两阶段填补方法会议纪要

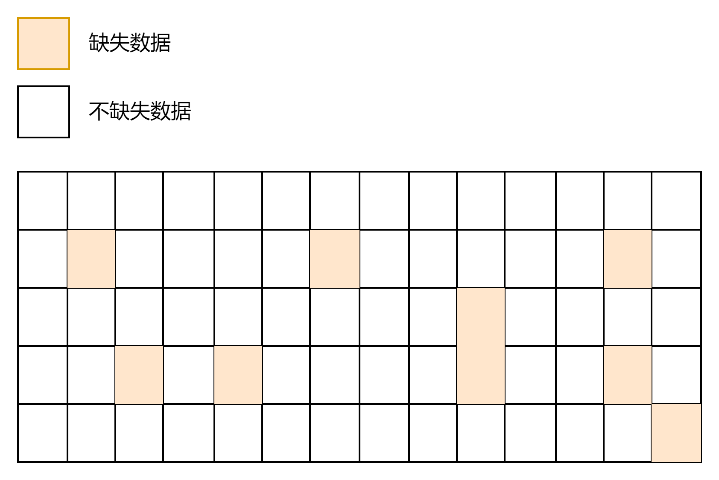
# 概述

初步策略是先利用完整的数据列补全缺失部分，然后结合其他数据进行进一步填补。根据RMSE评估指标，这个"两阶段填补"方法，即先填补部分缺失数据，再结合其他数据进行补全，能有效提升填补效果。

将一个完整的数据集划分为多个部分，并先填补其中一部分的缺失数据，然后再与其他数据合并填补后续数据，这种方法相比于直接使用完整的数据进行填补效果更好。之前进行了验证实验，将数据集划分为三个部分，并先填补其中一部分的缺失数据，然后再与其他数据合并填补第二部分的缺失数据，实验结果表明，这种“两阶段填补”的方法优于单独填补或直接联合填补的方式。

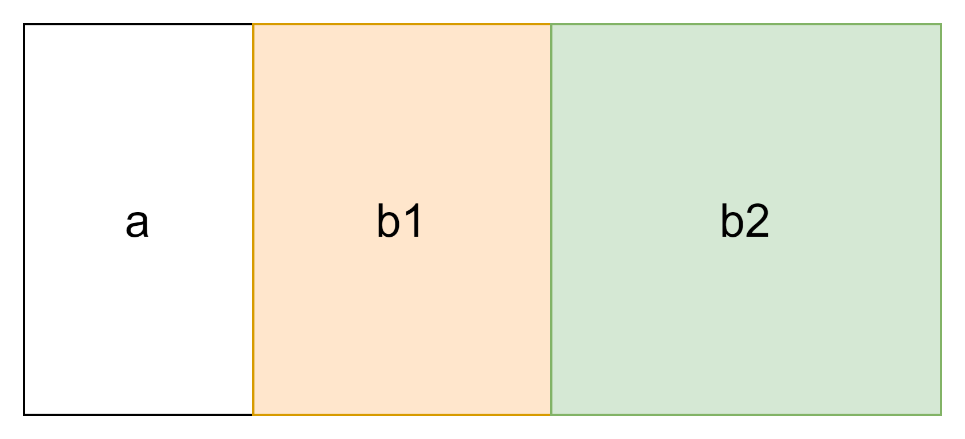
# 方法描述

假设有一个数据集D，其中包含一些缺失值。由于不知道缺失值的位置，它们可能在不同的样本或特定位置上出现，无法确定缺失可能出现在哪个位置，需要一列一列进行寻找。



将数据集划分成三个部分：a、b1和b2，通过以下步骤来划分并处理：

* 第一步，寻找没有缺失值的列，在给定的数据集中，可能存在一列或多列完全没有缺失值。提取这些没有缺失值的列，并将它们作为基准，将这个新的子数据集命名为a。
* 第二步，找出一部分存在缺失值的列，将这些列组成数据集b1，并进行填补处理。
* 第三步，将剩余的缺失率较高的列组成数据集b2。结合数据集a和b1，对b2进行联合填补。



# 方法细节一：缺失的列的选择方式

针对方法描述中的第二步，在处理存在缺失的列时，需要按照一定的规律进行查找，而不是随意选择，现考虑了三种选择方式：

1. 根据缺失率进行选择。
2. 根据相关性进行选择。
3. 根据机器学习的任务或目标进行选择。

## 根据缺失率进行选择

优先选择缺失率较小的列进行填补，因为较小的缺失率通常会导致更好的填补效果，相比于缺失率较高的列。之前已经验证过，当缺失率越高时，填补的效果应该会变得越差。因此，在第二步中，选择在一定范围内的缺失率，并对其进行填补处理。可以设置缺失率的范围，缺失率比较小的就放到b1里面。

## 根据相关性进行选择

在寻找缺失率较低的列，作为数据集b1的过程中，需要考虑这些列与a的相关性，如果这些列与a中无缺失数据的相关性较大，那么我们将其纳入b1。这种相关性的计算可以基于a。

## 根据机器学习的任务或目标进行选择

填补缺失数据的主要目的是为了进行分类，或者是为其他机器学习模型提供数据，以及生成新的样本。在选择缺失数据的列时，如果数据集存在标签，我们可以基于这些标签计算相关性。如果数据集不存在标签，则我们可以采取类似无监督学习的方式，尽可能寻找与目标变量相关性较强的列。这些列将被选入b1，以便在整个数据集中使用。

# 方法细节二：b1和b2的填补方式

根据方法描述，先单独填补b1，然后，将填补好的b1和a联合起来，对于缺失率较高的b2进行填补。在这个策略中，b1和b2可能包含多个列。假设b1有m1个列，b2包含m2个列。因此我们需要验证的一个问题是，在编写算法时，是否需要循环遍历b1中的每一列，逐个进行单列填补，或者将b1中的所有列合并在一起进行填补。

根据之前的实验结果，联合填补可能并不一定好。因此，先循环进行m1个列的单独填补，然后再联合a一起处理b2。先对b1这m1个列进行单独填补，然后再联合a对一起对b2进行填补。这样的顺序可能更有效。

如果做单列循环的填补，可能要用生成方法，如果一起填，才能用填补方法。

# 方法细节三：缺失率比较高时

如果缺失率比较高，比如达到80%、90%，找不到没有缺失的列，可以考虑先寻找缺失率少于一定阈值的列，采用之前的先生成或填补，再联合填补的方式。

# 方法细节四：单独填补b1时的细节

b1那部分可以先聚类再填补