1. 缺失值处理

官方提供的 count 文件中不直接含有缺失值，但将 count 文件按 2017 年至 2020 年逐日展开时，会出现很多日期的病例数量缺失，且在 count 文件中没有记录。经过统计，count 文件中不存在病例数量数据为 0 的情况，但实际生活中肯定会没有病人就诊，因此我认为这些没有病例数据记录的日子，要么是当天就没有收治病人，要么是当天忘记记录。此处我将所有缺失值填充为 0，同时对于 2017.11.1 号之前这部分明显忘记记录的数据予以删除。（济南市儿童医院没有记录 2017.11.1 之前的就诊数据，济南市第四人民医院则记录了 2017.11.1 之前的就诊数据，count 文件给出的病例数量数据是两个医院病例数量的加和）

2. 数据变换

原始数据呈现长尾分布

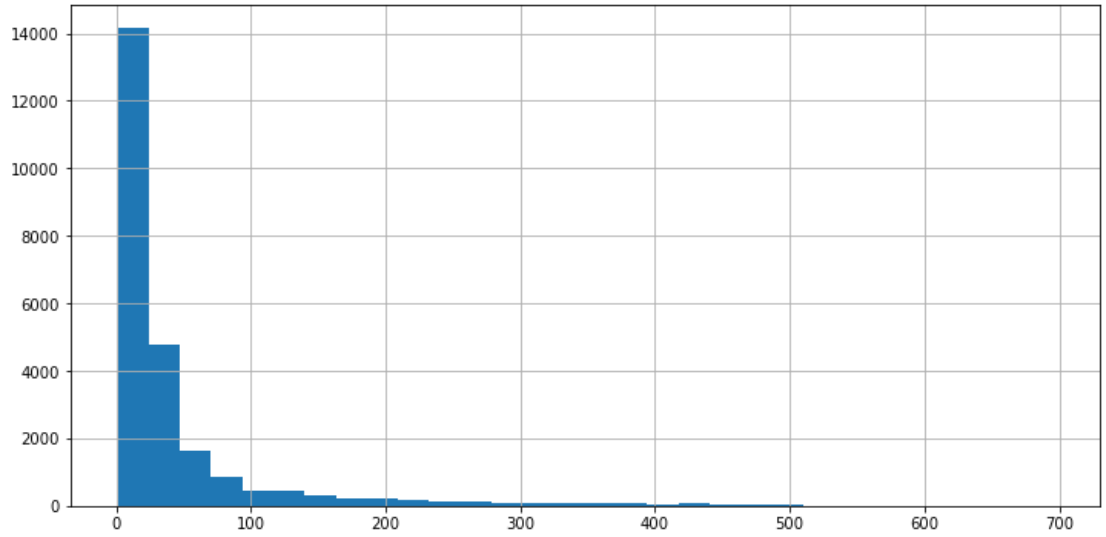

图 1 原始数据的分布

我对原始的病例数量数据进行 log(1+x)的变换，将其近似转化为正态分布，有利于模型的拟合。

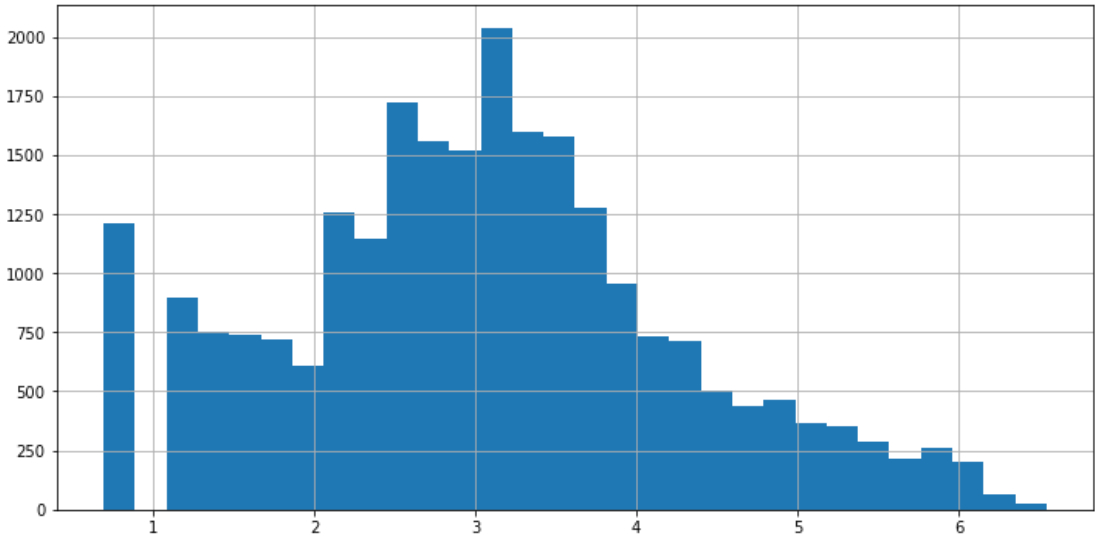

图 2 log(1+x)变换后的数据分布

3. 总体思路

因为需要预测 30 个病种在 2019 年至 2020 年 6 个时间节点之后 14 天每天的病例数量，每个病种单独建模数据量较少，效果不是很好。因此我选择将 30 个病种的数据融合到一起进行训练和预测，相当于模糊了疾病的概念，更多的数据可以让算法更好的学习特征与标签之间的关系。

同时我用 14 组模型、14 个独立的数据集来训练和预测未来 14 天的病例数量，采用第 1 组模型和第 1 个数据集训练和预测未来第 1 天的病例数量；采用第 2 组模型和第 2 个数据集训练和预测未来第 2 天的病例数量；依次类推。14 个独立数据集采样的时间段都是 2017.11.1 以后所有时间。每组模型采用七折交叉验证的方法，使用该组数据集七分之六的数据进行训练，七分之一的数据进行验证，最后将该组七个模型对该组测试集的预测结果取平均，即可得到当天的病例数量预测结果。

最后对数据进行适当的后处理即得到最终的预测结果。

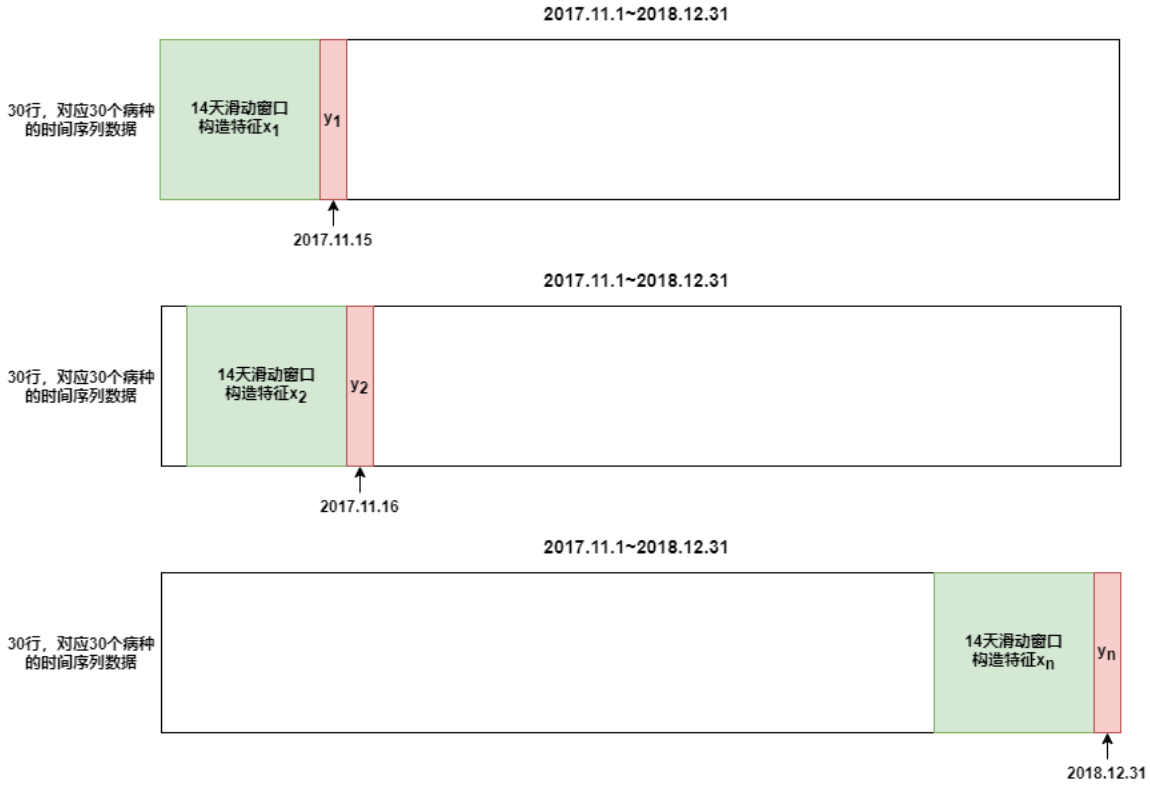

图 3 训练和预测未来第 1 天病例数量的数据集的采样方式

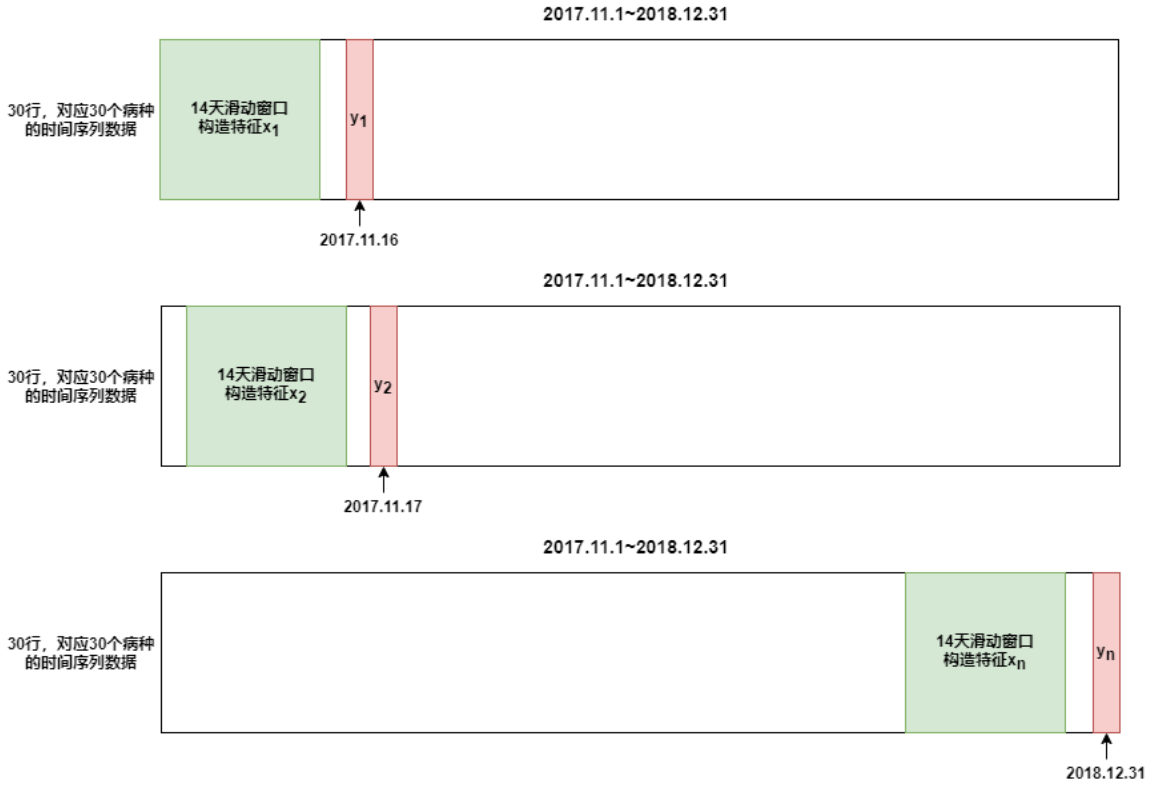

图 4 训练和预测未来第 2 天病例数量的数据集的采样方式

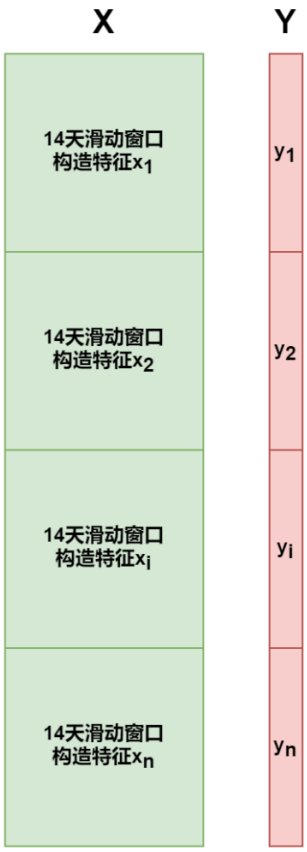

图 5 数据拼接得到对应数据集

如图三所示，不断地在时间序列上滑动时间窗口来构建数据，然后将不同时间节点构建的数据像图 5 一样竖向堆叠在一起，即可得到数据集。为了作图方便，图三和图四的只用了 2017.11.1 至 2018.12.31 之间的数据，但其实 count 文件中 2019 年和 2020 年的数据同样也会用滑动窗口来构建数据。

4. 基于 14 天滑动窗口的特征构建

在时间序列上不断地滑动时间窗口来构建数据集，时间窗口每滑动到一处，就以时间窗口的最后一天为基准，按如下方式构造特征：

过去 3、5、7、10、14 天内病例数量的均值

过去 3、5、7、10、14 天内病例数量的平均绝对误差

过去 3、5、7、10、14 天内病例数量的最小值

过去 3、5、7、10、14 天内病例数量的最大值

过去 3、5、7、10、14 天内病例数量的中位数

过去 3、5、7、10、14 天内病例数量的 20%分位数

过去 3、5、7、10、14 天内病例数量的 40%分位数

过去 3、5、7、10、14 天内病例数量的 60%分位数

过去 3、5、7、10、14 天内病例数量的 80%分位数

过去 3、5、7、10、14 天内病例数量等于 0 的数量

过去 3、5、7、10、14 天内病例数量的指数加权平均

过去 7-10 天、7-12 天、7-14 天内病例数量的指数加权平均

过去 2、4、6、8、12 天内同类别疾病病例数量和的均值（**同类别疾病详见下文“疾病的类别”**）

过去 2、4、6、8、12 天内同类别疾病病例数量和的中位数

过去 2、4、6、8、12 天内同类别疾病病例数量和的最小值

过去 2、4、6、8、12 天内同类别疾病病例数量和的最大值

过去 2、4、6、8、12 天内比值时间序列的均值

过去 2、4、6、8、12 天内比值时间序列的加权乘积

过去 12 天比值时间序列每天的值(**比值时间序列详见下文“比值时间序列”**)

过去 2 周里每周 k 的均值（k=1,2,3,4,5,6,7）

过去 2 周里每周 k 的方差

过去 14 天病例数量的值

过去 3、5、7 天内相邻两天病例数量差值的均值

过去 3、5、7 天内相邻两天病例数量差值的方差

过去 2、3、4、5、6 相邻两天病例数量差值的值

疾病的编号（30 个病种，每个病种对应一个 id）

时间窗口最后一天是星期几

时间窗口最后一天是当月的几号

时间窗口最后一天是今年的第多少天

时间窗口最后一天是今年的第多少周

5. 疾病的类别

30 个疾病，每个疾病都会对应一个时间序列，考虑到数据量较少，因此将这些数据按分位数分箱成 0-24 这样 25 个类别（此处分箱前的时序数据不经过 log(1+x)的变换，而是采用原始数据）。

然后将这些疾病对应的分箱后时间序列数据作为 30 个 sentences，分箱后的每日病例数据作为 sentences 中的 word，使用 ngram=1-3 的 TF-IDF 算法将每个时间序列转化成高维稀疏向量。

进而使用主题数量为 12 的 NMF 矩阵分解算法将每个时间序列对应的高维稀疏向量转化为 12 维稠密向量。至此，每个疾病都会被表征为一个稠密向量。

最后对每个疾病对应的稠密向量取 argmax，得到稠密向量最大值所在的列数，以此作为该疾病的类别。

6. 比值时间序列

根据未经过 log(1+x)变换的原始时间序列，可以计算得到比值时间序列。

假设疾病 1 在 2017 年 11 月 1 日的病例数量为 A，在 2017 年 11 月 2 日的病例数量为 B，在 2017 年 11 月 3 日的病例数量为 C，在 2017 年 11 月 4 日的病例数量为 D。

则疾病 1 对应比值时间序列上 2017 年 11 月 3 日的值为 C/(A+B)，2017 年 11 月 4 日的值为 D/(B+C)。分母为零这些特殊情况的处理请详见代码。