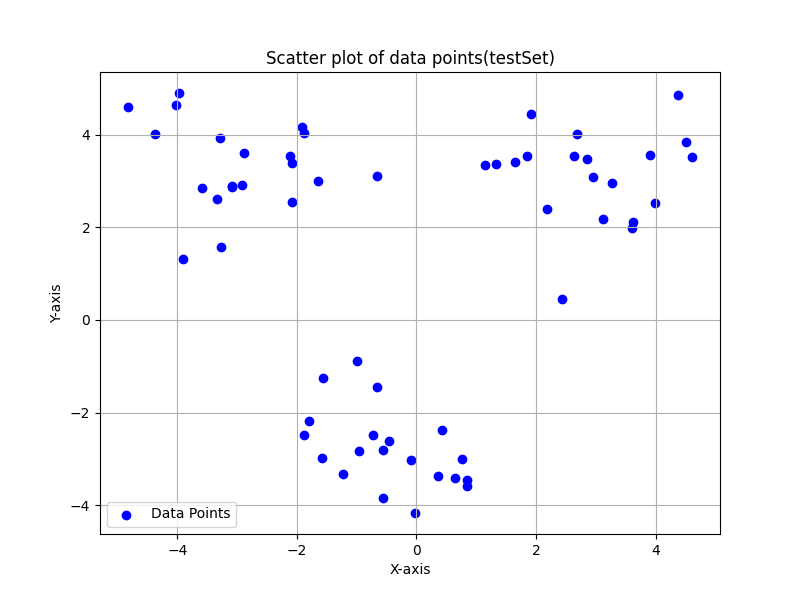
聚类算法比较

# 四种聚类算法简介

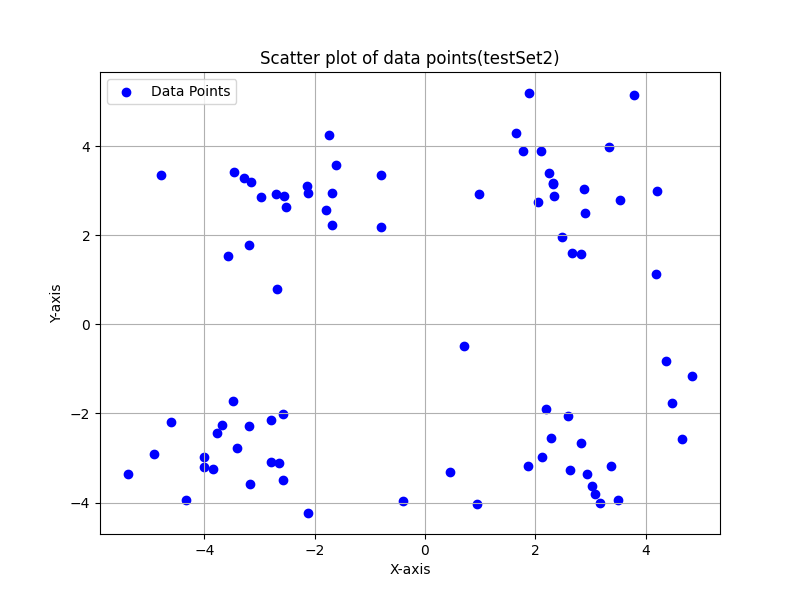
1. **K-means 聚类**：这是经典的划分算法，适用于球状的簇，算法依赖于用户指定的簇数。
2. **自组织映射神经网络（SOM）**：这种算法使用神经网络的方式进行无监督学习，将高维数据映射到低维空间，适合处理复杂数据结构。
3. **AP 聚类（Affinity Propagation）**：这是一种基于“消息传递”的算法，不需要预设簇的数量，能够根据数据点之间的相似度自动确定簇数。
4. **密度聚类（DBSCAN）**：该算法能发现任意形状的簇，适合处理带有噪声的数据集。

# 实验过程

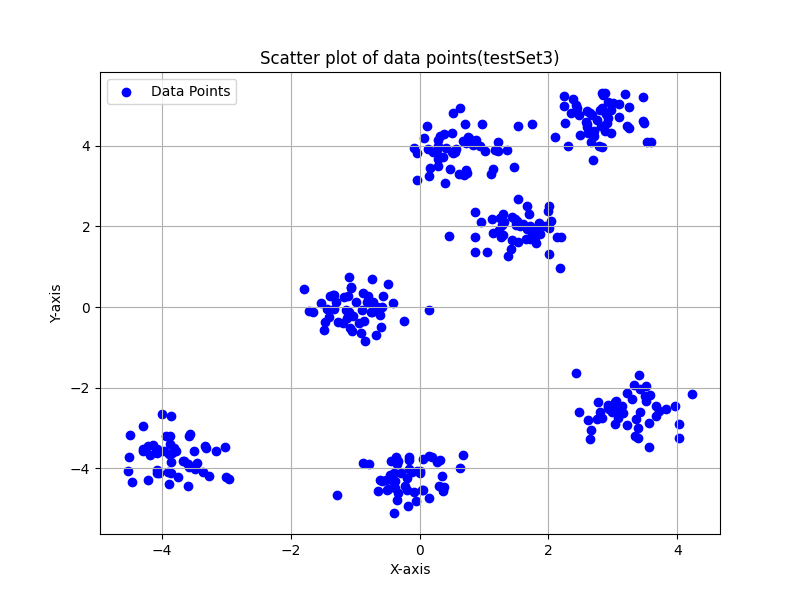
我们使用三个随机生成的测试集，测试数据如下图所示：



图表 1测试集一



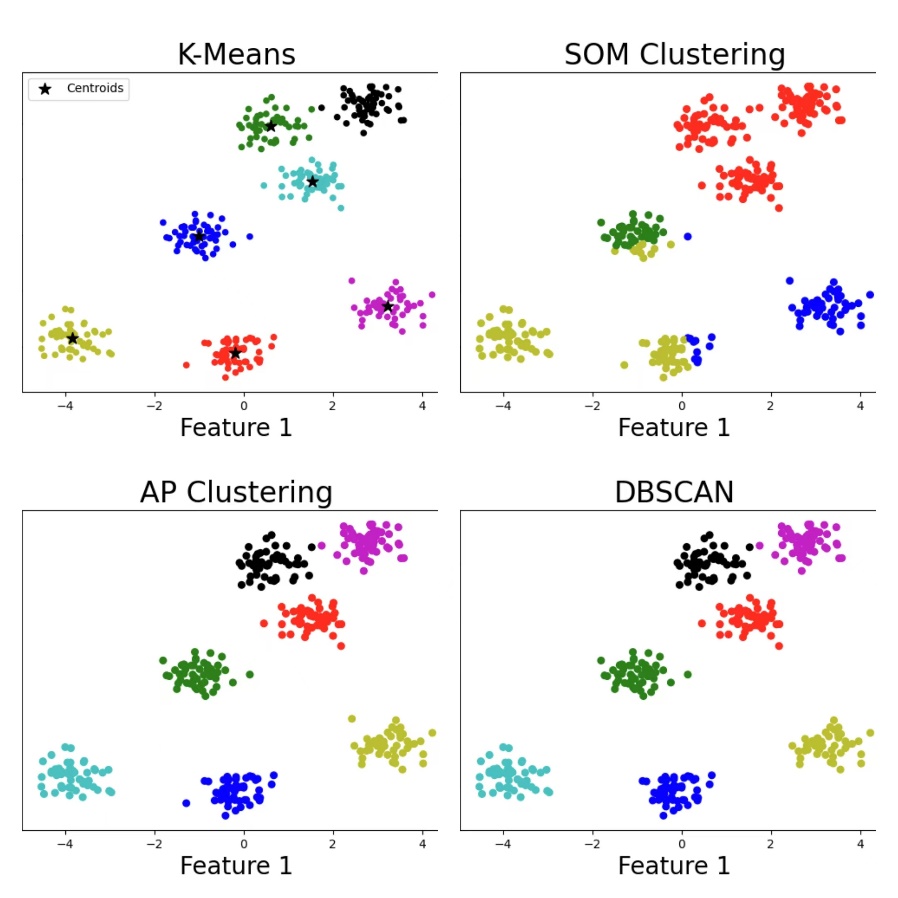
图表 2测试集二



图表 3测试集三

通过sklearn中的算法实现四种聚类代码，将聚类结果可视化并计算他们的DB指数，结果如下图所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 聚类算法 | TestSet.txt(3类) | TestSet2.txt(4类) | TestSet3.txt(7类) |
| k-means | 0.3982 | 0.4294 | 0.3438 |
| SOM | 0.6464 | 0.4253 | 0.5776 |
| AP | 0.3982 | 0.4253 | 0.3438 |
| DBSCAN(eps=1.5,min\_samples=4) | 0.3943 | 0.4202 | 0.3505 |



图表 4四种不同算法在测试集三下的聚类结果

# 结果分析

综合上面的数据，我们发现k-means算法综合最优，其次为AP算法，SOM网络表现较差，密度聚类在合适的参数上表现也很不错。

K-means假设数据是球形的，并且每个聚类的大小相似，所以对于聚类中心分布较均匀、具有明显类间分隔的数据集，K-means会有较好的表现。由于我们测试集的生成过程是按照随机生成几个中心点，然后再在中心点附近随机生成测试点，因此k-means表现不错。

SOM表现一般，主要是因为SOM是一种无监督学习算法，它保持数据的拓扑结构，但它并不总是最适合低维空间上的数据，尤其是当类的数量较少时，它的表现可能不如K-means和AP。但它不总是适合低维数据，尤其是类内和类间差距不显著的数据集。

AP不需要提前指定聚类的数量，自动选择聚类中心。这可能使它在某些情况下表现得比K-means更好，尤其是对于不规则的聚类结构。

DBSCAN擅长处理形状不规则的聚类和噪声，但在类间分布差异较大的情况下，DBSCAN可能表现得不如K-means。DBSCAN对噪声点的处理较好，因此在包含噪声的数据集上，DBSCAN的效果可能优于其他算法。但是DBSCAN得性能取决于参数eps 和 min\_samples。在特定参数下，它能接近K-means的效果。