proposal

daisy

2021/5/20

2.5 数据分析

考虑到该系统的不稳定性，有可能出现脱靶、碱基错配、错位编辑等多种现象，数据可能会带有比较大的噪音。为识别数据是否被噪音污染，我们试图采用机器学习的方法搭建二分类模型，对三代测序的数据进行预处理，以便提高后续分析精度，提高系统整体的鲁棒性。

2.5.1 数据的预期形态

在无信号的情况下，我们假设作为参照时钟的正常序列如下所示：

系统一：时间记录仪

AAAAAAAAAAAAAAAAAAAABBBBBBBBBBBB

其中， A 表示 TOE（编辑后的时间信号）； B 表示 INT（未编辑）。A 的长度和系统记录的时间对应。

系统二：信号记录仪（以系统一为基础） C 表示 SIR（记录的信号）。 C 的数量表示信号出现的时间长度。一个可能的序列形态是：

AAAAAAAAAAAAAAACCCCAAAAABBBBBBBB

针对这两种系统，我们有一个朴素的假设，即系统二在系统一的基础上记录信号，若该系统记录时间信息的原件不出现错误，那么系统二出现的错误与记录时间的元件无关，只与信号输入有关；反之，系统二的错误来源于系统一的错误。因此，我们主要分析系统一可能出现的噪音并进行分类排除。

2.5.2 预期噪音

我们预期的可能出现的数据噪音有两类，一类是系统误差，如碱基错配、脱靶等情况，一类是随机误差，类似但不局限于以下几种情形（许多可能的噪音我们无法从原理上做出预判）。

a. 沉默

某次编辑不成功导致此后不能再进行编辑。

ABBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

b. 重启

A) 在沉默后，由于随机错误在下游某位点重新开始编辑

错误发生： ABBBBABBBBBBBBBBBBBBBBBBBBBBBB

错误延续： ABBBBAAAAAAAAAAAAAAAAAAABBBBB

B）正常情况下，在下游某位点开启新的编辑，并且能够同时持续下去

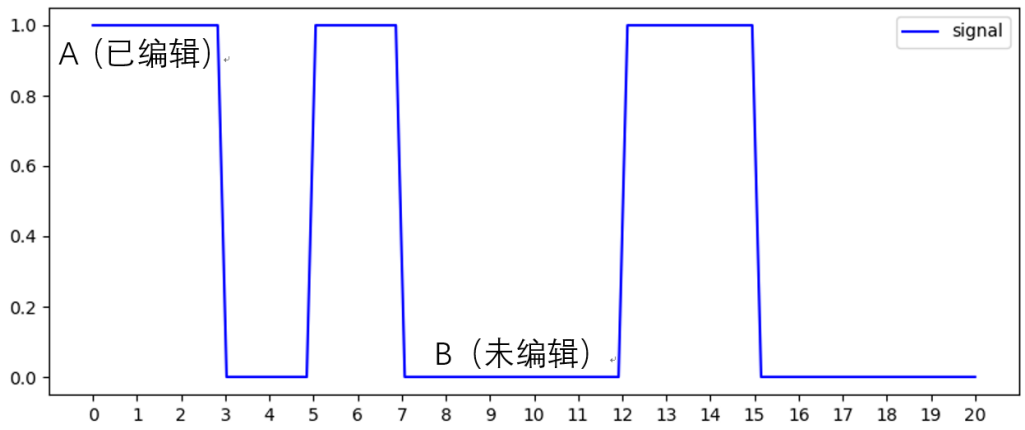
错误发生： ABBBBBBBBBBBBBBBBBBABBBBBBBBBBBB

错误延续： AAAAAABBBBBBBBBBBBBAAAAAAABBBBBB

由于未进行实际实验操作，且考虑到存在其他多种未知的干扰因素会扰乱数据的稳定性，因此，有许多可能的噪音类型我们目前无法进行预判。但是通过以上正确序列和错误序列的比较，我们可以发现序列本身存在一定的规律和特点，针对这些特点，我们使用特征向量来对序列信息进行描述。

$$\omega_i = [谷数_i, 平均谷宽_i, 起始长度_i, 末尾长度_i]^T$$

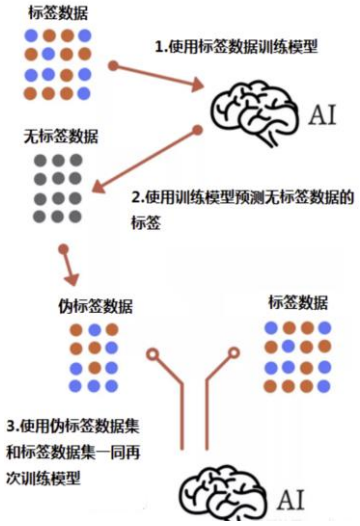
为解释上述特征向量中的每一特征，现需要定义”谷“的概念： 如下图所示



如图所示，其中纵坐标为 1 的点表示 A（已编辑的序列），纵坐标为 0 的点表示 B（未编辑的序列），横坐标上一个单位长度表示一个信号单元（A 或 B）。“谷”指的是前后均为 A 信号的谷，指的是可能出现的编辑错误区段。 特征向量ω_i中，”谷数”指的是序列全长中出现谷的个数，在本图所示案例中为 2；“平均谷宽”指的是所有谷的平均长度，本例中为$(2+5)/2 = 3.5$；起始长度指序列开始端到第一谷出现前的距离，本例中为 3；末尾长度指最后一个谷出现到序列终止端的距离，本例中为 5。因此，以本图信号为例，特征向量为$\omega = [2,3.5,3,5]^T$。通过此特征向量的描述，可以较好地描述一个序列反应的信息。

2.5.3 机器学习模型选取

本系统生成的噪音特征有两点重要性质：一，可以人为预估部分特征；二、部分特征未知，不可人为预估。因此，我们选择半监督学习，在已知部分数据标签（是噪音或不是噪音），绝大多数标签未知的情况下，来根据训练数据进行聚类。我们计划采用的算法是半监督支持向量机（TSVM）。该算法的简要原理是先进行监督学习，对已有标签的部分训练集构造划分超平面，然后根据该超平面对无标签的训练数据进行打标，对新形成的有标签的所有训练数据重新进行超平面划分，不断迭代直到找到最佳的输出（如下图）。



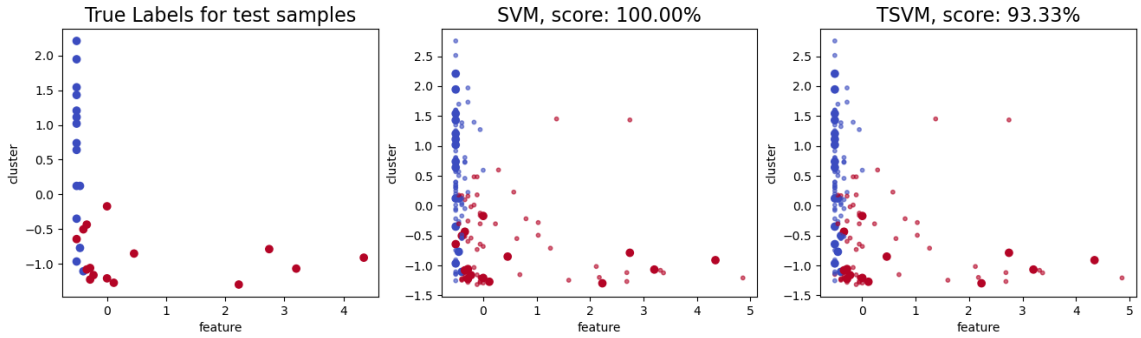
2.5.4 模型可行性验证

为验证我们选择的模型是合理可行的，我们构造了伪随机数据集，如下表所示，其中 label 列表示标签，“0”代表正确编辑，“1”代表错误编辑。数据集的正反例我们各构建了 75 例，共 150 例，对正反例分别来说，各取 15 个数据作为测试集，10 个数据作为带标签的训练集，50 个数据作为不带标签的训练集。

label	peak	Peak_length_ave	Up_length	Down_length	possible situation:
0	0	0	600	400	正确编辑
0	1	1	345	654	正确编辑
1	0	0	120	880	沉默 ABBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
1	23	30	192	211	重启 ABAABBBAAAAABBBAAABBBAAAABBBAAA
1	3	209	121	77	重启 ABBBBAAAAAAAAAAAAAAAAAAAABBBBB

在模拟训练后，我们得到了如下图所示的结果，从左至右依次为数据的真实标签、使用支持向量机（SVM）进行监督学习后的结果、使用半监督支持向量机（TSVM）进行半监督学习后的结果。可以看出，在原本的数据集上，无论是 SVM 或 TSVM 模型泛化能力都达到了 90%以上，说明我们采用的机器学习模型是合理的。但是此处监督学习泛化能力好于半监督学习，其原因为输入的数据集是我们进

行人工预判模拟的，而无不能人工预判的数据作为输入，如进行实际实验，半监督学习模型将会好于监督学习模型。



2.5.5 数据处理结果

如下图所示，为计划从三代测序数据处理至模型训练的分析流程。经过降噪处理，我们可以筛选出有效的信息，并读取其信号以反映时间信息（系统一）及后续的信号记录信息（系统二）。