**数据挖掘 实验报告**

院系＿＿软件学院＿ 专业＿软件工程＿

年级 2015级 课程班级＿

实验时间 2018 年＿5 ＿月 11＿日

实验名称 **聚类分析**

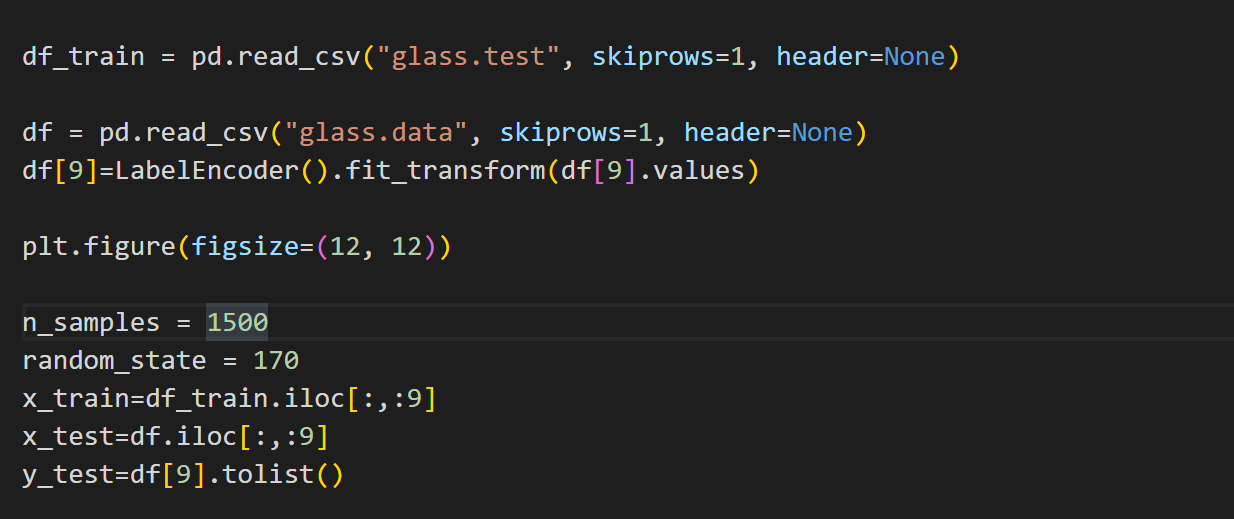
指导老师 刘昆宏

# 第一部分 实验内容

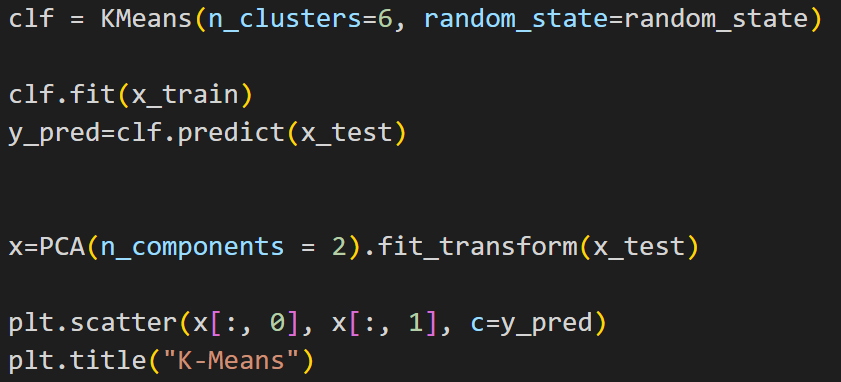
* 使用kmeans，dbscan，层次聚类，AP聚类算法，分别设置不同的参数，分别对两组数据进行聚类，比较结果的差别；
* 分析聚类获得的结果，请使用scikitlearn 中提供的聚类算法测度来比较不同聚类算法在数据集上性能的差异。

# 第二部分 实验过程描述

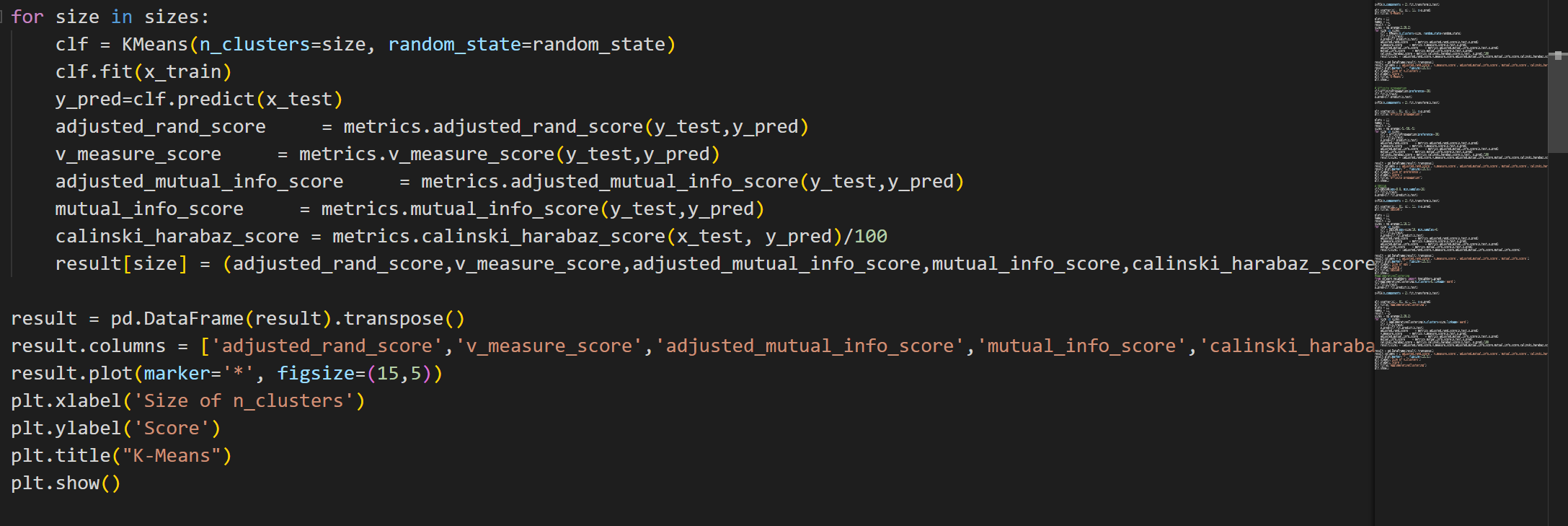
1. 读取数据



1. 聚类结果输出



1. 测度分数输出



# 第三部分 结论与分析

聚类结果输出：

分别选取了对应的数据的类别方便观察：glass为6，iris为3

测度说明：

**adjusted\_rand\_score：**

Rand index adjusted for chance.真实的群集标签和由聚类算法预测出的群集标签之间，相互调整至与另一方完全相同所需要的代价。相似比分在-1.0 和1.0 之间，1.0代表完美匹配。

**v\_measure\_score：**

V-measure cluster labeling given a ground truth. 是均匀性和完整性之间的调和平均值。

* **均匀**性: 每个群集只包含单个类的成员。
* **完整性**: 给定类的所有成员都分配给同一群集。
* 比分在0.0 和1.0 之间。1.0 代表均匀完整的标签

**adjusted\_mutual\_info\_score：**

Adjusted Mutual Information between two clusterings.

和**adjusted\_rand\_score**类似,衡量两个数据分布的吻合程度

当两个标签相同 (即完全匹配) 时, AMI 返回值1。

**mutual\_info\_score：**

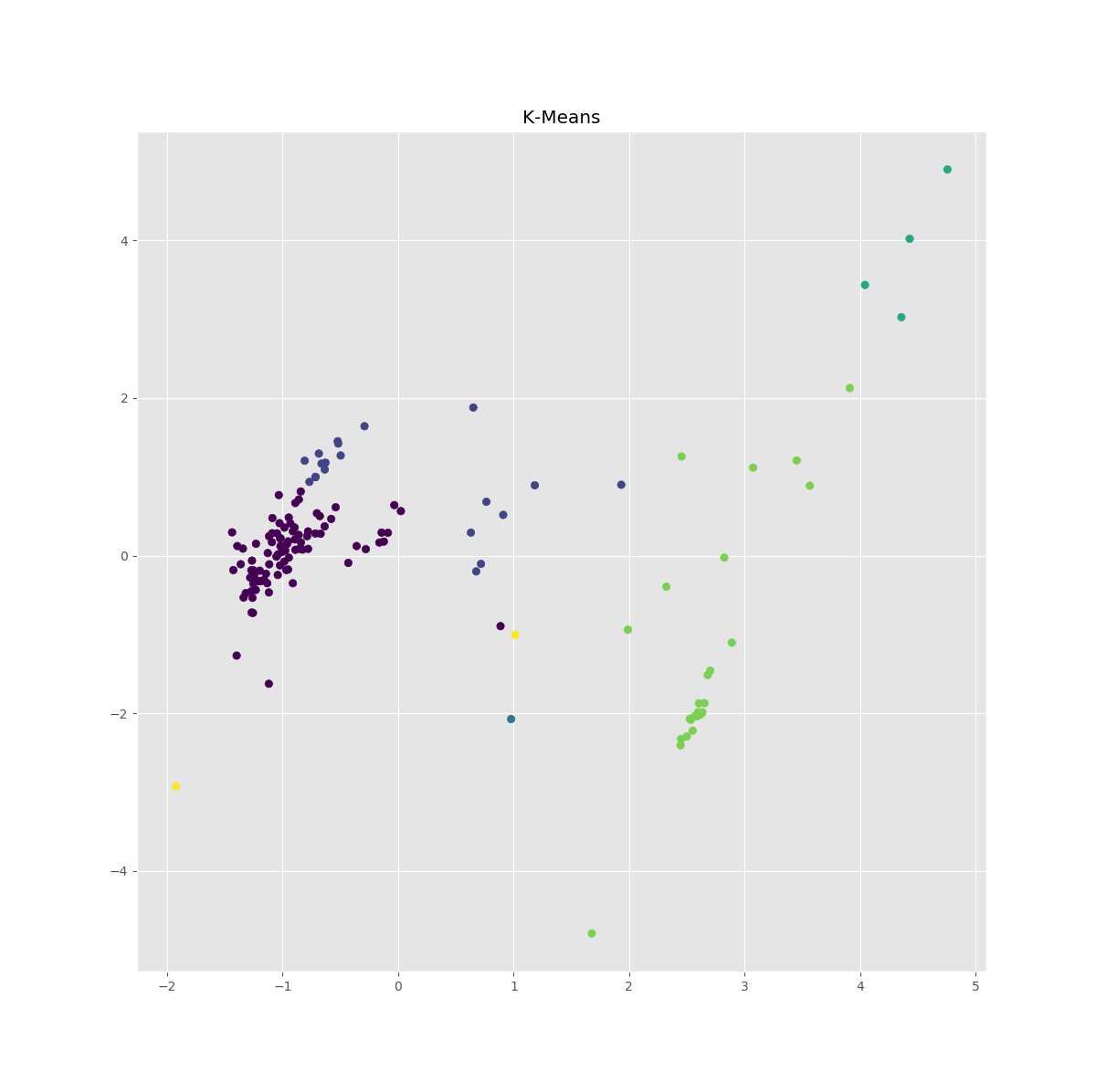
是对同一数据的两个标签之间的相似性的度量

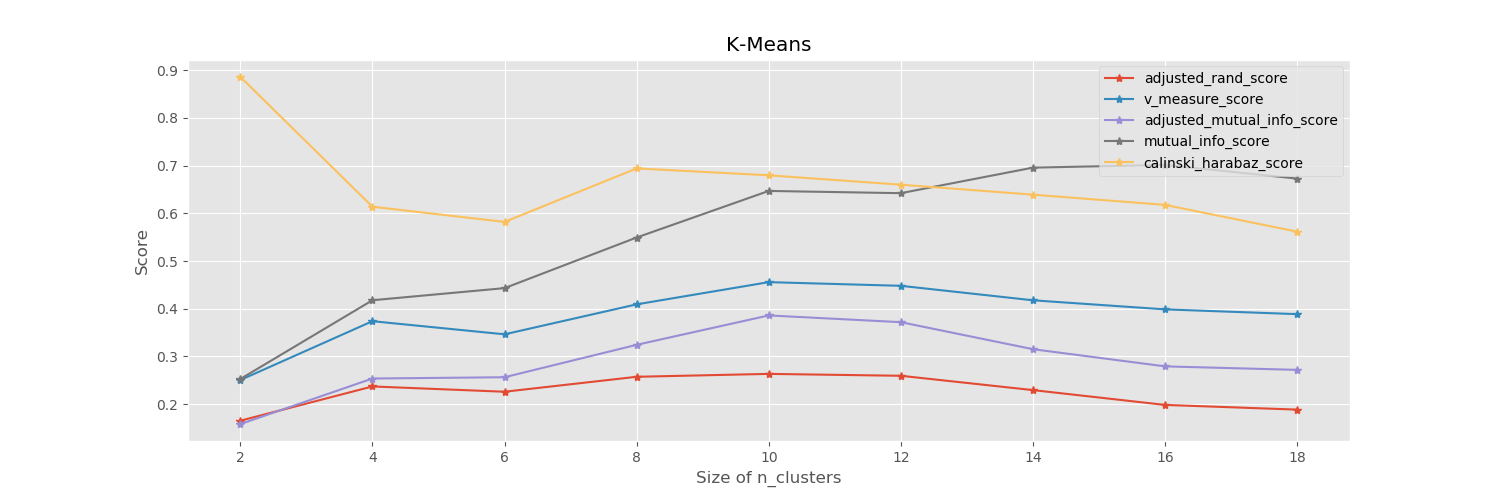
**calinski\_harabaz\_score：**

得到的Calinski-Harabasz分数值ss越大则聚类效果越好，因为数值较大，这里可视化时分别/100与/500

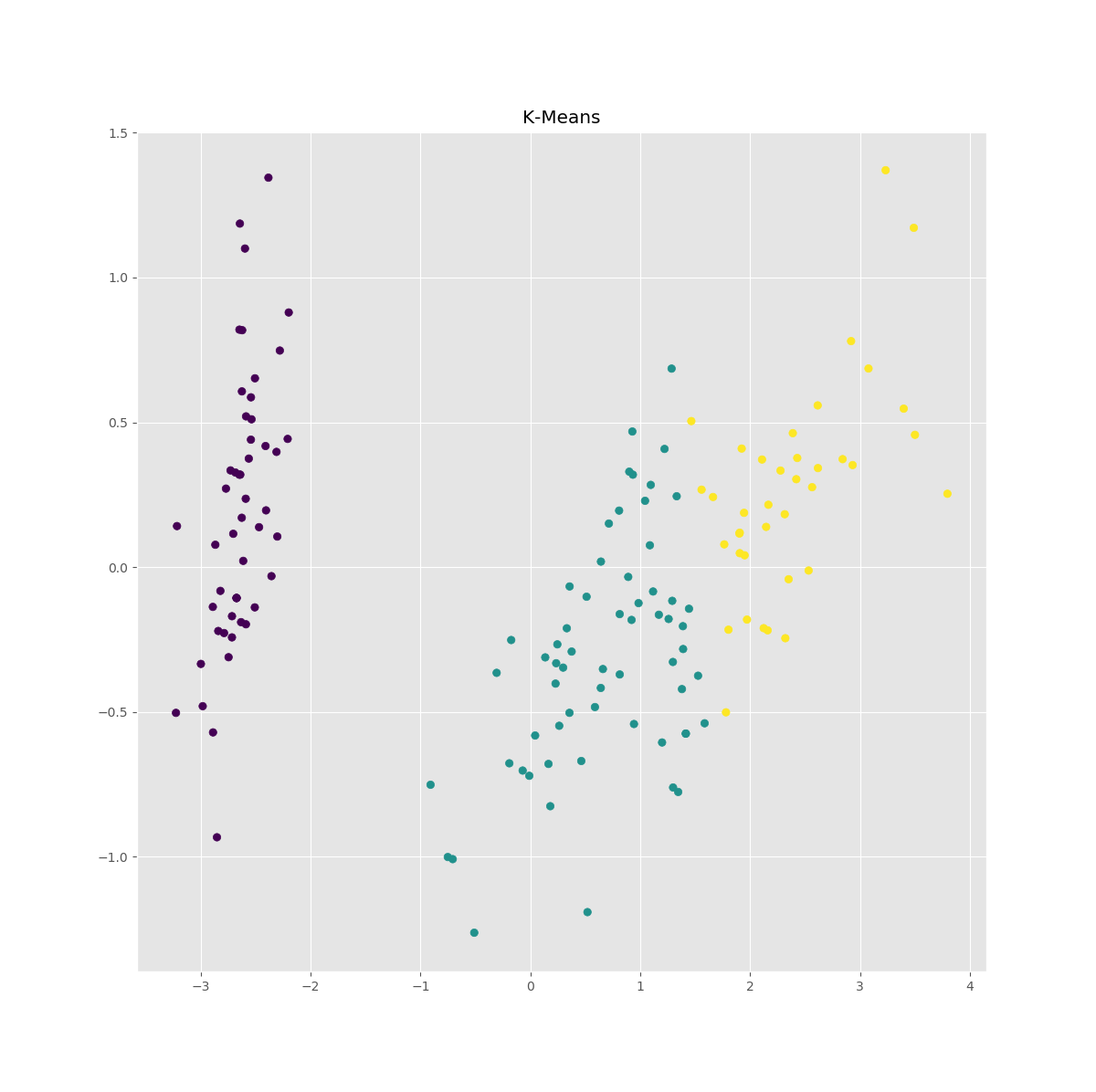
1. K-means

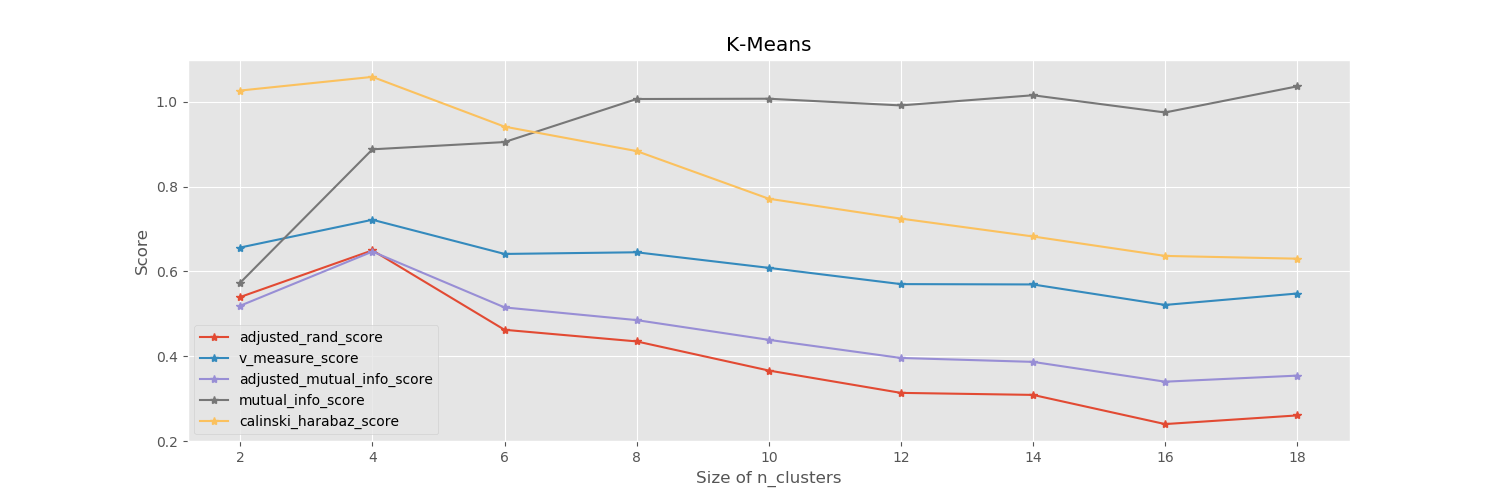
Glass：





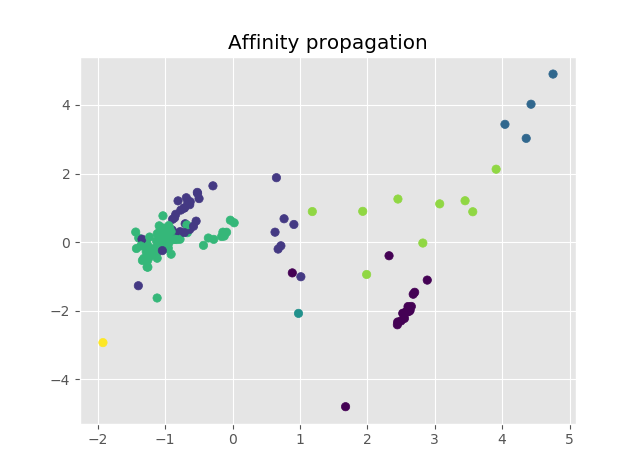
Iris：

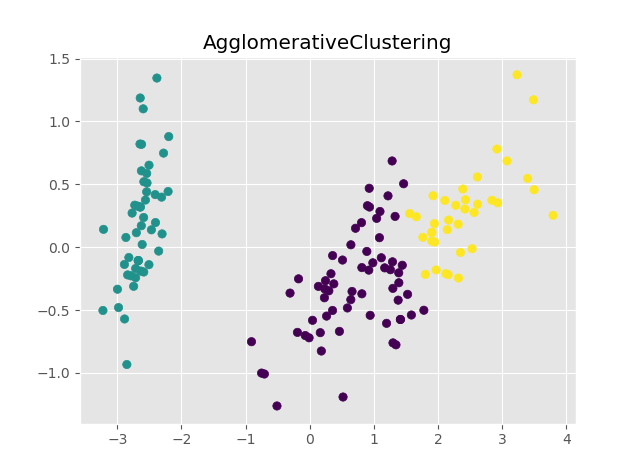


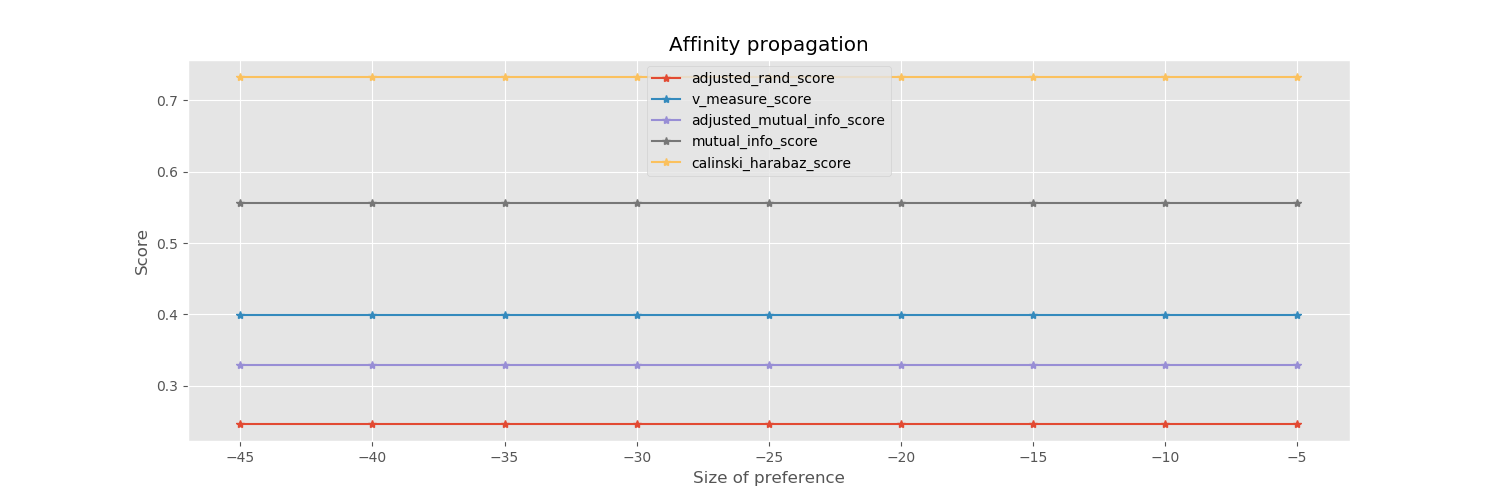


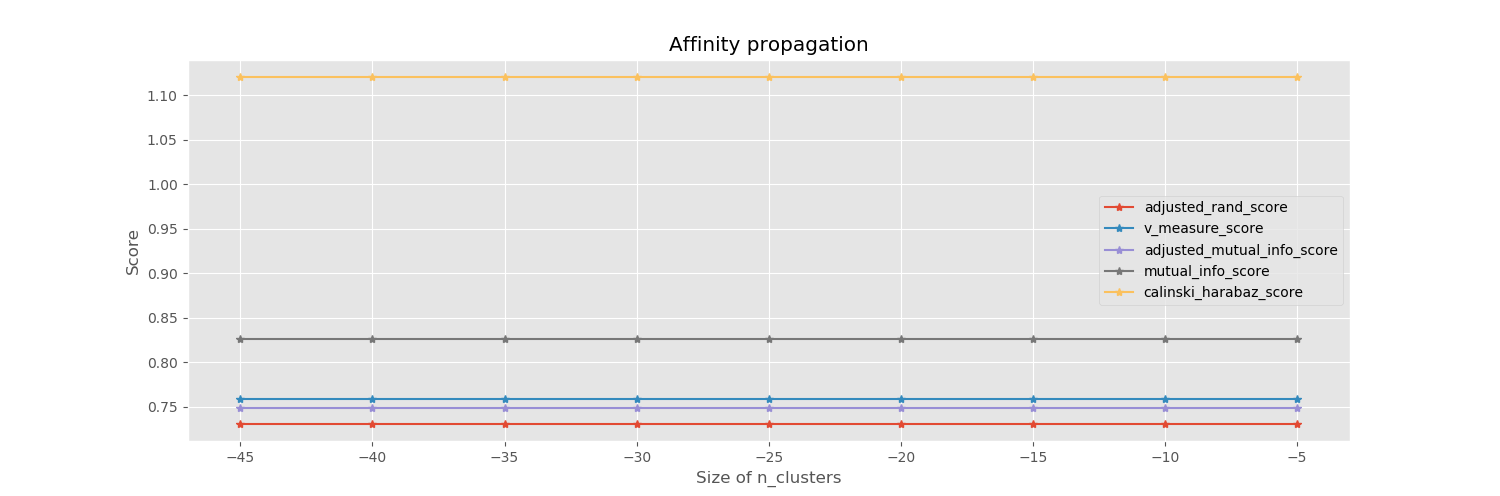
K-Means适用于均匀的簇大小，在iris的数据中，各个簇的大小总体是比较接近，因此有良好的分类效果，同时在避免了没有必要的分类后，在n\_clusters为4这个点上，K-means也有较为良好的表现。而glass的簇大小并不均匀，所以其效果较差。通过PCA降为后的数据，依然没有较为明显的分开，glass数据的特征不够明显。

1. Affinity propagation



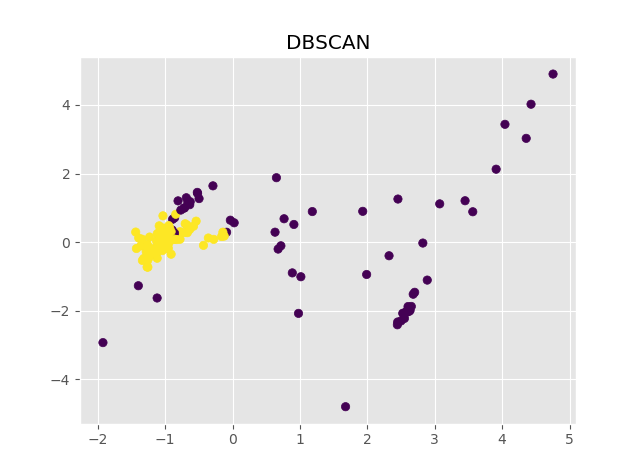


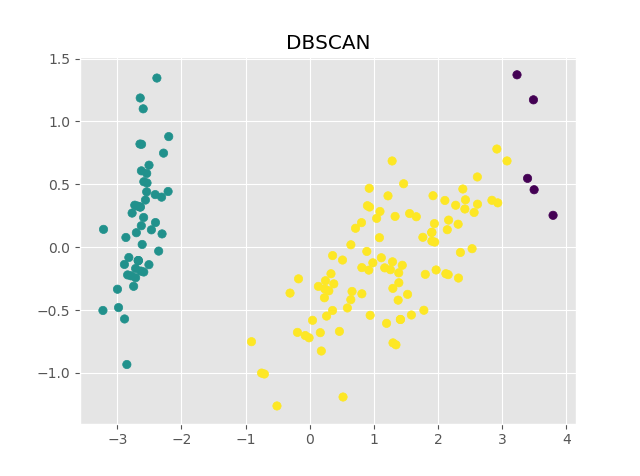


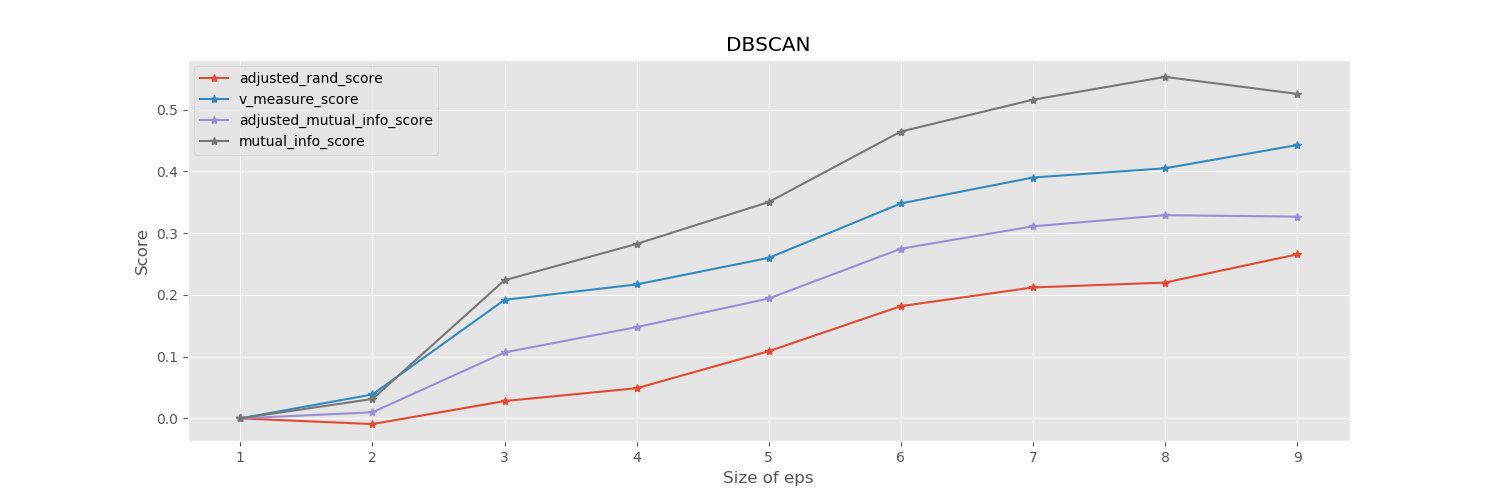


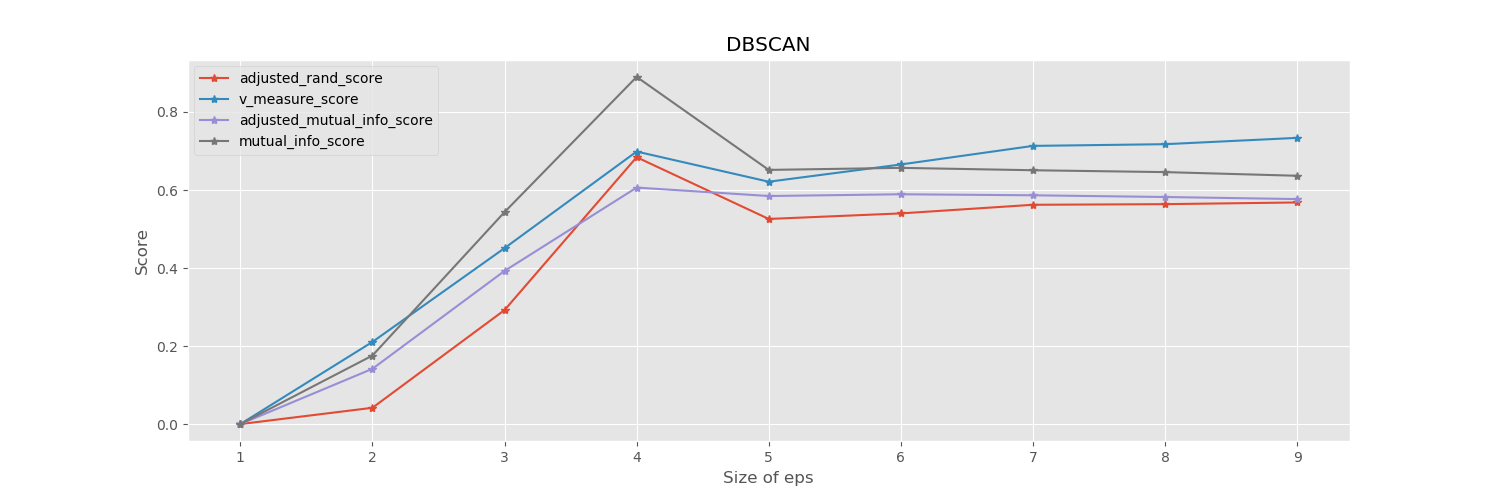
无法指定聚类的数量，参数preference决定了聚类数量的多少，值越大聚类数量越多，在preference较小范围的变化上没有太大的改变，同样对于簇较为均匀且特征明显的数据有着良好表现（iris）。

1. DBSCAN



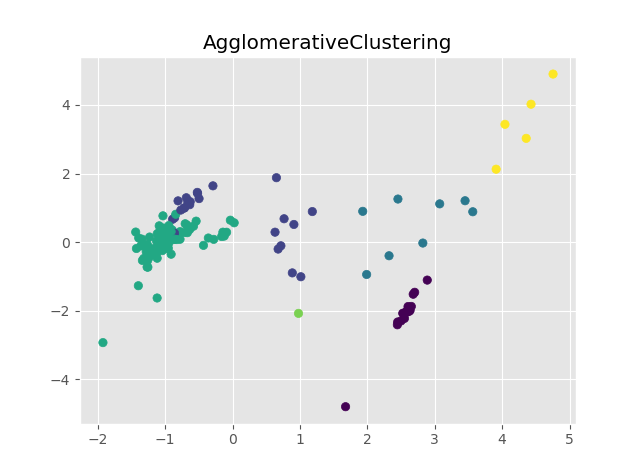


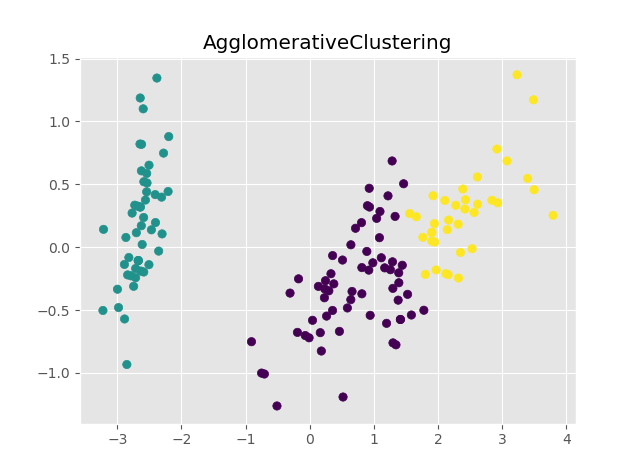


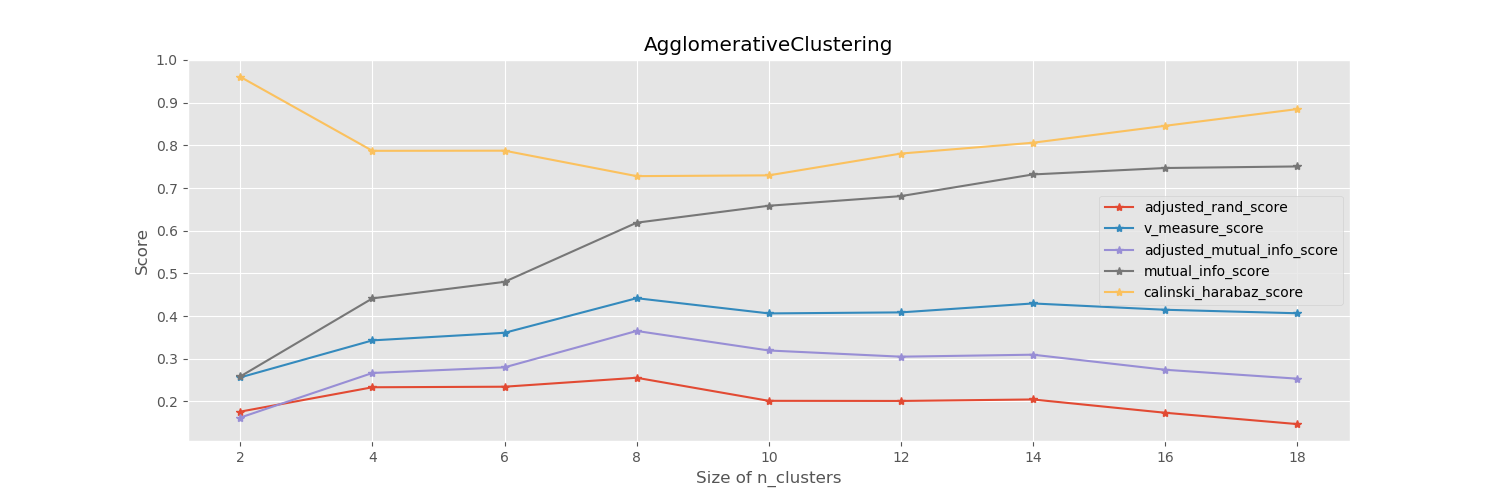


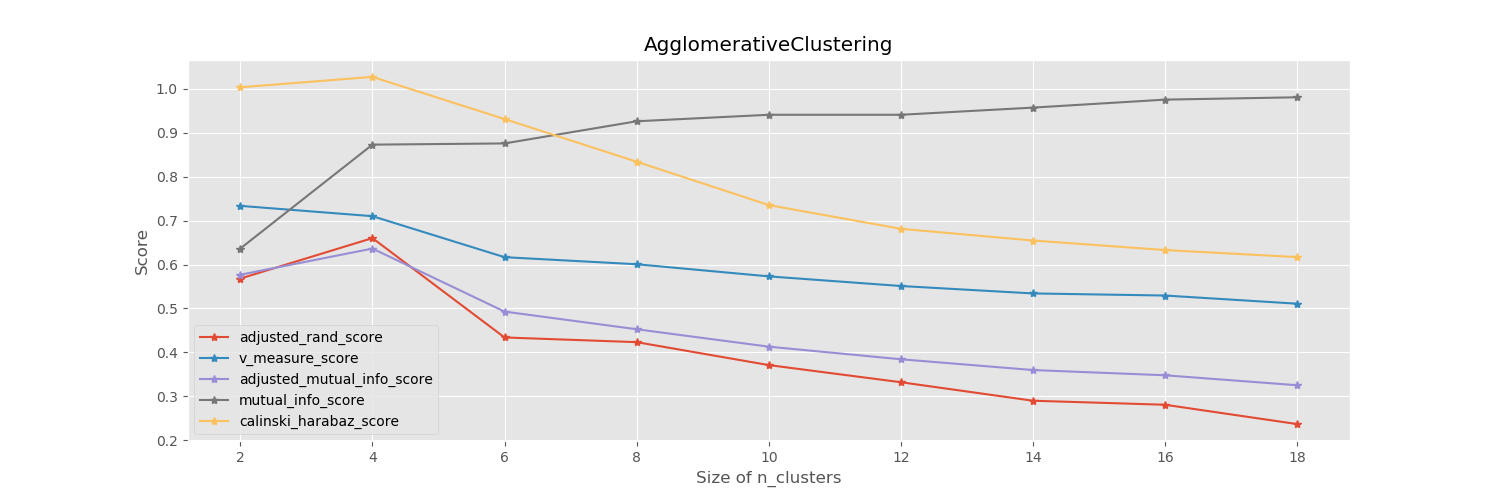
DBSCAN需要设置的两个参数，eps和min\_samples完全决定了聚类的分类效果，这两个参数需要经过大量的调试才能得到和其它聚类算法相同的结果，否则在表现上相对于其它算法差强人意。

1. Agglomerative Clustering









分层聚类总体上比K-means表现优秀，无论是iris还是glass数据集都有更高的分数和更好的效果。

# 第四部分 相关代码

