**八、问题四的模型与建立**

**8.1 问题分析**

在本题中，自变量为婴儿的整晚睡眠时间，睡醒次数，以及婴儿的入睡方式，因变量为婴儿的睡眠质量，分为优，良，中，差四类，需要根据综合评判，得到对应的婴儿睡眠治疗的指标。

应用评价型模型的难点在于，婴儿的年龄不相同，入睡方式不同，且数据繁多，难以统一标准，因此，本文采用了无监督分类模型的Kmeans，并规定分成的簇的数量为4，即对应优，良，中，差，共4类，再根据4堆簇的中心点指标，去判断该簇属于优，良，中，差中的一类。

**8.2 Kmeans算法**

**8.2.1 Kmeans算法原理**

已知数据集，Kmeans聚类要把这个数据划分到个集合中

，使得组内平方和最小，它的目标是找到使得下式满足的聚类

其中是中所有点的均值。

**8.2.2 Kmeans算法步骤**

（1） 对数据集进行标准化和归一化，避免均值和方差大的数据对聚类产生决定性影响

（2）选择初始化的个样本作为初始聚类中心

（3）针对数据集中每个样本，计算它到个聚类中心的距离，并将其分到距离最小的聚类中心所对应的类中

（4）针对每个类别，重新计算它的聚类中心，即属于该类的所有样本的质心

（5）重复(3)(4)步，知道达到某个中止条件（迭代次数，可允许最小误差等）

其伪代码为：

获取数据 n 个 m 维的数据

随机生成 K 个 m 维的点

while(t)

for(int i=0;i < n;i++)

for(int j=0;j < k;j++)

计算点 i 到类 j 的距离

for(int i=0;i < k;i++)

1. 找出所有属于自己这一类的所有数据点

2. 把自己的坐标修改为这些数据点的中心点坐标

end

**8.2.3 Kmeans模型求解**

通过Python的可视化库matplotlib可以直观地展现出聚类散点图，由于变量数大于2个，即取主成分分析(PCA)降维后前两个主成分来绘制散点图，在一定程度上可查看聚类效果。

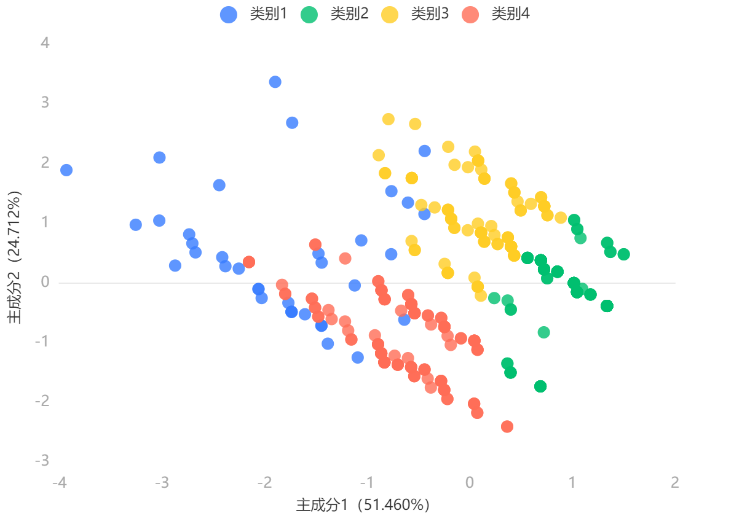


图4.1 聚类散点图

表4.1 聚类中心点坐标



根据附件可知，研究的婴儿年龄均为1~3个月大小，并通过查阅相关资料表4.2，可以知道睡眠质量优，良，中，差分别为聚类中心点中的1，3，4，2。

表4.2 0~1岁婴儿睡眠规律图



8.3 KNN预测

和问题二预测同理，这里仍然使用有监督的分类算法KNN，以母亲的身体指标，心理指标为自变量，将婴儿的睡眠的睡眠质量作为因变量，并枚举超参数寻优

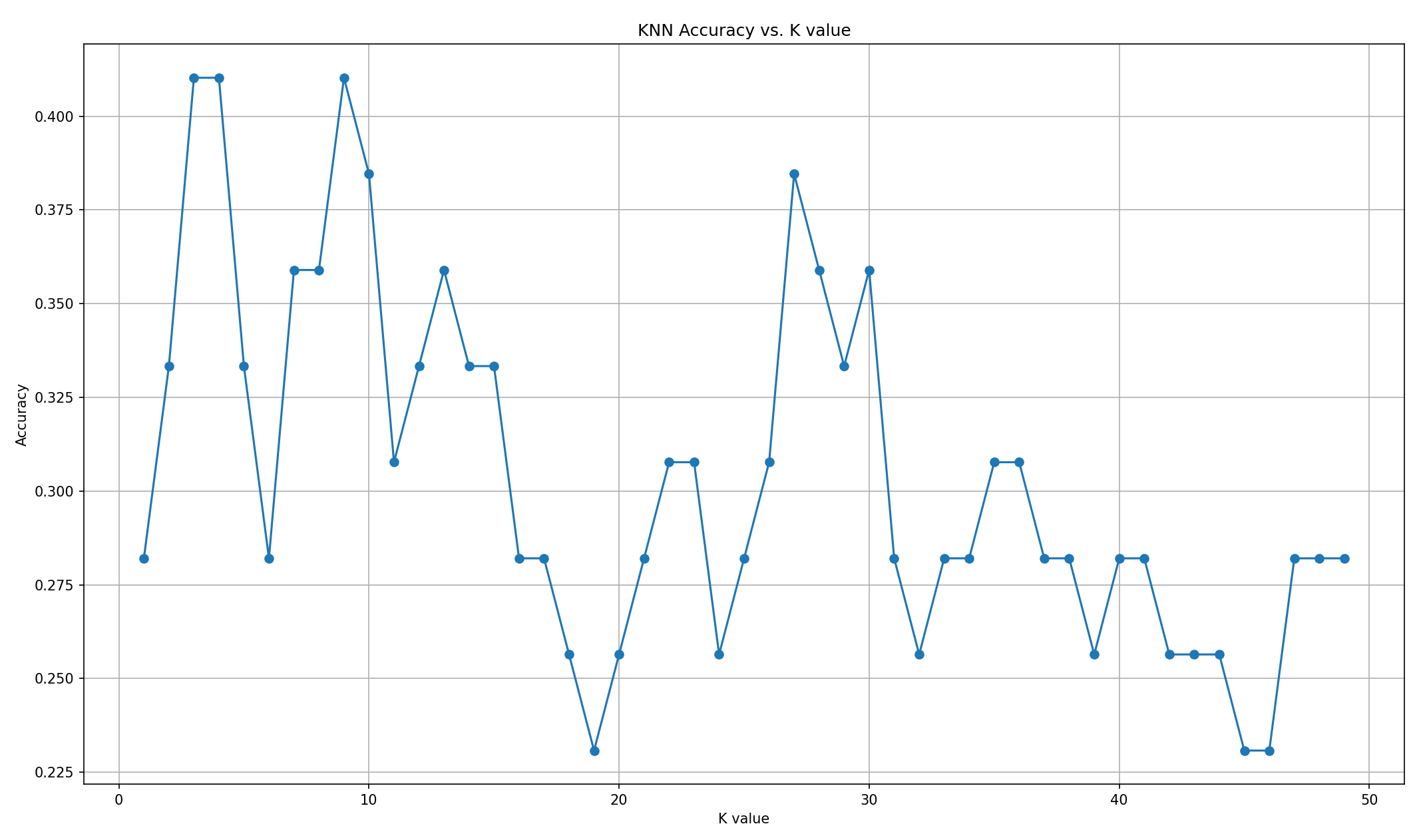


图4.2 KNN中K值寻优时的测试集准确率

得到最后20组（编号391-410号）婴儿的综合睡眠质量如下：

表4.3 编号391-410号婴儿的综合睡眠质量预测结果表

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 391 | 392 | 393 | 394 | 395 | 396 | 397 | 398 | 399 | 400 |
| 分类 | 中 | 中 | 差 | 中 | 中 | 优 | 差 | 中 | 中 | 中 |
| 编号 | 401 | 402 | 403 | 404 | 405 | 406 | 407 | 408 | 409 | 410 |
| 分类 | 良 | 良 | 良 | 差 | 中 | 优 | 中 | 中 | 中 | 中 |