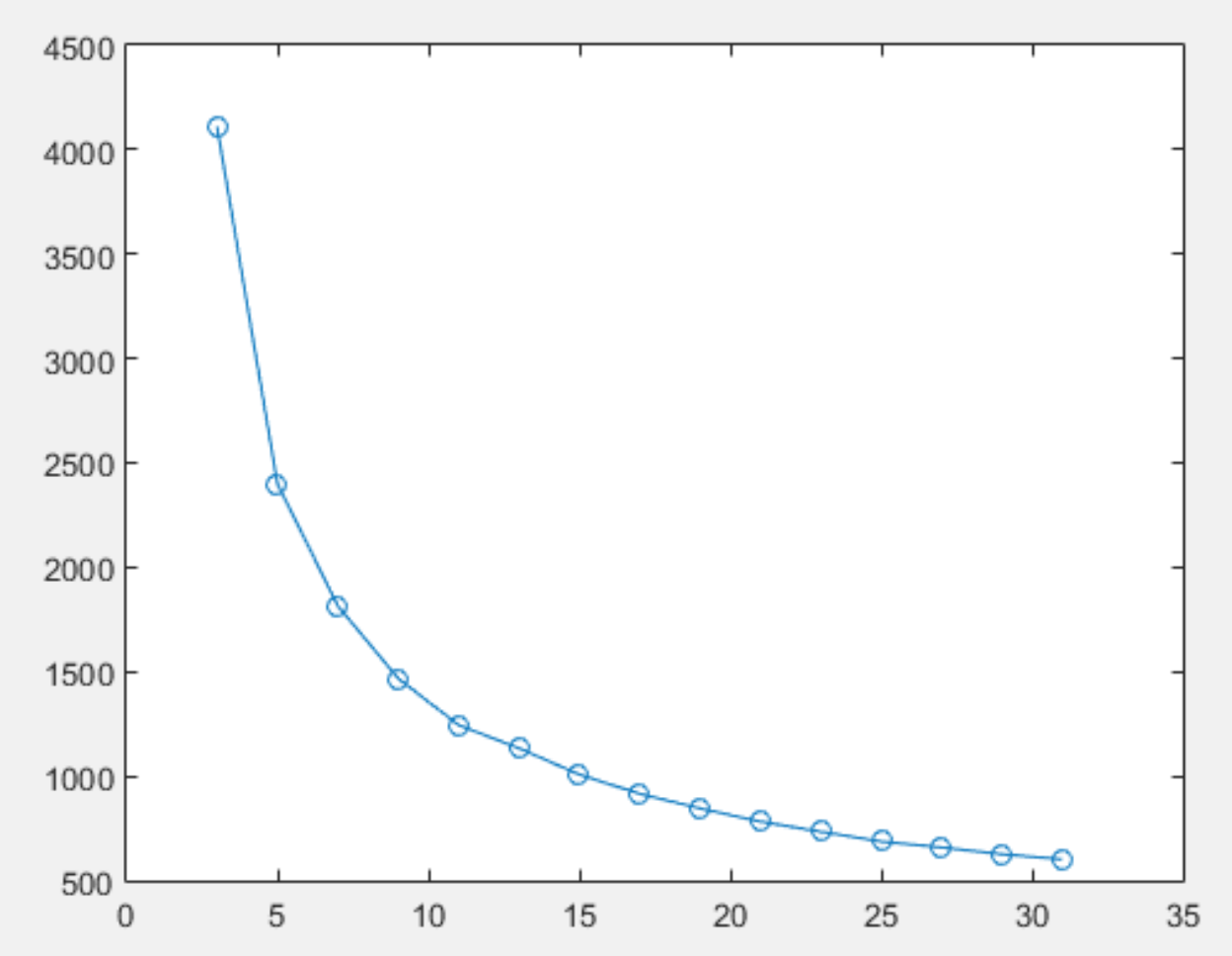




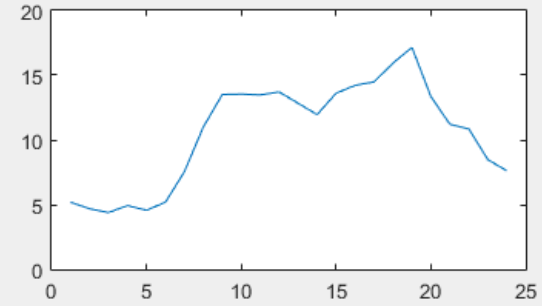
11种类别用户用水分布

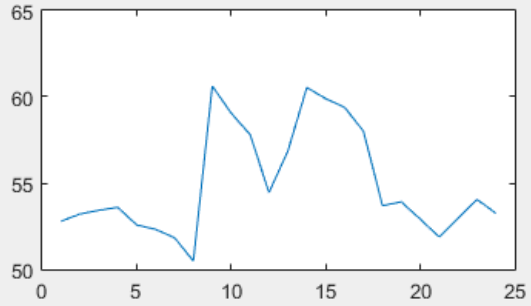
工厂用水的第一次聚类

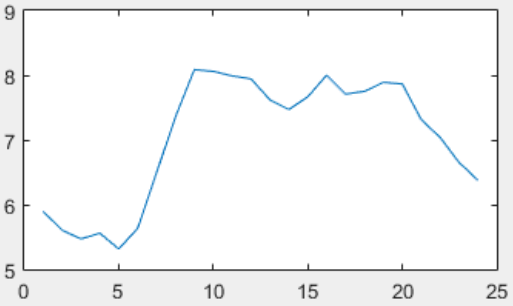
- 肘点位于 k = 9的位置

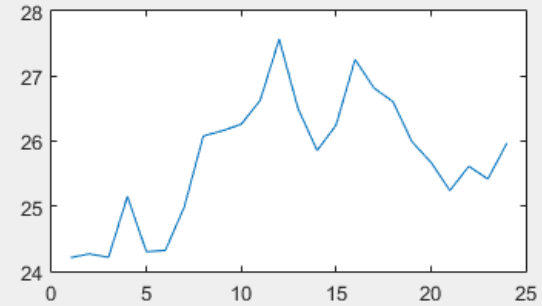


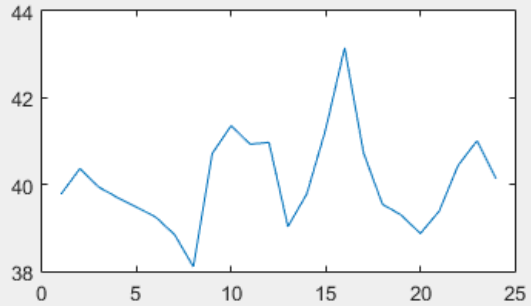
9种类别的工厂日用水图像

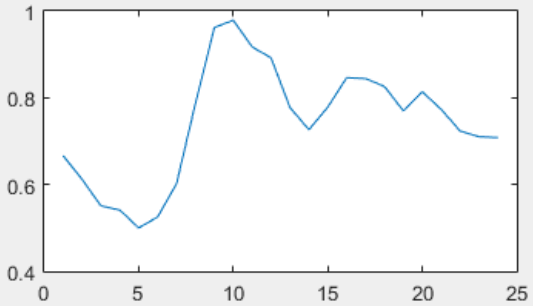


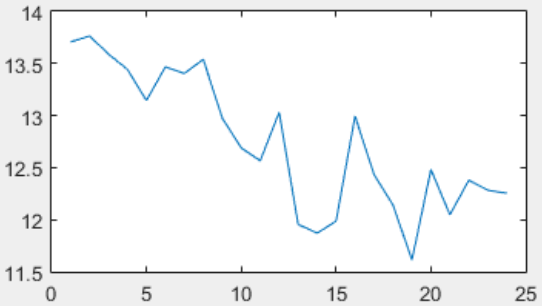


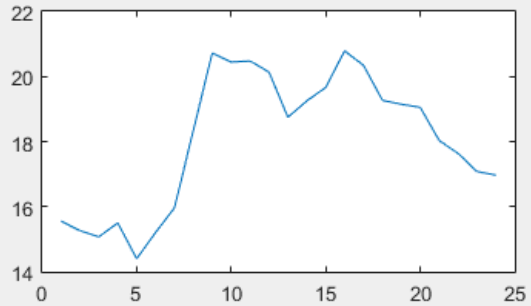


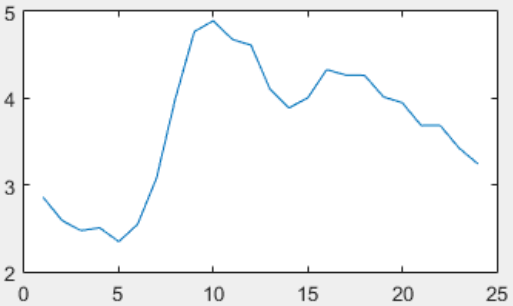






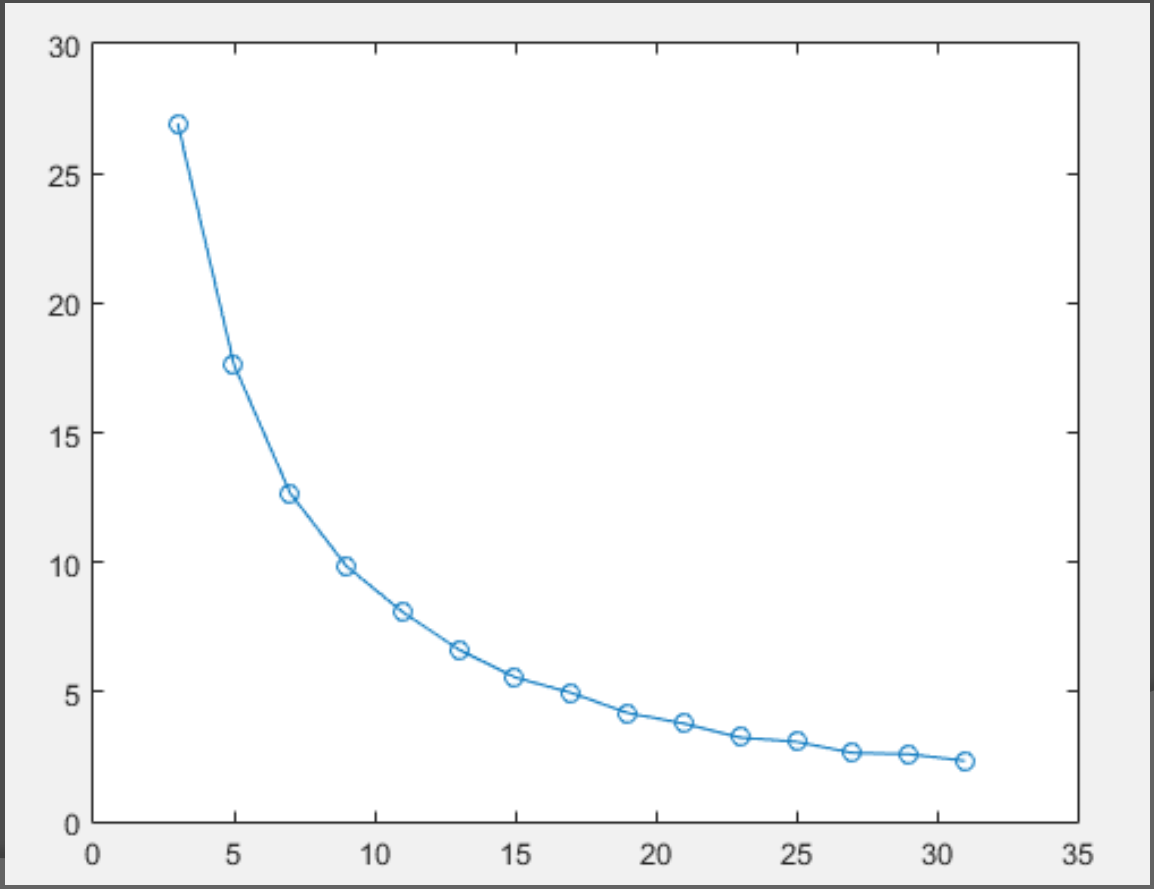




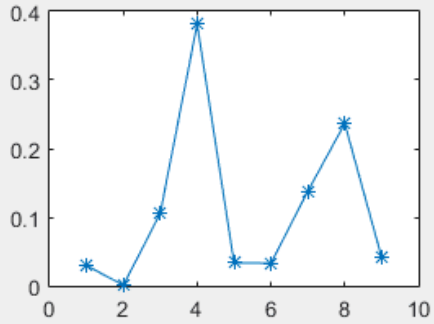
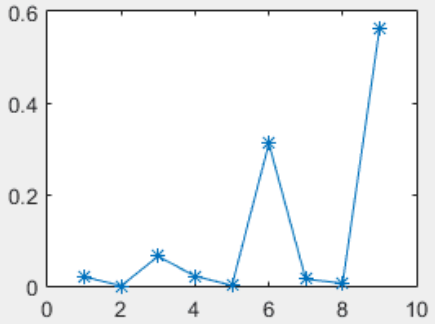
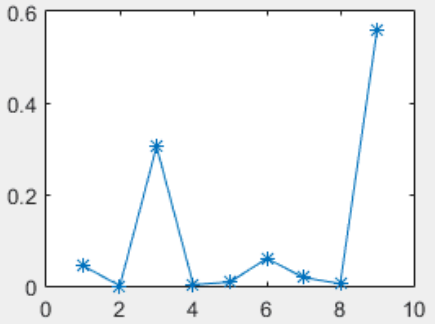
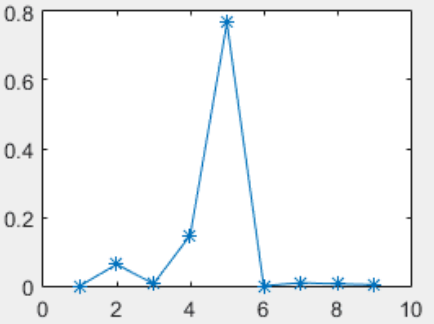
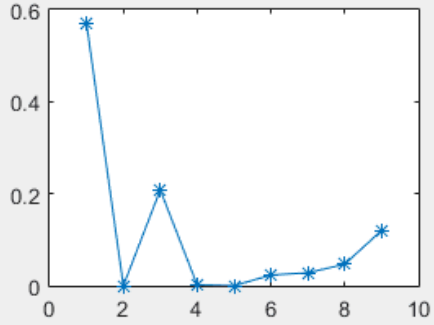
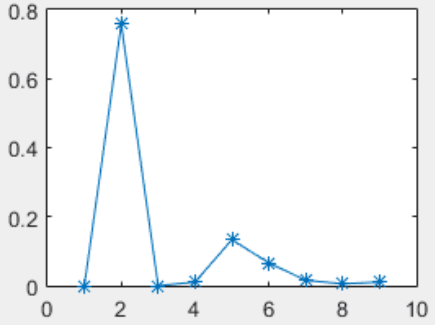
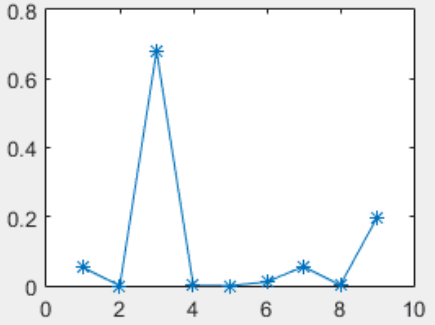
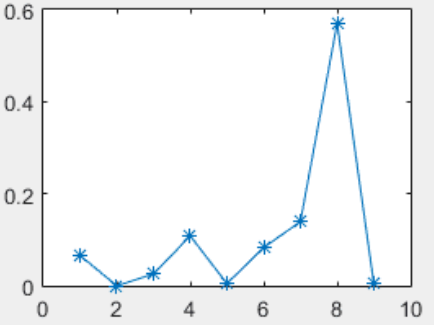
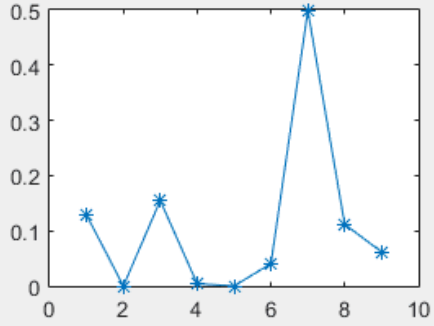
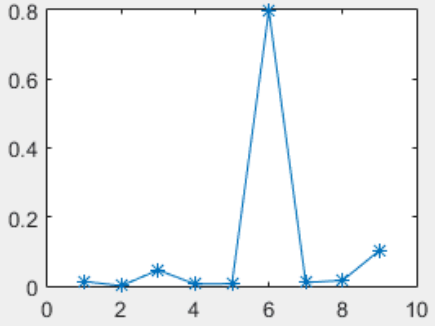
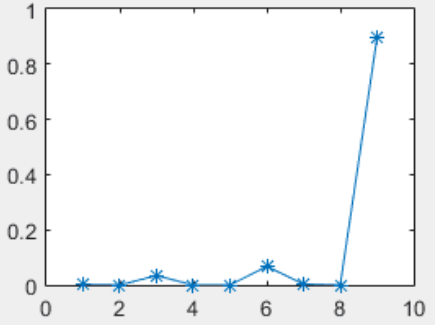


工厂用水的第二次聚类

- 第二次聚类的目的在于将用户划分到不同的类别中去，这里取K = 11

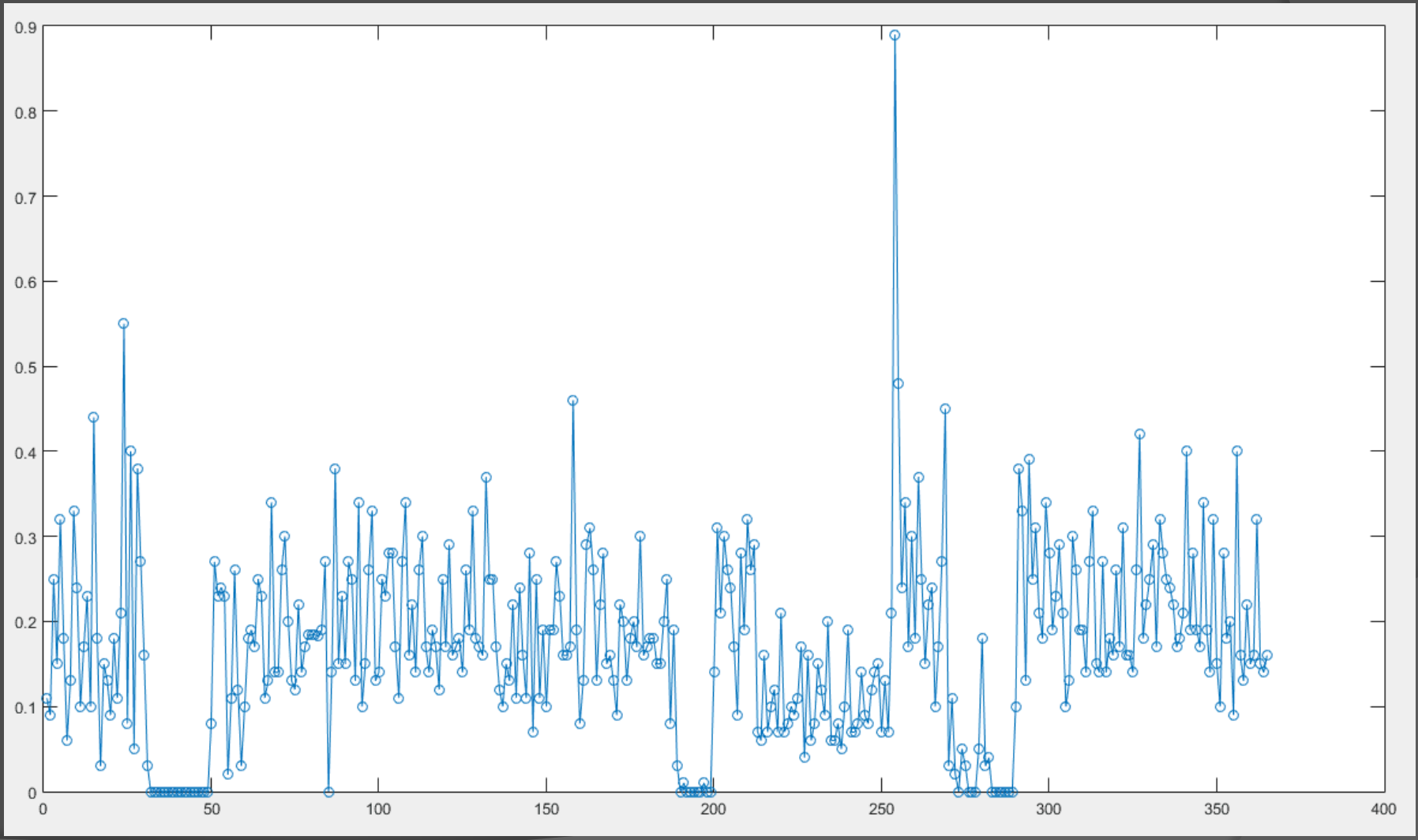


11种工厂用水分布情况（簇中心）

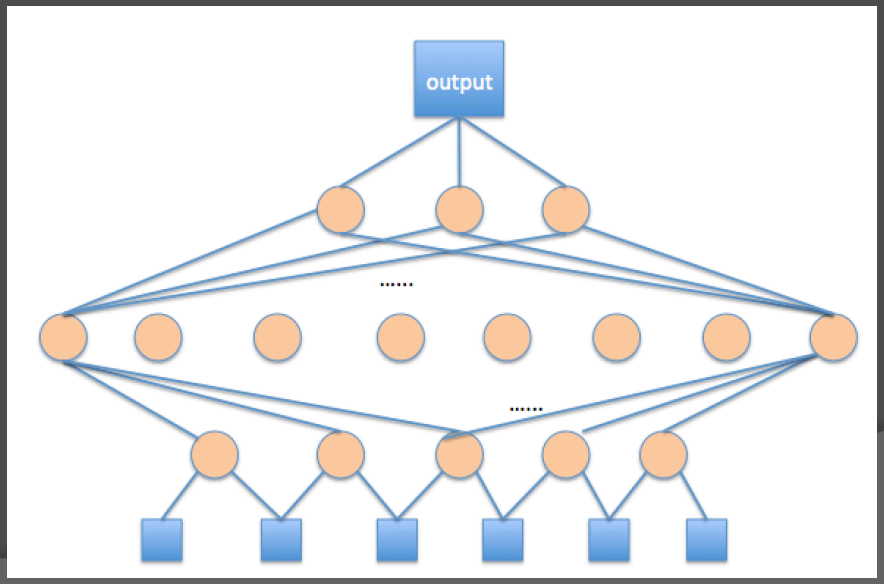
预测与回归

某个用户一年的用水图像



神经网络回归方案

- 对某个用户构建一个训练集，其中每行数据是一个六维向量，分别表示去年同期用量和前3、2、1天的用量以及当日用量。
- $V = [a_{lastyear}, a_{-4}, a_{-3}, a_{-2}, a_{-1}, a_0]$

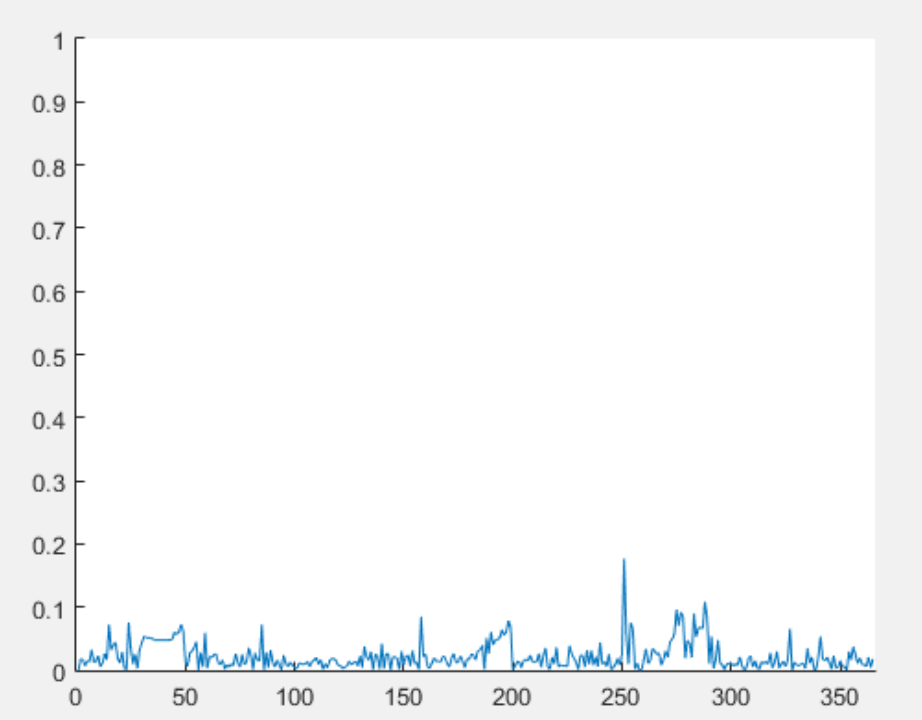


使用神经网络对用户一年用水进行拟合

- 交叉验证误差

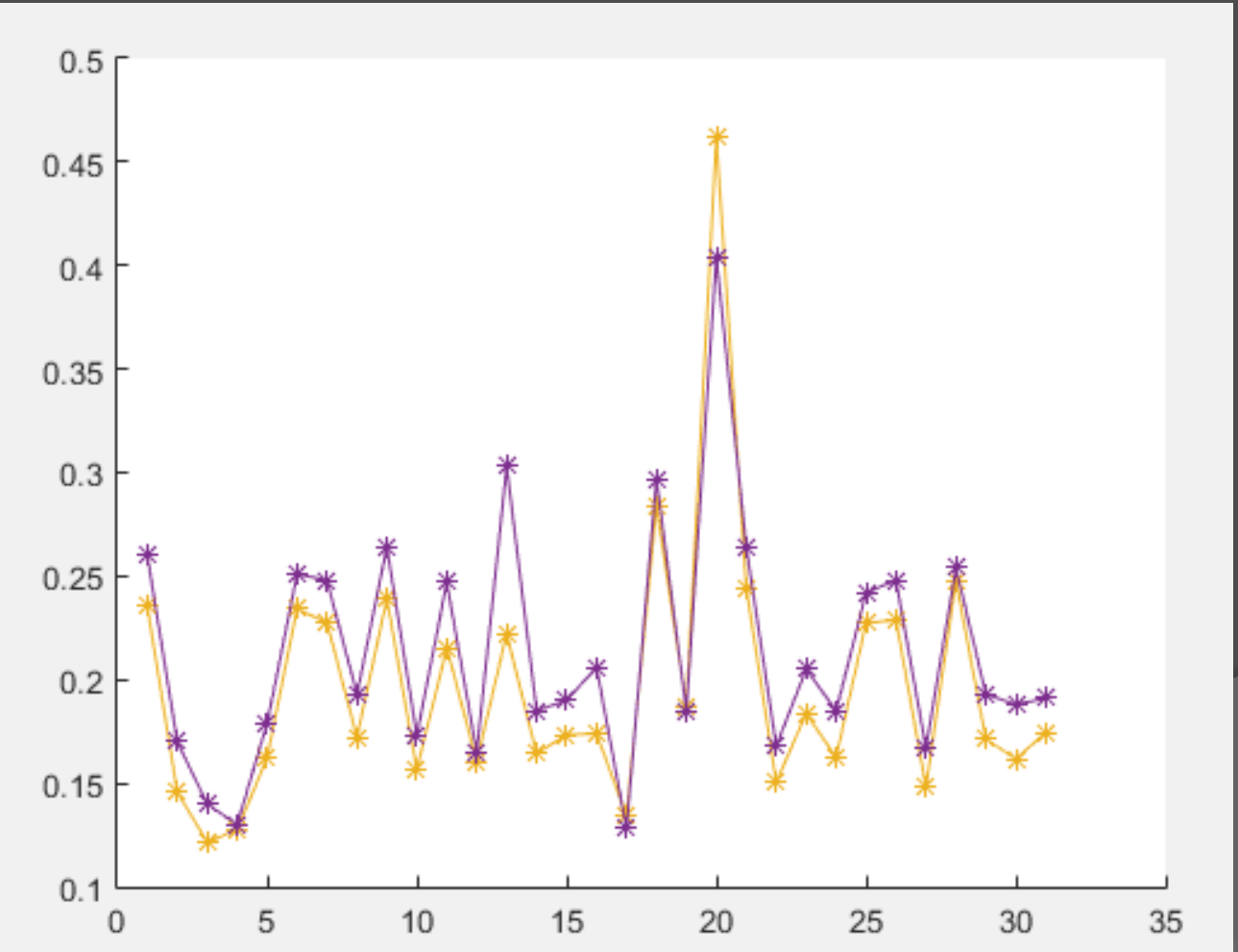
```
Training error: 0.003913106557767807
Validation error: 0.005394808097225701
```

- 误差曲线



对一个月的用户用水的预测情况

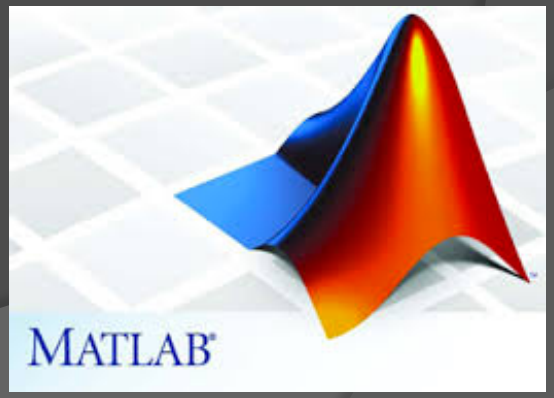
- 橙色表示实际量，紫色表示预测量



其他

实验用到的工具

- Spark Mlib：Kmeans聚类
- Encog：神经网络回归
- Matlab：实验结果可视化
- Spark SQL：数据清洗与数据准备



实验分工

- 李克西
- 谢尚广
- 刘兴

算法设计与实现

- 杨巧杰
- 钱伟
- 夏杰

数据清洗