相似性距离的合并促进交互时间序列分析

# 摘要：

在时间序列数据中寻找相似性趋势在从财务规划到政策制定等领域发挥重要作用。这些多面关系的检测，特别是不同长度和对齐的时间序列的时间扭曲匹配，是非常昂贵的计算。为了在大时间序列数据集上实现实时响应，我们提出了一种称为时间序列在线探索（ONEX）的新范例，采用强大的一次性预处理步骤，对关键的相似关系进行编码，以支持后续的快速数据探索。由于对于所有可变长度时间序列片段的大量成对相似性关系的编码是不可行的，因此我们的工作依赖于以下重要的见解：利用便宜的点到点距离的聚类，例如欧几里德距离可以支持随后的时间扭曲匹配。我们的ONEX框架克服了与更坚固的弹性距离相关的禁止的计算成本，即DTW通过在令人惊讶的紧凑知识库而不是原始数据上应用它。我们的比较研究表明，ONEX的精度提高了19％，比最先进的技术快几倍。除了是一个高度准确，快速的领域独立的解决方案，ONEX提供了一个真正的互动探索体验支持新的时间序列操作。

# 介绍

## 动机

在金融，商业，医学到气象，时间序列数据是普遍存在的，表现为股票波动，心电图，降雨量等。

动态实例，让我们来看一个现实生活中的例子：揭示在寻找和利用数据相似性的决策中涉及的挑战。2013年，马萨诸塞州政府废除计算机软件服务的销售使用税，被认为会对本州的经济产生负面影响。数据驱动证据分析显示各种税率随时间的变化以及从大量政府公共网站获得的社会和经济因素的波动，所有这些都是以时间序列为模型。期间遇到了许多困难，主要的是找到和解释以时间序列表示的经济指标之间的相关性。

（1）来自不同领域的数据在特定时间间隔内的存在需要比较不同长度和序列的时间序列，因为税收变化的影响可能需要不同的持续时间才能变得明显。这种时间感知比较必须使用鲁棒距离来执行，例如动态时间规整（DTW）。这些措施的“权力”，在精度方面是有益的，被它们的计算复杂度所遮蔽，因此即使对于中等大小的数据集的时间响应也很慢。

（2）在这个过程中，分析师使用具体指标，如增长率，来评估引入新税的潜在影响。例如，他们“设计”了表示税收积极影响的样本增长率时间线，并在所有州中搜索匹配。在这种情况下，序列可能存在也可能不存在于数据集中。 如果发现完美匹配，则反映那些特定状态下的类似影响，而紧密匹配表示特定状态的影响略有不同。因此，分析人员需要能够通过使用可能存在或可能不存在于这些数据集中的样本序列来探索大时间序列数据。

（3）分析人员必须解决如何找到最佳匹配序列的复杂问题。例如，他们搜索重复的相似性模式，例如几年内一个国家的增长率以及类似的增长率和不同国家在特定时间长度上的其他经济指标，表明类似的“短期”影响。

（4）来自不同域的数据需要使用不同的参数设置，例如相似性阈值，导致每个参数设置的重复和冗余计算。 例如，用于研究人口统计数据的相似性的最合适的阈值不同于用于生长速率的阈值。

## 当前面临的研究挑战和局限性

1. **高数据基数导致响应性降低**

诸如财务记录或ECG数据收集的时间序列数据集往往是巨大的。更糟的是，适应不同时间粒度的需要要求我们考虑不同长度的时间序列。对于包含N个时间序列的数据集，每个长度n，要考虑的子序列的总数是N n（n-1）/ 2。例如，一个基准数据集，如来自UCR时间序列数据挖掘Archive2的StarLightCurves，具有9236个时间序列，每个长度为1024，由4.83e9个子序列组成。现实世界的数据集往往比这个数量级大3。对所有这些子序列进行相似性比较显然是不切实际的。

许多最先进的技术试图解决这种缺乏瞬时响应性。滞后的响应性在勘探过程中可能失去分析师的注意。最先进的技术面临着准确性和时间响应之间的权衡，特别是在探索非常大的数据集时。一些系统提供精确或高度准确的解决方案[1]，[19]，[26]，而牺牲及时的响应。这可能对在医学和财务中的应用是有害的，这取决于对行动的即时回答。其他[2]使用预处理步骤来提高及时响应。然而，它们对设置许多不同参数的要求限制了它们的效率[22]