frog 实验

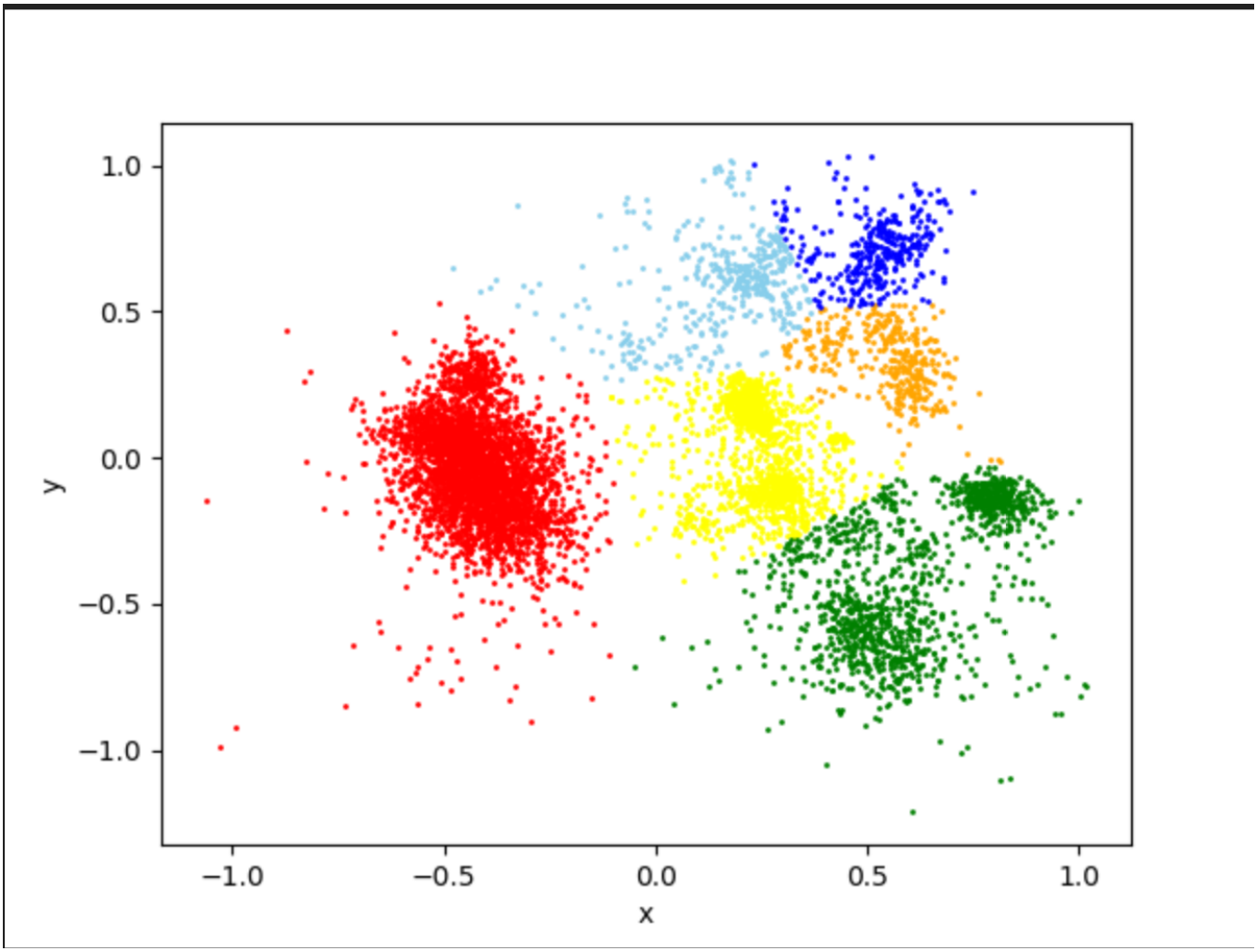
PB16060674-归舒睿

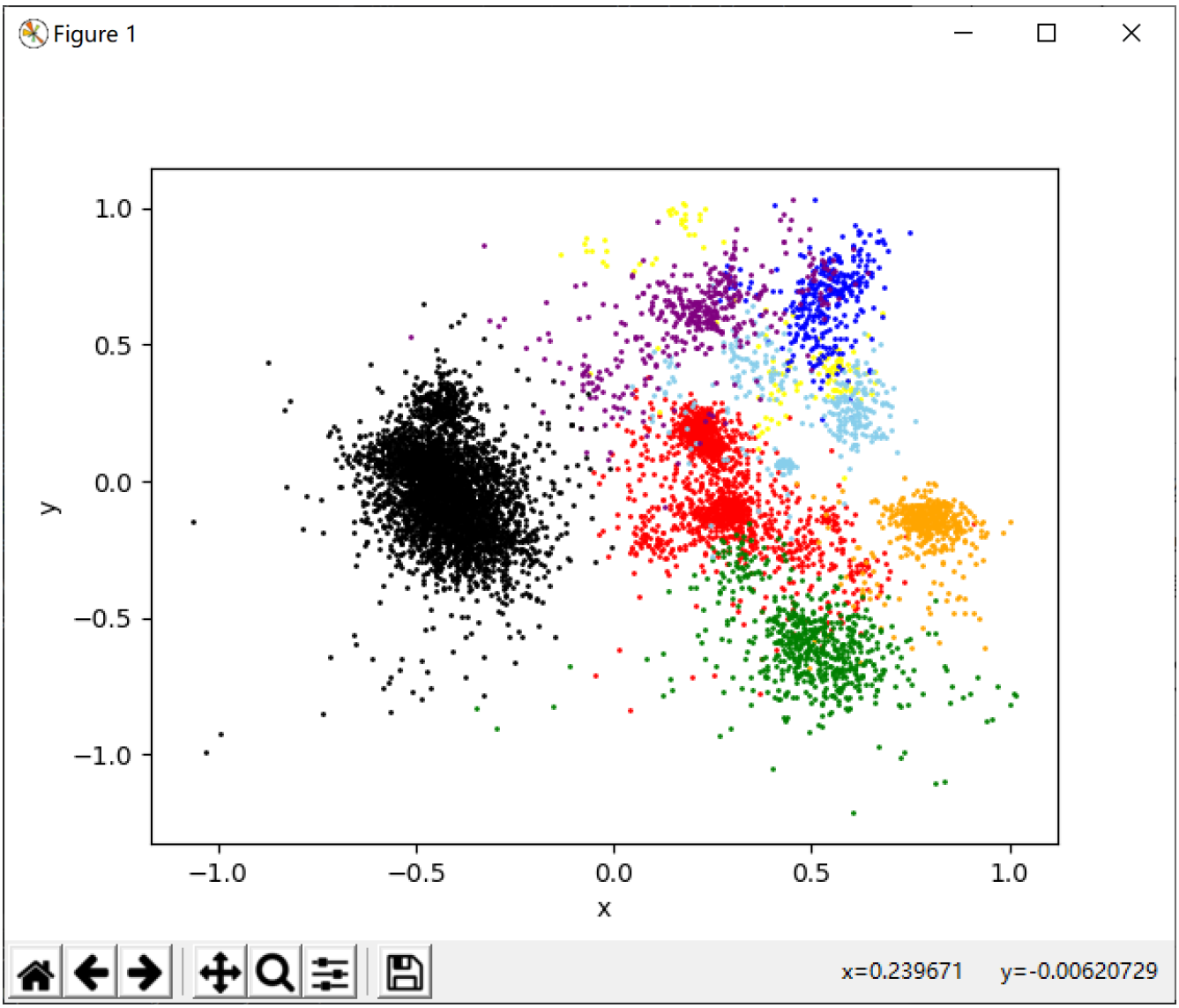
Kmeans算法k选择的依据

遍历k从1到9，得到每组的purity和RI值，选出RI最高的k作为选中的k。

Kmeans评估结果

类型	最好的k	purity	RI
PCA前	9	0.9149	0.7773
PCA后（阈值0.5）	6	0.8435	0.7691
PCA后（阈值0.9）	8	0.8878	0.7750


1.0
0.5
0.0
−0.5
−1.0
y
−1.0
−0.5
0.0
0.5
1.0
x

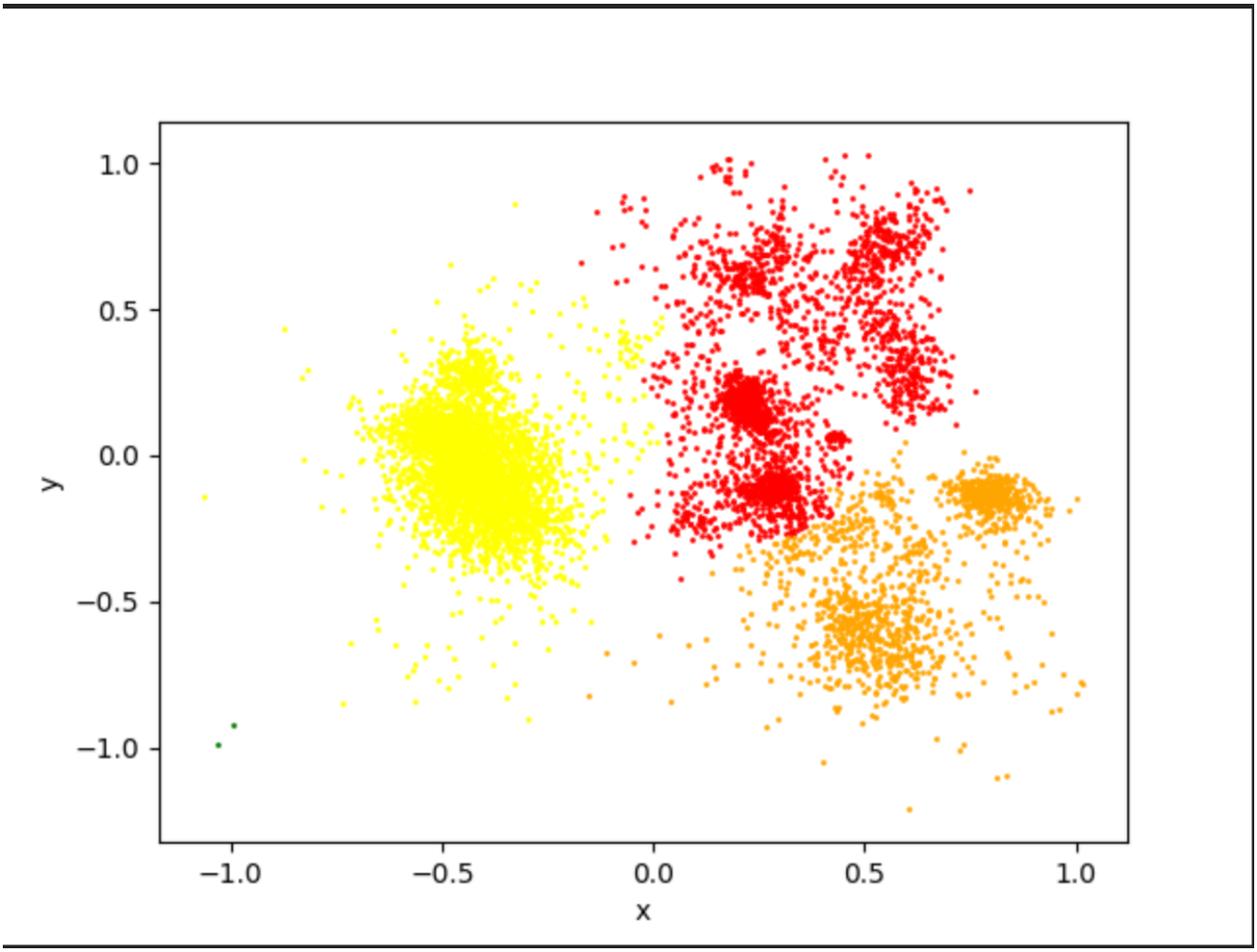


分析：PCA后训练速度明显加快，而且兰德指数没有受到太大影响，纯度一定程度的下降，说明虽然信息有所缺失，但是主要的信息都保留下来了。threshold为0.5时，保留的信息不如为0.9的时候多，所以，阈值为0.9时效果更好

层次聚类分析

层次聚类的评估结果，为了获得好的可视化结果，利用PCA进行了降维。

类型	数据量	k	purity	RI
层次聚类（距离为中心距离）	100%	4	0.7668	0.7221



层次聚类的缺点在于，如图上，左下角的两个绿色点为一类，这是由于，离群点很容易自成一类，这是比Kmeans不好的地方，这次总体效果来说没有Kmeans好（也有没有选出最好的k的缘故）

时间上层次聚类最久，消耗了接近一个小时：全数据，优化后。