改进模型总结

1、如何针对 k-means 算法改进模型参数？

k-means 算法在具体实现中包含了 11 个可以调整的参数:

n_clusters	init	n_init	max_iter	tol	precompute_distances
verbose	random_state	copy_x	n_jobs	algorithm	

可以根据聚类结果修改模型参数，从而改进模型，得到尽量好的聚类结果

本次案例中，仅通过对 n_clusters 这个参数进行修改来改进模型，分别设置 n_clusters = 2，3，4，5 实现多次聚类，并对每次聚类结果进行筛选

注：更多 k-means 算法介绍可访问如下网址：

http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

2、如何针对 k-means 算法改进数据选取？

改进数据选取就是调整聚类所使用的数据，在对每次聚类结果进行解释的同时，会发现数据集中的数据可能存在改进的必要性。

例如案例中的“亲戚为官数”一列在综合考虑后进行舍弃，舍弃的理由：

1）每次聚类结果中不同类别的官员之间“亲戚为官数”差异性极小

2）政治关系数具有 6 个维度，而“亲戚为官数”仅 1 个维度

3）通过韦恩图发现“亲戚为官数”仅少数官员具有该值

3、如何判断最佳聚类结果？

k-means 聚类的评估方法有计算轮廓系数等。较为简单的方法是结合领域/业务知识或者邀请相关专家，人为观察聚类结果，衡量其聚类结果的可解释性。可解释性越高，说明聚类结果越好