新算法概述

2015/4/21

一、使用环境：

1、数据特征：面板数据或 temporal-spatial 数据的时序（不一定连续，可能有间断）特征无法被现有的时间序列模型提取（已经远远小于 80%，拟合效果没有生产意义），且提取出的微小特征也是不稳定的（极有可能是白噪声）。

2、时间上：涉及的时序数据是稀疏的，这种数值上的稀疏性可以理解为两种状态中的非关注状态频繁连续出现，而我们关注的状态却断续出现，并且这种断断续续的规律性无法在一个有效的观察期内被发现，甚至表现为无规律性。

3、截面（空间）上：当采用构造预测变量对响应变量的回归模型来拟合样本数据时，横截面或空间记录由于数据的稀疏性会带入模型更多的噪声干扰，进而会出现拟合不稳健、预测失效的风险。并且，在选择预测变量上属性意义的相关性不明了，或者说相关性不稳定、不充分。

4、需求呈现上：我们的需求是预测某一截面（或空间）对象在未来某一时间出现某种状态的概率，然后提取出概率值排序靠前的 topN 个截面对象作为未来的当时需要重点观察的对象。这种表示状态的概率必须是同质的（或者说是归一化的），即各个截面对象遵循相同的计算条件、计算法则，否则排序取 topN 是无意义的。

5、模型选择上：根据业务数据的特征分析和需求呈现分析，我们很

容易联想到分类回归模型，比如决策树、逻辑斯蒂回归、支持向量机、神经网络、其他自适应算法。但是，调用这些算法，需要具有很好的预测变量，然而，寻找预测变量，将预测变量转换成有效的形式，也是目前比较难实现的，因为，尽管从属性意义上判断某种预测变量对响应变量的状态有相关性，但是，由于样本数据的稀疏性特征和出于计算性能的考虑，我们无法用线性可加模型的形式将那些未经变换的预测变量直接加入模型。

二、背景假设：

1、横截面对象（空间个体）的时序特征具有记忆性。

注：此处的“记忆性”不等于时间序列中的周期性、平稳性，而是一种用比率（概率）表示的历史上某种状态对当前状态的影响程度，这些比率值构成一个所有空间对象的比率值矩阵。这样的话就弱化了规范性的数学模型中按时序衰减或递增的强假设。

2、变量属性间的相关性可以转化成其他变量以及该变量自身的滞后或延迟算子对该变量某一状态的影响程度来度量。

注：这里相当于只研究相关性的一个方面（前提是两个变量都有正向与负向数值意义，只考虑类别变量）。并且，使得影响程度的数值矩阵能够反映一段时间内的稳定特征。由于这个矩阵不是线性代数中的矩阵意义（因为该矩阵极有可能是奇异的），而是形式上相像，故我们称之为比率矩阵集。

3、以上滞后或延迟的影响程度在时序上（以及计算层级上）是独立、

可加的。

注：此处的时序独立可加是经过频次统计后时序 id 上的独立可加。

4、截面对象（空间个体）是独立的。

注：这里的对象或个体独立是指该对象或个体之间的相对位置是固定的（即我们项目中的格子 id），即我们只研究每一个对象或个体的时序特征。在具体做法上，我们将个体间的相关性转化成一个可加的变量来单独计算其对响应变量的影响比率，并对所有个体逐个做相同流程的计算，以保证计算环境、对比尺度的一致，进而便于横向比较个体对象排列次序。这是最基本的假设环境。

三、执行计算

1、相关层（**×格子数**）：最外层

相关层，表示对该格子在考察期内有无案情的记录进行分离，然后再计算考察期内该格子有无案情的比重。以下所有计算中，我们关注的都是“有案情”这个侧面。

例如，某个格子在某个考察期内的案发相关性可以表示为：

$$rate_{grid_i} = count_{yes}/(count_{yes} + count_{no})$$

其中，$count_{yes}$表示在考察期内发生案情的天数，$count_{no}$表示在考察期内没有案情的天数。

注：从年度、季节、月份等因素来看，这个仅依赖单一考察期计算出的$rate_grid_i$不一定是无偏的，即，可能存在季度或年份上的失效，相关比率是不正当的。那么，这就看如何选择考察期长度和训练集的

更新频次了。Ps：如果我们将考察期也逐天推移计算，然后再计算出所有推移结果的期望，这样做的话会造成更新数据集时计算复杂度的提升，与此同时，噪声也会加剧（因为数据太稀疏、时序回归拟合不足）。目前，我是以单一考察期计算出的$rate_grid_i$来表示对某格子案发相关性的估计。

2、格子层（×**推移长度**）：中间层

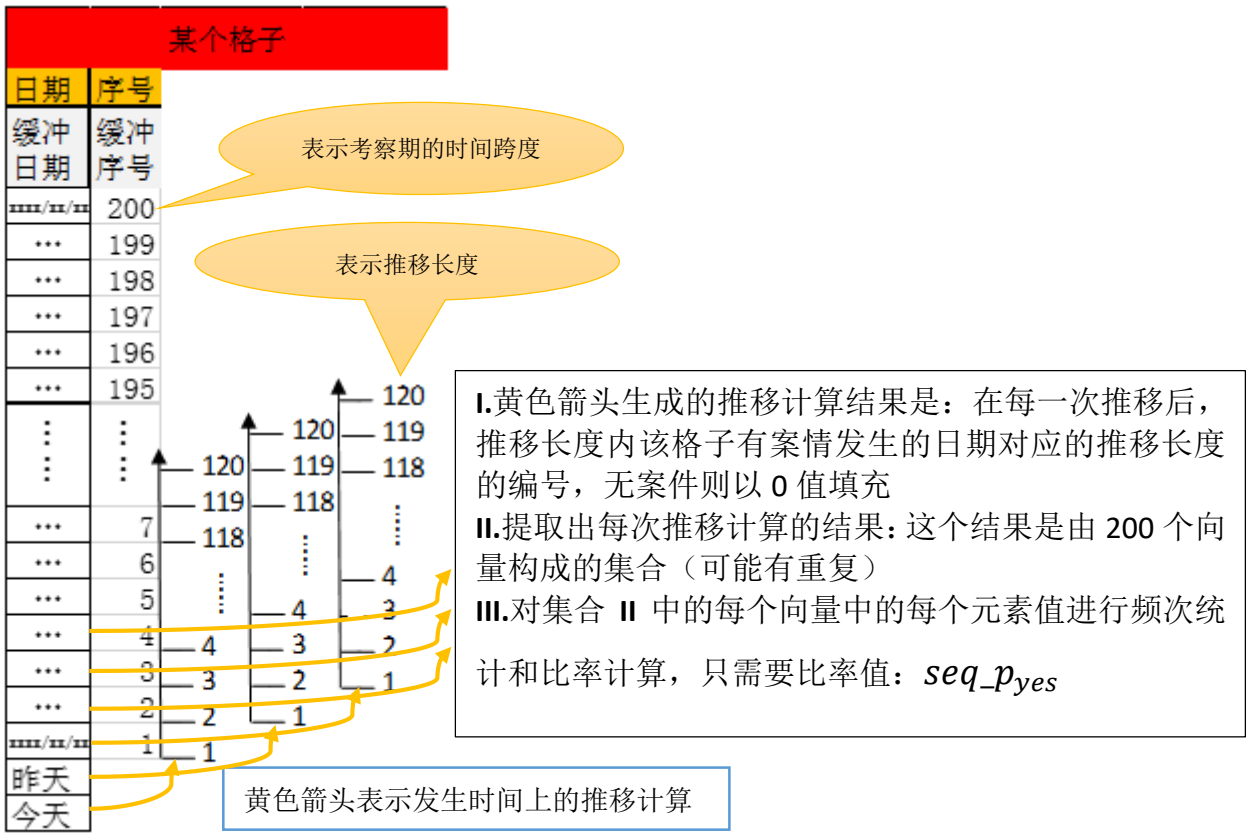
对相关层中分离出的两种情况进行这样的统计：统计出每一次推移提取到的元素（推移期编号）个数，再进行如下计算：

$$rate\_len\_seq_{yes} = len\_seq_{yes}/(len\_seq_{yes} + len\_seq_{no})$$

其中，len_seq_{yes}表示在考察期中有案情记录的某格子在某次推移中也有案情的记录个数，len_seq_{no}表示在考察期中没有案情记录的某格子在某次推移中有案情的记录个数。

3、时间层（×**格子数**）：最内层

对相关层中分离出的两种情况进行这样如下步骤的计算：



4.构造比率集：

$$rate\_set_{grid\_seq_i} = rate_{grid_i} \times rate\_len\_seq_{yes} \times seq\_p_{yes}$$

注：以上公式中矩阵相乘的维度表示略。

四、运行预测

1、取出预测数据集，时间跨度要足够发生推移计算，即满足推移期限定。

2、对预测数据集进行推移计算，并根据“执行计算”出的比率集进行匹配计算。

注：我目前使用的匹配计算方式是简单的独立相加策略，即，当有多个影响因素时，我先是单独计算该因素对本格子案情与否的影响（执行计算），然后，将所涉及到的所有影响因素的比率相加起来。

这样可能会出现比率值大于 1 或等于 0 的情况。

3、出格子和对应的权重比率

注：这里的权重比率是对当天所有格子根据匹配计算后归一化调整后的值。同时，添加了热点补救，即，当提取的 topN 的比率值中有 0 时，再用热点计算一遍，补充需要推送的格子。

4、评估（按策略，按班次）