An overview of the Chinese version [8]

Clustering v.s. Classification

Clustering 是一种无监督学习方法，它将相似的个体（如样本或基因）聚类到一起，并将不相似的个体分离开来。Clustering 的目的是发现数据中的潜在模式和结构，而不需要预先定义样本的标签或分类。Clustering 通常用于探索性数据分析、发现样本间的生物学相似性、检测异常样本等。常见的 Clustering 方法包括 K-means、层次聚类等。

Classification 则是一种有监督学习方法，它用已有的样本标签或分类信息来训练模型，并将新的样本分配到先前定义好的类别中。Classification 的目的是构建预测模型，用于预测新的样本的分类或标签。Classification 通常用于生物学分类问题，如分类不同组织中的基因表达谱、预测肿瘤的治疗反应等。常见的 Classification 方法包括支持向量机、逻辑回归等。

总体来说，Clustering 适用于没有先验知识或标签的数据集，帮助探索数据中的结构和模式。而 Classification 适用于已经有标签或分类信息的数据集，目的是构建预测模型，对新的样本进行分类。当然，两种方法也可以结合使用，如使用 Clustering 方法对数据进行聚类分析，然后使用 Classification 方法来验证聚类结果的生物学意义。

K-Means clustering (Hard)

K-Means clustering 是一种常见的无监督机器学习算法，用于将数据集分成 K 个不同的、相互独立的类别或簇。K-means 算法包括两个主要步骤：E 步（Expectation Step）和 M 步（Maximization Step）。

在 E 步中，首先需要为每个样本点指定一个初始的聚类中心。具体来说，算法会将所有的样本点按照离它们最近的聚类中心进行分组，这一过程也被称为“聚类分配”。在这个阶段，K-means 算法将根据当前的聚类中心，计算每个样本点与各个聚类中心之间的距离，然后将每个样本点分配到距离它最近的聚类中心所属的聚类中。

在 M 步中，根据 E 步中分配的聚类标签，计算每个聚类的新的聚类中心。具体来说，对于每个聚类，算法将计算所有属于该聚类的样本点的平均值，然后将这个平均值作为新的聚类中心。这一过程也被称为“聚类中心更新”。

E 步和 M 步交替进行，直到聚类结果收敛为止。在每一次迭代中，E 步和 M 步都会更新聚类中心和样本点的分组，并计算当前聚类结果的误差。最终，K-means 算法将根据聚类结果的误差，选择最优的聚类中心和聚类标签。

下面是 K-Means clustering 算法的详细步骤：

1. 首先，选择 K 个聚类中心（也可以随机选择），这些聚类中心可以是数据集中的点。
2. 接下来，计算每个数据点与所有聚类中心的距离，并将每个数据点分配到最近的聚类中心的类别中。
3. 对于每个聚类，计算其所有数据点的平均值，并将该平均值作为新的聚类中心。
4. 重复步骤 2 和步骤 3，直到分类不再发生变化，或者达到预定的迭代次数。

K-Means clustering 算法的优点是易于实现和计算，适用于大规模数据集，并且对于分布较为明显的数据效果较好。它的缺点是对于数据分布不均匀、密集程度不一致的数据集效果较差，聚类结果可能会受到初始聚类中心的影响。

在 genomics 中，K-Means clustering 可用于数据的无监督分类和聚类，例如基因表达数据的聚类分析、细胞群体的分型、突变谱分析等。

Random forest algorithm

Random forest algorithm（随机森林算法）被广泛应用于分类和回归问题的解决。其基本思想是建立多个决策树模型，每个模型使用不同的样本和特征进行训练，最终将它们的结果进行集成，以提高模型的鲁棒性和泛化性。

在分类问题中，Random forest algorithm 通常用于基因表达数据中基因表达模式的分类，如肿瘤和正常样本之间的区分。在这种情况下，输入特征通常是基因表达值，输出结果是肿瘤或正常样本的标签。每个决策树的节点都是一个基因，每个分支都表示该基因在该样本中的表达水平是否高于或低于阈值。通过评估每个节点的基因，随机森林可以确定哪些基因对于分类任务是最具有信息性的。

在回归问题中，Random forest algorithm 可以用于生物数据的拟合和预测，如蛋白质折叠和 RNA 二级结构的预测。在这种情况下，输入特征通常是序列、结构或其他生物特征，输出结果是相应的生物性质或属性。随机森林可以通过构建多个决策树来评估输入特征与输出结果之间的关系，并用于预测未知样本的属性。

Gaussian Mixture Model (soft)

GMM（Gaussian Mixture Model）是一种基于概率密度函数的聚类算法，它假设数据是由若干个高斯分布组合而成的混合分布。通过最大似然估计，可以对混合分布的参数进行估计，从而得到对数据的聚类结果。GMM 的主要思想是对数据进行一定的分布假设，然后通过迭代算法，不断优化参数估计，得到数据的最优分组方式。

GMM 的使用步骤包括：

1. 首先需要对数据进行标准化处理，使得各个特征具有相同的量纲和方差。
2. 确定聚类数量 K。可以通过 BIC、AIC 等指标选择最优的 K。
3. 初始化模型参数。包括每个高斯分布的均值、协方差矩阵和混合系数。
4. 迭代计算每个数据点属于每个高斯分布的概率，即 E 步。
5. 根据计算出来的每个数据点属于每个高斯分布的概率，重新估计模型参数，即 M 步。
6. 重复迭代 E 步和 M 步，直到模型收敛或达到迭代次数。
7. 根据每个数据点属于每个高斯分布的概率，将数据进行聚类。
8. 对聚类结果进行后处理，如剔除离群点等。

$$\mathbb{P}(x_i \mid C_i = 3) = \frac{\pi_3 f_3(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i) + \pi_3 f_3(x_i)}$$

Formula for calculating the probability of point x_i assigned to the 3rd gaussian distribution, where f_k is the Gaussian pdf, π_k is the class specific weight.

Weighted mean: $\bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i}$;

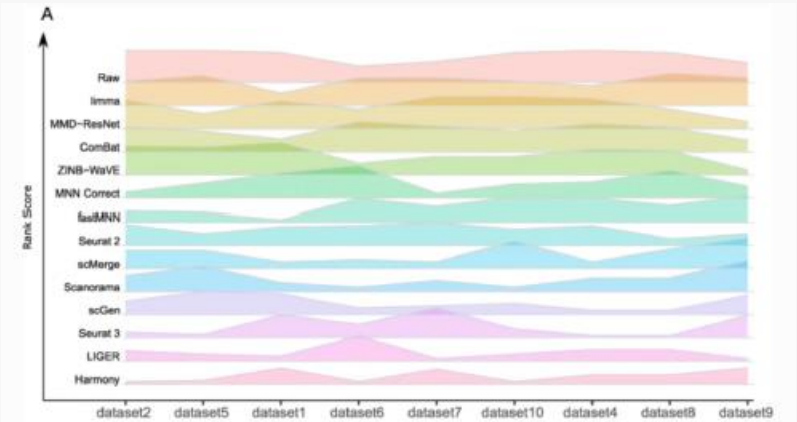
Weighted variance: $s_w^2 = \frac{1}{d} \sum_i w_i (x_i - \bar{x}_w)^2$

Harmony

Harmony 是一种用于单细胞 RNA 测序中批次效应校正的算法。它通过非线性降维（如 t-SNE、UMAP 等）学习每个样本之间的相似性，并利用共变性的角度建模批次效应。以下是 Harmony 在批次效应校正中的步骤：

1. 基于原始的表达矩阵计算非线性降维的坐标。Harmony 支持多种降维算法，例如 t-SNE、UMAP 等。
2. 对于每个细胞，计算一个矢量表示其降维坐标。
3. 对于每个批次，计算一个权重矢量，该矢量表示该批次中样本在降维空间中的分布与其他批次中的样本之间的差异。该权重矢量反映了该批次中的共变性。
4. 通过调整每个样本的矢量，Harmony 通过最小化批次效应来调整每个样本的坐标，从而在整个数据集中获得更好的相似性。

通过对数据的降维和批次效应建模，Harmony 可以在维持尽可能多的生物学变异的同时，减少批次效应的影响。它已经成功地应用于各种单细胞 RNA 测序数据集中的批次效应校正，并成为一个常用的工具。



FDR and FWER

FDR 和 FWER 是用于多重假设检验中的两个概念，下面对它们进行定义、计算方式和作用的解释。

FDR (False Discovery Rate)：假阳性发现率

FDR 是指在多重假设检验中，被拒绝的零假设中实际为真的比例。换句话说，它是指在所有被拒绝的假设中，有多少是错误的拒绝，即被拒绝的假设中实际为真的比例。FDR 的计算方式为：

FDR = (被拒绝的零假设中实际为真的数量) / (被拒绝的总数)

例如，如果有 100 个假设被拒绝，其中有 10 个实际为真，则 FDR 为 10/100 = 0.1。

FDR 的作用是控制多重比较中的错误率。当进行多次假设检验时，会面临多重比较问题，即进行多次比较会增加假阳性（错误拒绝）的概率。FDR 可以作为一种控制方法，帮助确定哪些拒绝假设可能是假阳性，从而减少这种错误。

FWER (Family-wise Error Rate)：家族式错误率

FWER 是指在多重假设检验中至少有一个假阳性的概率，即至少有一个错误拒绝的概率。FWER 的计算方式为：

FWER = P(至少有一个错误拒绝)

例如，在进行多次假设检验时，如果其中有一个假设为真但被错误拒绝的概率为 0.05，则 FWER 为 0.05。

FWER 的作用是控制多重比较中的整体错误率。与 FDR 不同，FWER 的目标是确保整个多重比较中的错误率低于某个严格的阈值，而不是控制错误的比例。