SSA和随机森林论文讨论

1. 2017-03-07

1.1. SSA 相关问题

1.1.1. 趋势项、波动项、噪声项 的 定义、由来

- `趋势项、波动项、噪声项` 的 `定义` 是什么?

> 第一个奇异值包含原始序列最基本的信息，而且能够被用于构造用来研究 股指长期趋势的趋势项。奇异熵去噪法被用于导出包含噪音的原始序列的 奇异值，这被用来构造出研究股市噪音影响的噪音项。然后剩下的奇异值 就被用来构造研究不同事件影响的股市波动项。

这是原文中对三个项的定义，我没看明白。奇异值是一个一维向量对不对， 那么“第一个”奇异值应该只是一个数，后面怎么变成一个时间序列的呢?

另外，为什么第一个奇异值是趋势项、另外的是波动项呢? 趋势、波动的含义是怎么定义的啊?

1.1.2. 这三项的经济学含义

> 然而，在图像中，趋势项表明深圳综指在2010年再次开始下降。
>
> 股市波动项的曲线表明从2010-2013年股市波动较为平稳，2013-2014年则 波动相对剧烈，这表明一些事件对股指产生的冲击往往会持续几个星期到 几个月不等，而重大事件的影响会持续更长时间。
>
> 噪音项主要反映了股市在股指上全体噪音的影响。

这几句话的逻辑我不明白。`趋势项`等指的就是图中的直线对不对？趋势 项这条直线的怎么变化，表明 "趋势项表明深圳综指在2010年再次开始下 降"呢？其余两项我也有相似的问题:P

1.2. 随机森林预测结果分析问题

1.2.1. 图表4 与 图表5 的区别联系

图表4使用的 1212 组样本是什么数据呢？与图5用的2015到16的数据有什么不同呢?

图表4这202个样本点的y值是什么呢?

什么量化指标能够用于衡量的图4、图5的优劣呢?

1.2.2. 结果分析

- 数学模型间的逻辑问题

  我们使用SSA将价格数据分成了3项，具体在随机森林中如何处理的呢？ 我们看到这三项与股指相关性是不同的，那么是否需要在随机森林的初 始化中体现出来?

- 方法的逻辑问题

  我们使用3项数据用来预测价格。我认为预测价格是非常困难的事情，单单 使用收盘价来，因为信息非常不充分，几乎是不可能达到预测价格这个 目标的

- 不同的时间维度上来研究其特征

  在文章中，哪里体现出是在不同时间维度的呢？或者`不同时间维度`的定义是什么?

1.3. 文章改进思路

大的思路有两条：

- 扩展数据集

  这点我们想的是一样的:D 单从收盘价时间序列出发，所包含的信息太少， 不足以对股价走势进行预测。我相信单单对数据集的扩充就一定能取得 更好的效果

- 理清文章中方法的逻辑、数学模型间的逻辑

  首先，我觉得我们不该把目标放在预测收盘价这么困难的问题上。收盘 价是由非常多信息综合的结果，甚至绝大多数信息是我们无法观测的。 我们是否可以将目标缩小一些，比如，对股市的涨跌趋势（上涨、下跌 两种情况）进行预测呢?

  其次，为了改进SSA，或者选用更好的方法，我们需要先明白什么是趋势 项、波动项、噪声项。它们背后的实际含义是什么？只有我们明白了这 些东西的实际含义，才能选取对应的模型挖掘这些信息。

最后，文中有些表述不准确的小问题。比如随机森林等方法可以被用来拟合任意曲线，但是随机森林只对 submodular 问题保证收敛到局部最优。如果想证明收敛问题我们首先得证明问题是 submodular 的。SVM 也只是对线性问题保证收敛到全局最优。对于凸问题只能保证收敛到局部，非线性问题是不保证收敛的。这些都是小问题，但是如果投论文的话我们还是表述精确些安全点好:P