一、聚类评价指标简介

（1）戴维森堡丁指数（Davies-Bouldin Index，DBI）

戴维森堡丁指数（Davies-Bouldin Index, DBI）是一种常用的用于评估聚类质量的指标。它通过比较聚类内的紧密性（即簇内点之间的相似度）和聚类之间的分离性（即簇与簇之间的差异）来衡量聚类的效果。

给定一个聚类结果，将数据分成 K 个簇（例如，簇 $C_1, C_2, \ldots, C_k$）。对于每个簇 C_i，可以计算该簇的“紧密性”：

$$S_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} d(x, \mu_i) \quad (1)$$

$$M_{i,j} = d\left(\mu_i, \mu_j\right) \quad (2)$$

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (3)$$

$$R_i = \max_{j \neq i} R_{i,j} \quad (4)$$

$$DBI = \frac{1}{K} \sum_{i=1}^{K} R_i \quad (5)$$

给定一个聚类结果，将数据分成 K 个簇（例如，簇 $C_1, C_2, \ldots, C_k$）。在（1）式中，计算簇内各点的均值向量作为簇的中心，计算簇内紧密性，定义为簇中说有点到簇中心的平均距离。其中，$|C_i|$ 是簇 C_i 中的点数，$d(x, \mu_i)$ 是点 x 与簇中心 μ_i 之间的距离。在（2）式中，主要是计算簇间分离性，定义为两个簇的中心之间的距离，其中，μ_i 和 μ_j 分别是簇 C_i 和 C_j 的中心。在（3）式中，对于每个簇 C_i，计算其与其他所有簇的相似性分数 $R_{i,j}$，其中，S_i 和 S_j 分别是簇 C_i 和 C_j 的紧密性，$M_{i,j}$ 是簇 C_i 和簇 C_j 之间的距离。在（4）式中，取簇对于其他簇的最大相似性分数作为该簇的分数。在（5）式中，将所有簇的分数平均得到整个聚类结果的戴维森堡丁指数。

DBI 越低，表示聚类的效果越好，因为低 DBI 意味着簇内紧密且簇间分离较大。DBI 越高，表示聚类效果较差，可能是因为簇之间不够分离，或是簇内分布较为分散。

（2）邓恩系数（Dunn Validity Index）

邓恩指数（Dunn Validity Index）是用于评价聚类质量的一个指标，它通过衡量类内紧密度和类间分离度来判断聚类的优劣。邓恩指数的目的是获得较小的类内距离（即每个聚类中的数据点彼此间距离较小）和较大的类间距离（即不同聚类之间的距离较大），从而得到更加分离和紧凑的聚类结果。

假设聚类结果为 $C = C_1, C_2, \ldots, C_k$，其中 k 是聚类的数量。邓恩指数的定义为：

$$D = \frac{\min_{1 \le i \le j \le k} \delta\left(C_i, C_j\right)}{\max_{1 \le l \le k} \Delta(C_l)}$$

其中，$\delta(C_i, C_j)$ 表示聚类 C_i 和 C_j 之间的类间距离。常见的类间距离定义包括最小距离（单链法）、最大距离（全链法）和质心距离（均值法）等。$\Delta(C_l)$ 表示聚类 C_l 的类内距离，通常计算该聚类中所有点之间的最大距离，或者计算点到质心的平均距离。

邓恩指数越大，说明聚类结果的质量越高，即类内距离小、类间距离大，聚类效果更好。邓恩指数对于聚类的数量和类别形状有一定的敏感性。

（3）方差比准则（Variance Ratio Criterion, VRC）

方差比准则（Variance Ratio Criterion, VRC）是聚类质量评估的一个经典指标，用于衡量聚类后的类内紧凑度和类间分离度的相对情况。它通过比较类内方差和类间方差来判断聚类结果的合理性。VRC 主要用于评估聚类的结果是否有较好的聚类效果。

对每个簇 C_i 计算该簇的均值 μ_i，然后计算 C_i 中样本数量 n_i 与该簇均值与总体均值 μ 的平方差，并将这些平方差加总，得到类间离差平方和（BSS）:

$$BSS = \sum_{i=1}^{k} (n_i \cdot \|\mu_i - \mu\|^2)$$

对每个簇 C_i 计算该簇内所有样本与簇均值 μ_i 的平方差之和，并将所有簇的类内平方差加总得到类内离差平方和（WSS）:

$$WSS = \sum_{i=1}^{k} \sum_{x \epsilon C_i} (\|x - \mu_i\|^2)$$

通过将 BSS 和 WSS 带入方差比准则公式，计算 VRC 值:

$$VRC = \frac{\mathrm{BSS}/(\mathrm{k}-1)}{\mathrm{WSS}/(\mathrm{n}-\mathrm{k})}$$

VRC 值越高，表示类间距离大，类内紧凑度高，即聚类效果更好。VRC 值越低，表示类间分离度差，类内紧凑度低，即聚类效果较差。

（4）三种聚类质量评定指标的对比

指标	DBI 指数（Davies-Bouldin Index）	邓恩指数（Dunn Index）	方差比准则（Calinski-Harabasz Index）
定义	衡量簇间分离度与簇内紧密度的比值	衡量簇间的最小分离度与簇内最大紧密度的比值	衡量簇间方差与簇内方差的比值
计算方法	计算簇内散度与簇间距离的比值，取最大值	计算簇间最小距离和簇内最大距离的比值	计算簇间离散度与簇内离散度的比值
值的范围	越小越好，最小值为 0	越大越好	越大越好，通常为正值
优点	计算简单	能有效区分不同簇，特别适合不规则簇形	强调整体聚类的分离度和紧密度，适用于大规模聚类
缺点	对噪声敏感，可能无法准确反映复杂簇形态的	计算复杂，且对簇形状的偏差较敏感	假设簇间方差差异大，对异常值敏感

	质量		
适用情况	簇形状规则且相对均匀的情况	形状不规则且密度差异大的簇	簇间差异明显且簇内紧密的情况
对簇形状的要求	假设簇为均匀球形或圆形	无需特定形状，适应不同簇形	假设簇间有较大方差差异

DBI 指数、邓恩指数和方差比准则是常用的聚类评价指标，各自有不同的优缺点。**DBI 指数**通过衡量簇内紧密度与簇间分离度的比值，计算简单，适用于均匀、球形簇的评估，但对噪声和复杂形状的簇较为敏感，且结果越小越好。**邓恩指数**关注簇内最大紧密度与簇间最小分离度，能够有效区分不规则簇，尤其适合处理形状不均的聚类，缺点是计算复杂且对簇形状的偏差敏感，值越大越好。**方差比准则**则通过比较簇间与簇内方差来评估聚类效果，强调聚类的整体分离度和紧密度，适合大规模聚类，且对异常值较为敏感，值越大越好。总结来说，三者各有侧重，DBI 适合简单的球形簇，邓恩指数适合不规则形状的簇，而方差比准则更关注群体的全局差异。

二、K-means 聚类及其质量评价

K-means 聚类是一种常见的无监督学习算法，旨在通过最小化数据点到其所属簇的距离，将数据分成预定数量的簇。它是基于迭代的思想，逐步优化簇内数据点的聚合性，并尽量使得不同簇之间的数据点尽可能地分离。

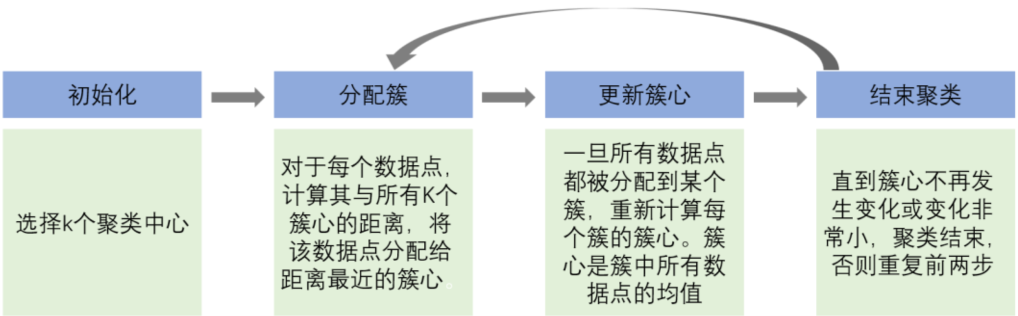

图 1：自制 K-means 步骤图

K-means 聚类算法的优点有两点，第一是计算效率高，K-means 的时间复杂度相对较低，通常为 $O(n \cdot k \cdot t)$，其中 n 是数据点的数量，k 是簇的数量，t 是迭代次数；第二是易于理解和实现：K-means 算法逻辑简单，非常适合用于大规模数据集。缺点主要有三点：第一是对初始簇心敏感，K-means 对初始簇心的位置敏感，不同的初始化可能会得到不同的结果，导致结果的不稳定；第二是无法处理形状复杂的簇，如果数据的簇具有不同的密度或形状，K-means 的聚类效果可能不理想；第三是对噪声和离群点敏感：K-means 对噪声和离群点比较敏感，因为它会试图将所有点分配到某个簇。

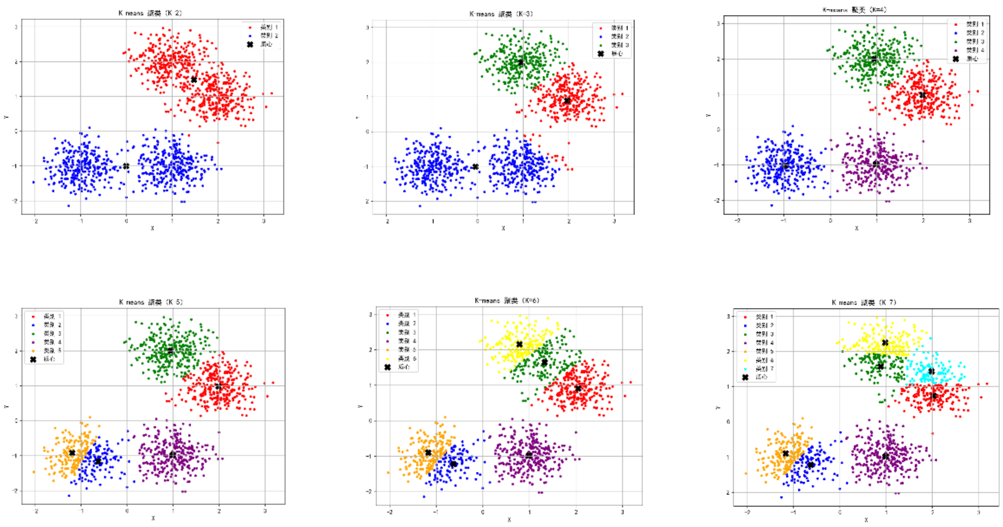

图 2-7：K-means 不同 K 数值情况下的聚类分布图

在 K-means 不同的 k 数值情况下，聚类情况有着截然的不同，可以看到，在 k 值较小的情况下，尽管一些点相隔很远，但是仍然被分为一类；而在 k 值较大时，尽管一些点相隔很近，但是仍然被分为不同的类，这种聚类实际上是我们不想看到的。在 K-means 聚类中，常使用手肘法进行 K 值参数选择。

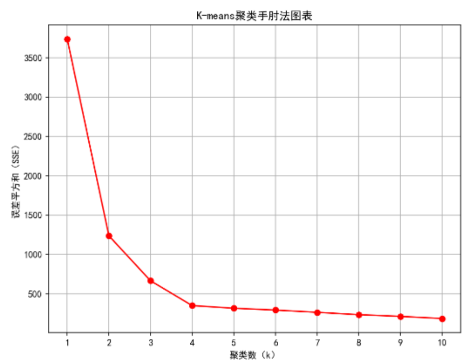

图 8：K-means 聚类手肘法结果

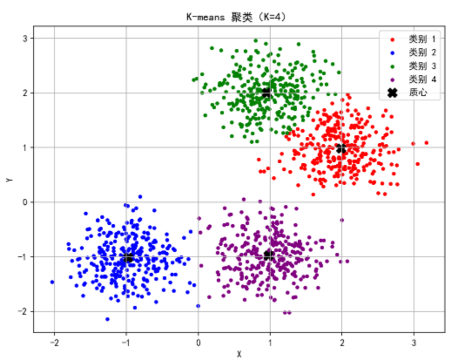

图 9：K-means 聚类（k=4）结果

K-means 算法的手肘法（Elbow Method）用来确定 K-means 聚类中最佳的簇数 K。它的基本思想是通过绘制不同 K 值对应的误差平方和（Within-Cluster Sum of Squares, WSS 或 Inertia），观察 K 值与误差之间的关系，从而确定最合适的簇数。

根据 K-means 聚类手肘法结果可以发现，误差平方和经历了一个先快速后缓慢的递减趋势，可以明显看出当 k=4 时，出现了一个明显的拐点，可以据此推断，k=4 即为比较合适的选取值，通过 K-means 聚类结果（k=4）也可以看到，k=4 时，数据点划分得较好，很符合人的直觉，有着四块较为明显的数据点集。

为了更好地评价聚类质量，我们可以同时参考多个聚类评价指标：

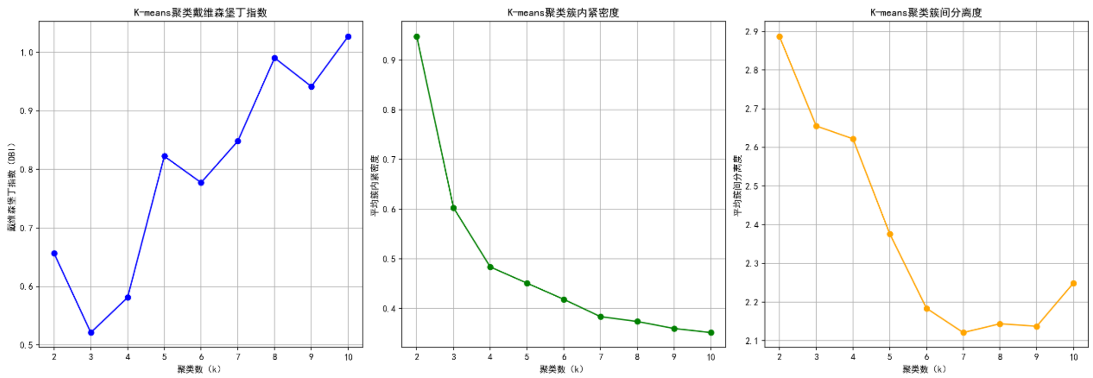

图 10-12：K-means 聚类戴维森堡丁指数、相应平均紧密度、相应平均分离度随 k 值变化图

首先，我们可以看到 K-means 在戴维森堡丁指数评价体系下的情况。就总算出的指数来看，经历了一个先减小后波动上升的过程，在 k=3 和 4 时，戴维森堡丁指数都比较低，说明此时的聚类效果较好。通过簇内紧密度和分离度也可以进一步分析，两者都是在递减（分离度在 k=10 的小幅度上升除外），但是紧密度的下降时从快到慢，而分离度虽然也是从快到慢，但是在 k=3 和 4 之间出现了一个缓降，随后才继续快速下降。聚类的重要标准是要尽可能使得簇之间的紧密度更大，而簇内的分离度更小，因此这也两条曲线在 k=3 和 k=4 的居中表现也直接决定了总的戴维森堡丁指数在 k=3 和 k=4 时保持了一个较好的平衡，因而数值较小，这和手肘法观测的结果是类似的。在前文提及过，戴维森堡丁指数对噪声敏感，适用于簇形状规则且相对均匀的情况，可能对不规则形状的聚类效果不好，但是由于数据点的分布形状较规则，轮廓没有异变，噪声较少，因此该指数的效果也较好。

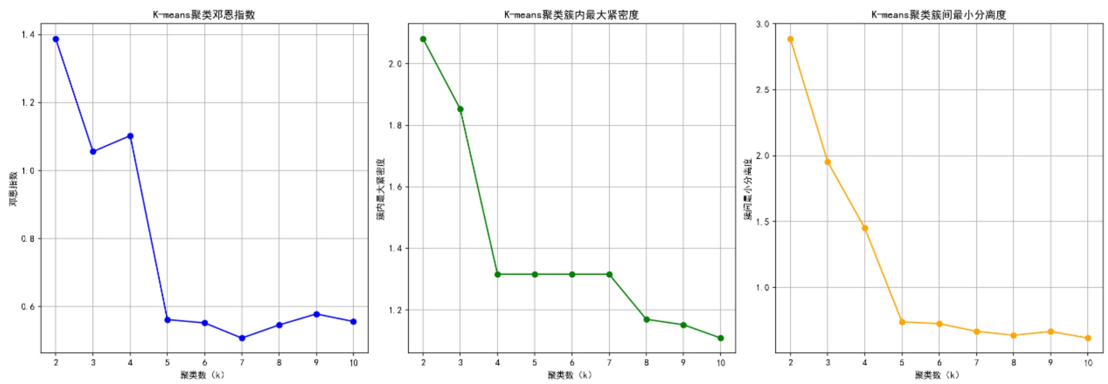

图 13-15：K-means 聚类邓恩指数、相应最大紧密度、相应最小分离度随 k 值变化图

随后，可以看到 K-means 在邓恩指数评价体系下的情况。就总算出的指数来看，经历了一个陡降随后平稳的过程，可以发现，与之前的判断结果不同的是，虽然后续的 k=3 以及 k=4 也在结果中显示为高值，邓恩系数在 k=2 阶段都非常高，结果显示为最佳的聚类。在邓恩系数中，我们仍然可以拆解其簇内最大紧密度和簇间最小分离度进行分析，它们的变化和总指标变化类似，经历了快速下降随后稳定的过程，只不过簇内最大紧密度的下降更快，在 k=4 时就已下降到较低值，而簇间最小分离度则是与总指标一样在 k=5 时才下降到较低值。从前面的描述中，我们可以知道邓恩系数比较适合形状不规则且密度差异大的簇，而这里测试的簇较为规则，且密度差异小，因此其准确度相对于戴维森堡丁指数来说会差一些。

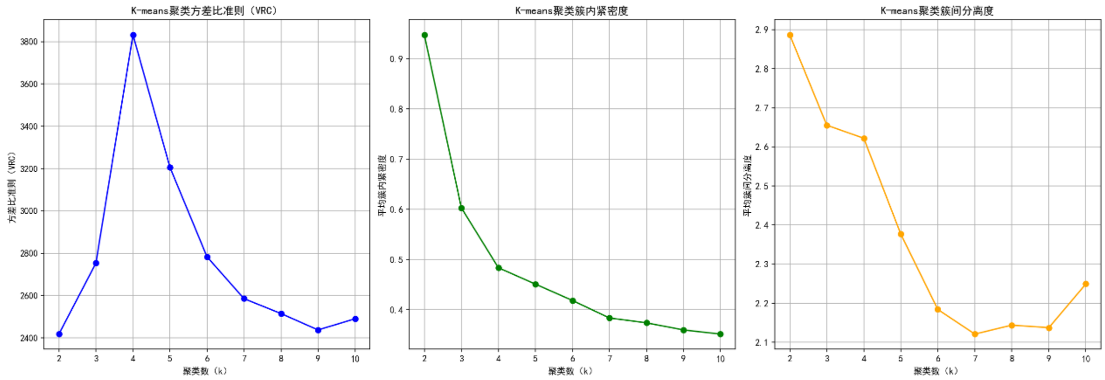

图 16-18：K-means 聚类方差比准则指数、平均簇内紧密度、平均簇间分离度随 k 值变化图

接着，可以看到 K-means 在方差比准则指数下的情况。可以看到就总 VRC 指数来看，呈现从 2 到 4 的快速上升到顶峰，随后快速下降的过程，在 9 至 10 的位置有轻微的翘起。另外，就簇内紧密度来看，整体曲线是下降速度从快到慢，而且变化速度比较均匀；而簇间分离度在 3 至 4 之间出现了曲线突然放缓的情况，随后的趋势是快速下降直至平稳状态。就指数来看，该项指标非常准确地反映了当 k=4 的时候聚类的质量最高，与手肘法的结果类似。VRC 指数比较适合于簇间差异明显且簇内紧密的情况，而我们采用的数据点比较符合这样的情形，因此 VRC 指数比较好地反映了聚类质量。

三、DBSCAN 聚类及其质量评价

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 是一种密度聚类算法，特别适用于发现任意形状的簇，并能有效处理数据中的噪声点。它通过评估样本的密度来判断数据点是否应该被归为同一簇，主要依赖于两个重要参数：eps（邻域半径）和 min_samples（核心点的最小邻域数）。

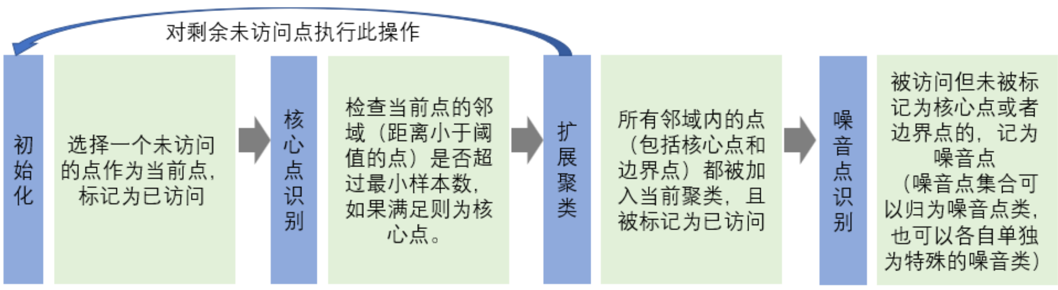

图 19：自制 DBSCAN 步骤图

DBSCAN 的优点在于一是能够识别任意形状的簇，适用于复杂分布的数据，二是不需要指定簇的数量，三是能有效识别数据中的噪声点，不受噪声影响。而其缺点一是在高维数据中，参数 eps 和 min_samples 的选择比较困难，二是在数据密度差异较大的数据集上效果不佳，三是对参数较为敏感，参数选取不当会影响效果。

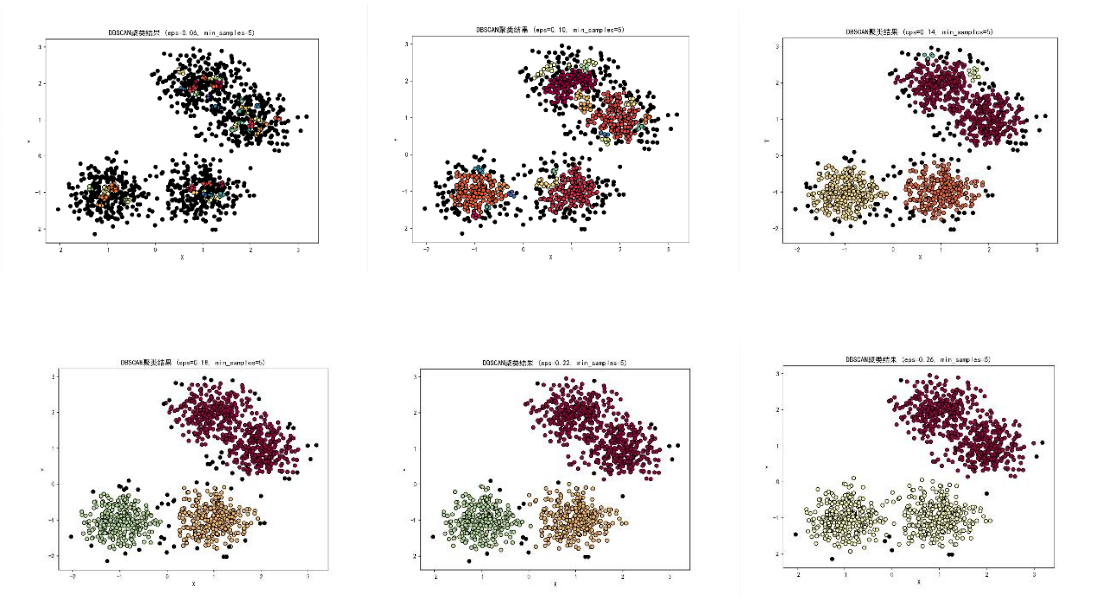

图 20-25：DBSCAN 在 min-samples 取 5 时不同 eps 下的结果

DBSCAN 的参数选择十分重要，这对于聚类的质量影响也非常显著，一般来说，在二维数据图情况下，选择 3 到 5 之内作为 min_samples 都是合适的，可以先拟定一个 min_samples，然后再选择 eps，这个时候 eps 就变成了一个较为关键的一步。可以看到，DBSCAN 在 eps 的不同设置下会出现明显的差异，在 eps 设置较小时，对于核心点的选择变得非常严苛，因此大部分点都会被标记为噪音点，而核心点非常少，不符合一般的聚类目标；而如果 eps 定的太大的话，则是噪音点的选择变得非常严苛，大部分的点都会成为核心点，仍然不是比较满意的结果，因此需要使得 eps 取到一个合适的值，才可以使得聚类质量提升。

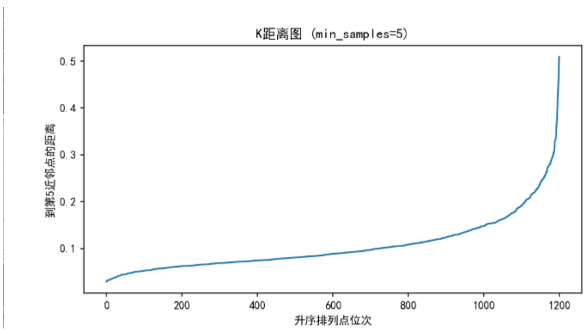

图 26：DBSCAN 在 min-samples 取 5 时的 K 距离图

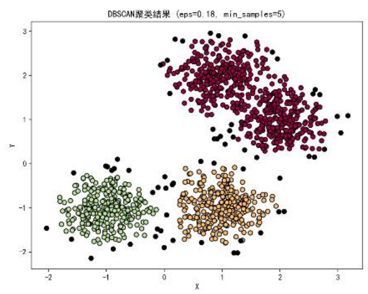

图 27：DBSCAN 较恰当的聚类图

在此，取 min_samples 为 5，进行 eps 的参数选择。DBSCAN 选择 eps 时，常会用到 K 距离图进行参数选取，可以帮助我们找到数据中的密度变化，进而选择合适的 eps，绘制 K 距离图的流程如下：

首先计算最近邻距离，对于每个数据点，计算到其第 min_samples 个最近邻的数据点的距离。随后计算最近邻距离，对于每个数据点，计算到其第 min_samples 个最近邻的数据点的距离。接着排序并绘制，将所有点的这些第 min_samples 个最近邻的距离按升序排序，然后在图上绘制这些距离值，这样生成的曲线就叫 K 距离图。

在 K 距离图中，通常可以观察到一个拐点位置。这个拐点之前，距离变化较小，说明点

的密度较高，而拐点之后，距离开始显著增大，表明点的密度降低。这个位置的距离值就是适合的 eps 初始选择，因为它能帮助 DBSCAN 将点密集的区域划分为簇，同时隔离密度较低的噪声点。

在我们生成的 K 距离图中，可以看到一条逐渐上升的曲线，其中的拐点时在 0.18 左右，因此可以大概确认 0.18 就是比较合适的 eps 值。

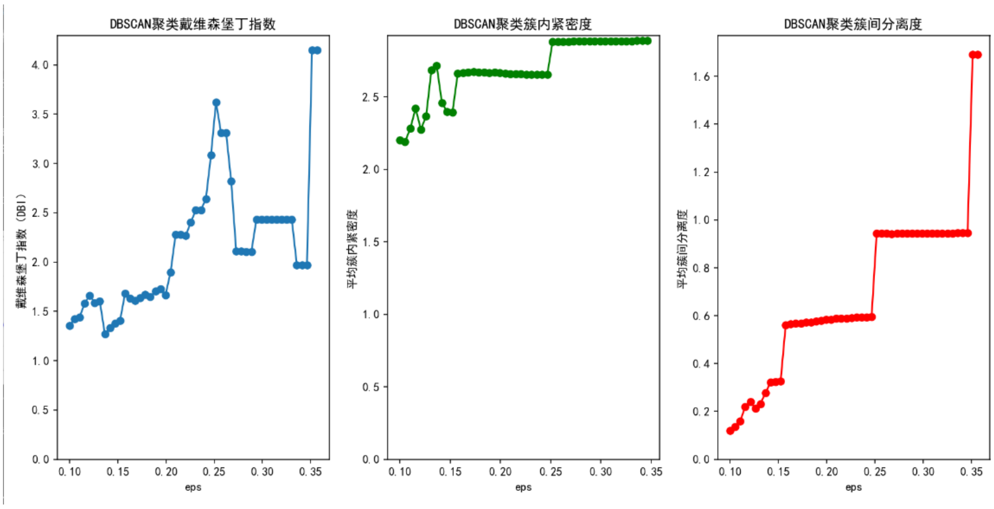

图 28-30：DBSCAN 聚类戴维森堡丁指数、相应平均紧密度、相应平均分离度随 k 值变化图

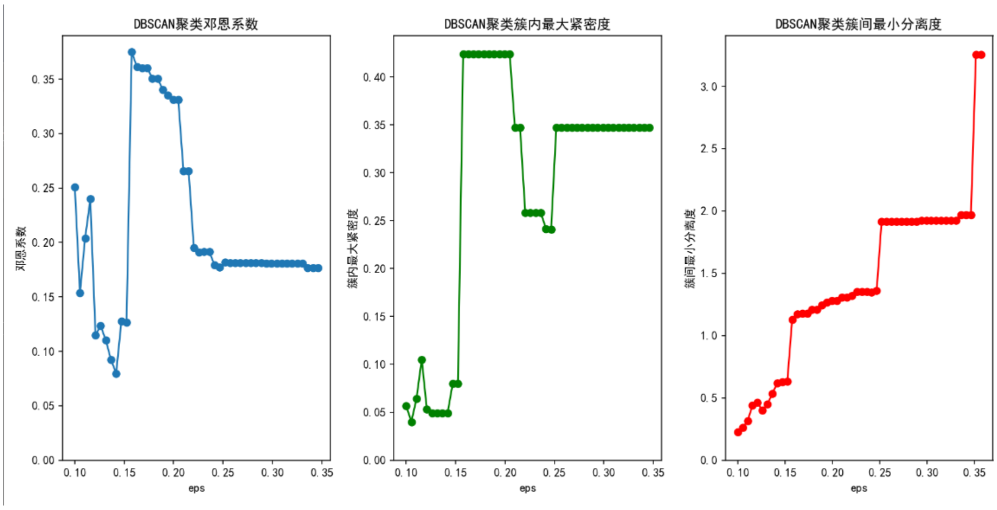

图 31-33：DBSCAN 聚类邓恩指数、相应最大紧密度、相应最小分离度随 k 值变化图

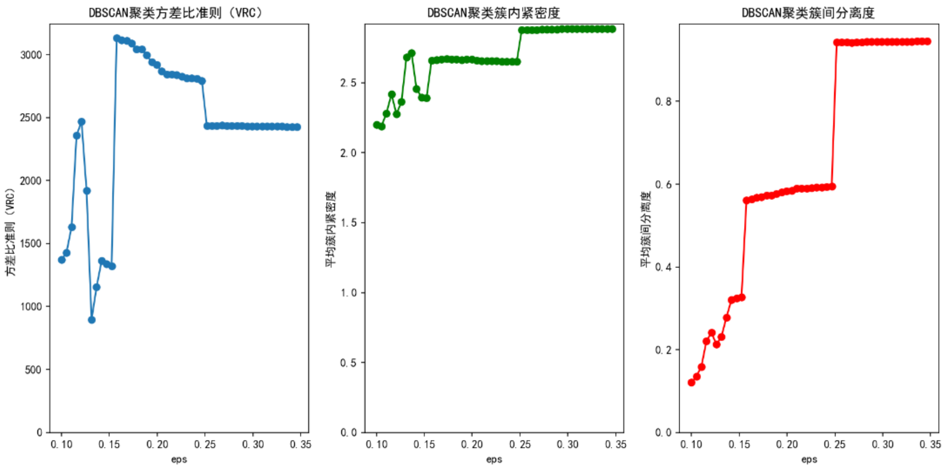

图 34-36：DBSCAN 聚类方差比准则指数、平均簇内紧密度、平均簇间分离度随 k 值变化图

首先，可以看到 DBSCAN 聚类在戴维森堡丁指数评价下的情况。就总指标趋势来看，呈现先低值分布，随后突增高，随后突降，在经历平稳期之后再次突增高。整体的最低值在 0.10-0.20 之间。此外看簇内紧密度，可以发现簇内紧密度经历了不稳定上升后经历一大段的平稳期，而看簇间分离度，可以发现簇间紧密度则是经历了不断的突增和平稳。可以看到戴维森堡丁指数还是比较准确的，我们选取的 0.18 也在其低值范围内。

另外，可以看到 DBSCAN 聚类在邓恩指数评价下的情况。就总指标趋势来看，呈现不稳定下降，随后快速上升，最后再度下降至平稳的趋势。整体的最低值在 0.10-0.18 之间。其中，由邓恩系数给出的簇间最小分离度和戴维森堡丁指数评价下的趋势差不多，但是由邓恩系数给出的簇内最大紧密度和戴维森堡丁指数还是有显著区别的，先是快速上升，随后才进入平稳期，这实际上反映了簇内紧密度虽然平均值变化不大，但是最大值变化异常显著，这甚至直接影响了邓恩系数总的分布变化。但是就之前给出的 K 距离图来看，邓恩指数评价认为最佳的 eps 数应当比 0.18 要稍微低一些，之前在 K-means 中也有提到邓恩系数比较适合形状不规则且密度差异大的簇，而这里测试的簇较为规则，且密度差异小，因此其准确度相对于戴维森堡丁指数来说会差一些。

最后，可以看到 DBSCAN 聚类在方差比准则指数评价下的情况。就总指标趋势来看，和邓恩系数相似，只不过在后半段的下降稍显缓和，其中簇内紧密度的趋势和簇间分离度的趋势实际上都和 DBI 指数相似，可以看到最低值也是集中在 0.10-0.18 之间，这里实际上也和之前的预测情况有一些偏差。

四、层次聚类及其质量评价

层次聚类（Hierarchical Clustering）是一种常见的聚类分析方法，旨在将数据集划分为多个层次结构的簇。其基本思想是通过计算样本之间的相似度或距离，将相似的数据点逐步合并（自下而上）或分割（自上而下）成簇，最终形成一个树形结构（称为**树状图**，Dendrogram）。层次聚类的特点是可以有效地展示不同数据点之间的层次关系。

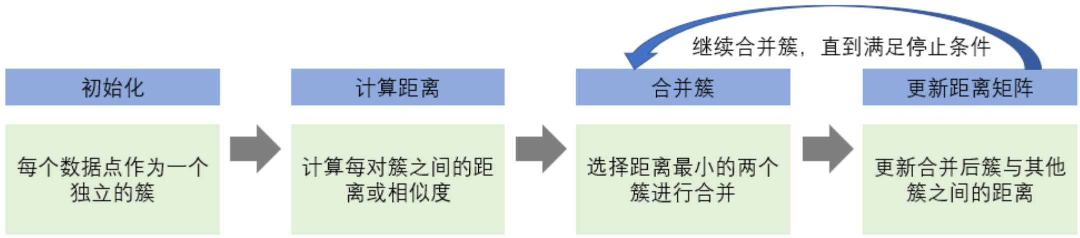

图 37：自制层次聚类步骤图

层次聚类的优点在于不需要事先指定簇的数量，结果呈树状结构，便于理解和分析，缺点在于计算复杂度较高，尤其是在大数据集上，可能需要较长的时间和较多的存储空间，同时对噪声和离群点敏感。

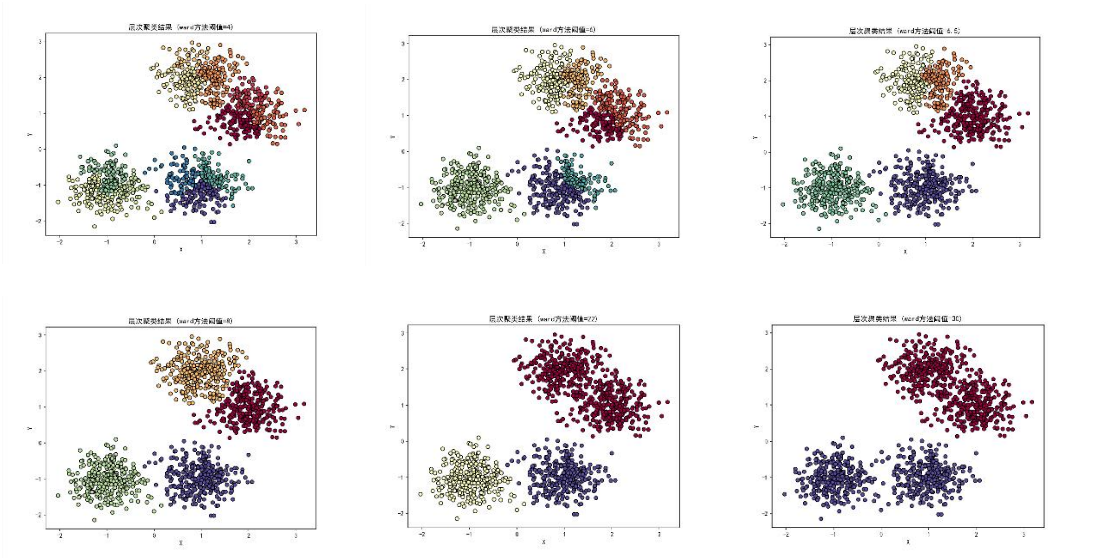

图 38-43：层次聚类在不同 ward 方法阈值情况下的结果

在层次聚类中，对于每对簇，需要计算它们之间的距离，通常，簇之间的距离有不同的计算方式：第一是单链接，计算簇之间的最小距离，即最接近的两个点之间的距离。第二是完全链接，计算簇之间的最大距离，即最远的两个点之间的距离。第三是平均链接（Average Linkage），计算簇之间所有点对的平均距离。第四是 Ward 方法，通过合并簇时最小化簇内方差来度量簇之间的距离。

Ward 方法能生成较为紧密和均衡的簇，这使得它在很多应用中是一个较为理想的聚类方法，尤其适用于数据本身有一定的结构，且希望簇内的点比较集中。单链接容易形成链状结构，全链接容易受到异常点的影响，而 Ward 方法通过最小化簇内的方差来解决这些问题，保证了簇的紧密度和分离度。因此，在本实验中，采用 ward 方法进行，可以看到，层次聚类在不同 ward 方法阈值情况下显示了截然不同的结果，在 ward 方法阈值较小时，很多排列紧密的点也被强行分为了不同的簇；而当 ward 方法阈值较大时，聚类数量太小，很多原本很远的点也被强行分到了一起，因此效果也不好，只有在 ward 方法阈值合适时，才可以有比较好的聚类效果。

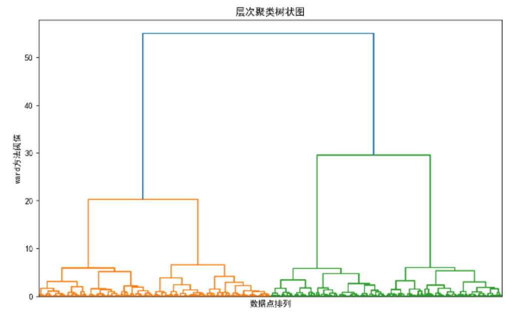

图 44：层次聚类树状图

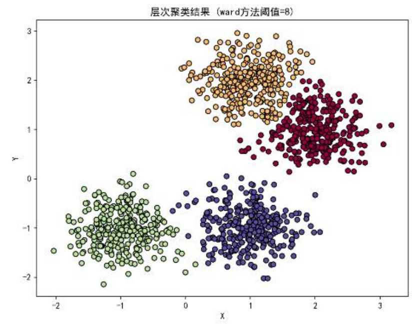

图 45：层次聚类较恰当的聚类图

在进行层次聚类时，常会使用树状图以辅助选择参数。树状图展示了聚类合并的过程，，横轴表示数据点或簇，纵轴表示 ward 方法阈值，树状图中的每一个分支表示从两个簇合并而成的新簇，合并的高度（纵坐标）表示这两个簇之间的 ward 方法距离。一般来讲要找到树状图中距离较大的垂直间隔，表示合并时有显著的差异，可以看到在 ward 方法阈值设定在 7-20 之间的间隔时，会比较好，我们取 ward 方法阈值为 8，可以得到一张聚类图，可以验证在该值对应下的聚类图确实比较恰当。

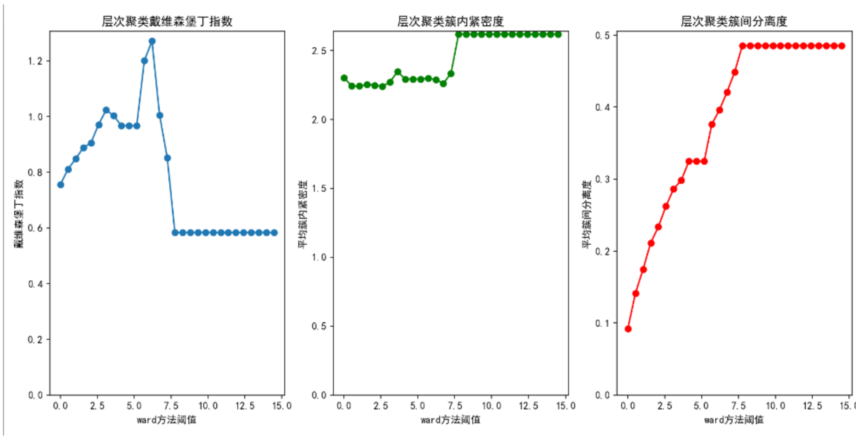

图 46-48：层次聚类戴维森堡丁指数、相应平均紧密度、相应平均分离度随 k 值变化图

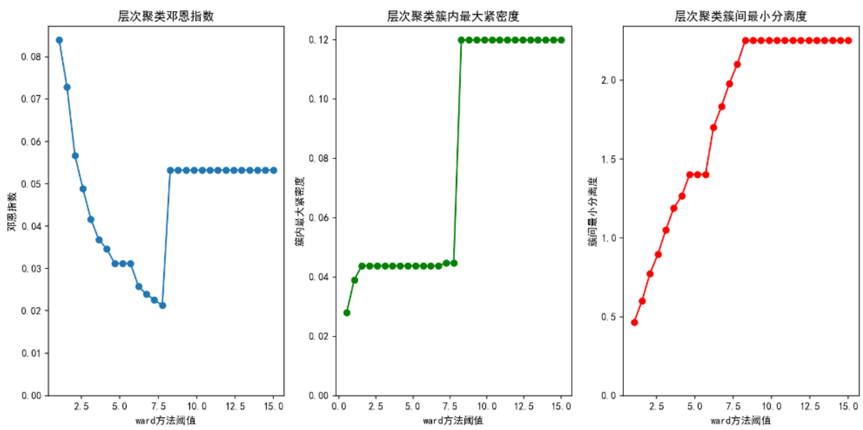

图 49-51：层次聚类邓恩指数、相应最大紧密度、相应最小分离度随 k 值变化图

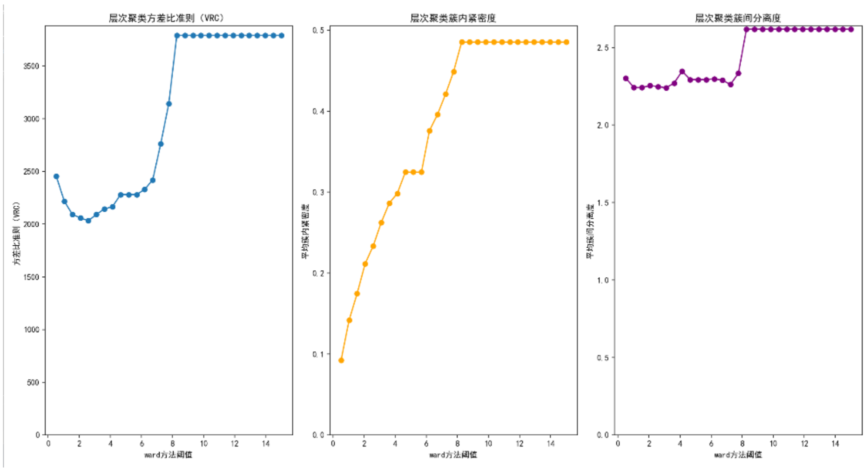

图 52-54：层次聚类方差比准则指数、平均簇内紧密度、平均簇间分离度随 k 值变化图

首先，可以看到层次聚类在戴维森堡丁指数评价下的情况。就总指标趋势来看，呈现不平稳上升，随后快速下降至平稳的趋势，整体在 7.5 稳定保持在最低值。此外看簇内紧密度，可以发现簇内紧密度经历了一段时间的平稳后有所上升，最后继续保持平稳，而看簇间分离度，可以发现簇间紧密度则是快速上升随后保持平稳。可以看到戴维森堡丁指数还是比较准确的，我们选取的 8 也在其低值范围内。

另外，可以看到层次聚类在邓恩指数评价下的情况。就总指标趋势来看，呈现不稳定下降，随后快速上升，最后保持平稳的趋势。整体的最高值在 2.5 一下时出现。其中，由邓恩系数给出的簇间最小分离度相对于戴维森堡丁指数评价而言，上升时更加剧烈，而由邓恩系数给出的簇内最大紧密度和戴维森堡丁指数差别不大。可以发现在 ward 方法阈值较低时邓恩系数呈现的超高值并不太正常，而后续在 7.5 之后回升的值会比较正常，这实际上反映了在 ward 方法阈值较低时邓恩系数不太准确，但是 7.5 之后高值的判断是较准确的。

最后，可以看到层次聚类在方差比准则指数评价下的情况。就总指标趋势来看，保持了一段时间的下降和上升之后突然快速增长，随后达到稳定值，高值在 ward 方法阈值为 8 及之后。簇内紧密度在 8 之前呈现持续快速上升，而簇间分离度呈现先平稳后小幅上升并平稳的趋势。方差比准则指数的情况还是比较符合最初的判断的，8 取值确实也是比较好的 ward 方法阈值。

总结

分别在 K-means 聚类、DBSCAN 聚类、层次聚类中分别使用了三种不同的聚类评价指标，结合参数的不同变化分析指标的变化，以探索三种不同评价指标的特征和区别。

参考资料

[1] 【机器学习-14】K-means 聚类算法：原理、应用与优化_改成随机化初始聚类中心-CSDN 博客
[2]数学建模：K-means 聚类手肘法确定 k 值（含 python 实现）-CSDN 博客
[3]DBSCAN 聚类算法——基于密度的聚类方式（理论+图解+python 代码）_基于密度的聚类算法-CSDN 博客
[4]聚类算法：Hierarchical Clustering 层次聚类_层次聚类算法-CSDN 博客
[5] Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, 20: 53–65.
[6] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on

pattern analysis and machine intelligence, (2), 224–227.
[7] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In 2010 IEEE international conference on data mining, 911–916.
[8] W. M. Rand (1971). Objective criteria for the evaluation of clustering methods”. Journal of the American Statistical Association. 66 (336): 846–850.
[9] Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 410–420.
[10] Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. Journal of the American statistical association, 78(383), 553–569.