Salient Subsequence Learning for Time Series Clustering

论文地址：https://ieeexplore.ieee.org/document/8386690

中文翻译：https://blog.csdn.net/SCS199411/article/details/90759528

源码：The source codes are available at https://github.com/BlindReview/shapelet.

Introduction

什么是shapelet?

shapelet是时间序列的子序列，具有判别性，其在某种意义上最大程度地代表类。

> A shapelet is a time series subsequence that is identified as being representative of class membership.

实际应用中的时间序列研究在诸如金融、医学、轨迹分析和人类行为识别等领域无处不在。从无监督时间序列中发现特征是非常有意义的，因为它不仅有益于时间序列聚类，而且还可以从原始时间序列中找到潜在的模式。例如，在2012年的一篇名为的论文，介绍使用选定的无监督特征(unsupervised features)对非侵入性胎儿ECG信号进行聚类可以改善诊断系统。心脏病专家已经证实，通过无监督方法获得的特征对应于ECG信号的p波，p波描述了发育中的胎儿心脏形态的变化。

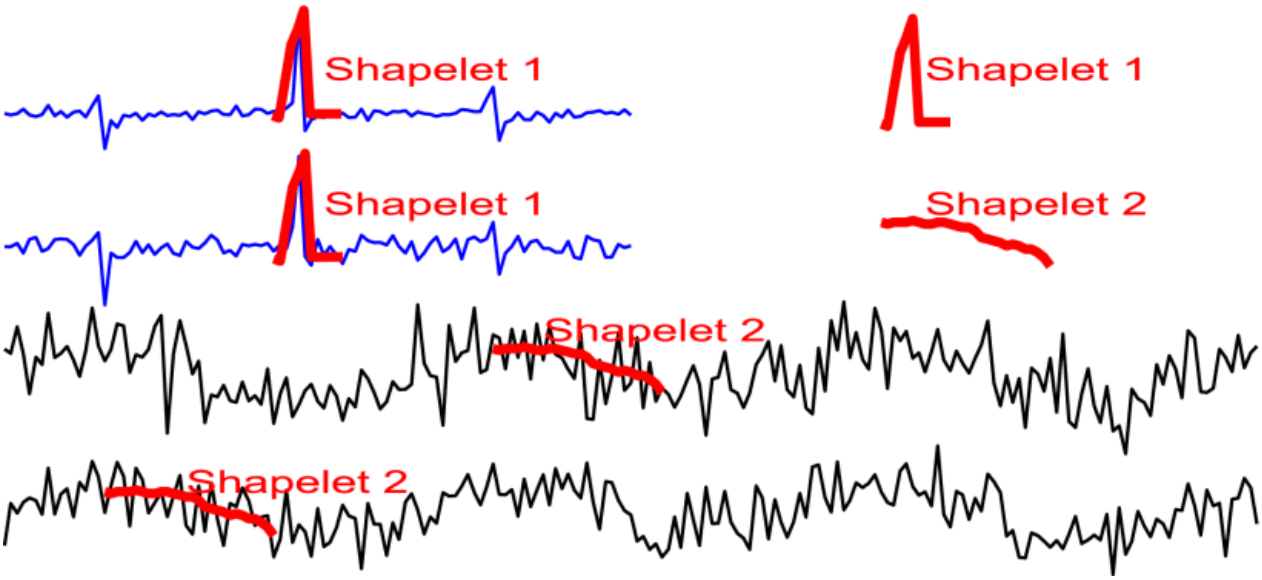
发展过程

最初，shapelet候选集是通过扫描整个时间序列来获取，时间复杂度很高，以Synthetic Control dataset 举例，它只包含600个时间序列， 长度为60，但是它的shapelet候选集中有 10^6数量级个时间序列。最近Grabocka等人提出一种基于回归学习的方法寻找shapelets，大大减少了时间复杂度，但是需要标记，引入了标签成本问题，因为它需要监督学习。这促使我们建立了一种更经济的shapelet学习模型，该模型可以自动学习shapelet，而无需人工标记。我们将学到的无监督shapelet称为lu-shapelet，lu-shapelet针对未标记的时间序列聚类具有区分的能力和出色的效用。

Ye和Keogh （《*Time series shapelets: A new primitive for data mining*》）所做的开创性工作通过全面扫描时间序列的所有候选片段来发现最佳的形状。根据预定义的距离度量（例如信息增益）对所有的Shapelets进行排名，并选择最能预测类标签的段作为Shapelets。然而，尽管已经努力去改进算法进行加快计算了，但该方法仍具有较高的时间复杂度。另外，所有这些方法都尚未克服大型段池（a large segment pool）的挑战。例如，综合控制数据集包含600个时间序列样本，每个样本的长度为60，这意味着所有长度的候选片段的数量为1.098×10^6如此众多的候选人可供选择，这对形状选择方法提出了巨大挑战。近几年，Grabocka人提出了关于时间序列中发现Shapelet的全新观点，该方法使用回归学习（Regression Learning）来查找小波，而不是在候选池中搜索小波。这种方法为学习从时间序列中学习shapelets（learning shapelets from time series）打开了新的大门。

论文提出了一种新颖的无监督显著子序列学习（USSL）模型，用于在未标记的时间序列数据中自动学习shapelet。首先，伪标签将无监督学习转换为有监督学习。然后，使用正则化最小二乘技术和光谱分析来学习lu-shapelets，伪类标记和伪类边界。接下来，添加一个shapelet正则项以避免学习相似的lu-shapelets。最后，它采用坐标下降算法同时获得伪类标签并以迭代方式学习最佳lu-shapelets。

不同于现有的通过穷尽搜索原始时间序列中的每个候选段来找到shapelet的无监督shapelet选择模型，我们的方法旨在直接学习最优shapelet（lu-shapelets），它可以最佳地协助从未标记的数据中进行时间序列学习。学习的shapelet与原始时间序列段不完全相同。



Contributions

1、USSL【无监督显著子序列学习】。 论文提出的模型综合了伪类标签（pseudo-class labels），频谱分析（spectral analysis），正则化最小二乘技术（egularized least-squares techniques）和Shapelet正则化（shapelet regularization），扩展了监督的Shapelet学习模型以处理未标记的时间序列数据。

2、通过对36个真实数据集和一个综合数据集进行实验，论文对USSL的效率和有效性进行了经验验证。测试结果显示出良好的性能，而且这个算法是非常先进的（state-of-the-art）。

Relative Work

非基于shapelet的无监督特征学习

研究人员还提出了一系列从无监督数据中选择信息特征的无监督方法。无监督特征选择的传统标准要求选择能够最好地保留原始时间序列特征空间的数据相似性和流形结构的特征。但是，这些方法不能包含数据中隐含的重要信息，因此不能直接应用于形状学习问题。现有的无监督特征选择方法分别评估每个特征的重要性，并逐个选择最佳特征，而**忽略了不同特征之间的相关性**。

用于无监督特征选择的最新方法通过同时利用**特征相关性和判别信息**来选择特征。无监督的判别性特征选择通过考虑其多种结构来选择最佳的信息特征来表示数据。

一种流行的方法是将伪类标签引入无监督的特征选择中，以预测良好的聚类指标。离群值和噪声数据是另一个重要因素，会严重影响最终的特征选择性能。现实世界中的数据通常不会以理想的方式分布：数据离群值和噪声是司空见惯的。因此，重要的是使不受监督的特征选择对异常值和噪声具有鲁棒性。

基于shapelet的无监督特征学习

shapelet在以前也有人将其用于时间序列的聚类问题了。*《Clustering time series using unsupervised-shapelets》*一文中就提出了一种称为“ u-Shapelets”的算法，该算法使用无监督的shapelet对未标记的时间序列进行聚类。该方法选择一组无监督的Shapelet，**以通过搜索和删除子集来分离原始数据集**。它迭代地重复直到没有数据剩余为止，并使用贪婪搜索算法来尝试最大程度地扩大被无监督的shapelet分开的不同组之间的间隔。

Paparrizos等人在*《k-shape: Efficient and accurate clustering of time series》*提出了一种新的基于形状的方法，称为k-shape，在保持域独立和有效的同时对时间序列进行聚类。 K-shape使用迭代可扩展的细化过程，该过程创建了分离良好且均匀的簇。具体来说，它采用归一化互相关（normalized cross-correlation）度量来定义两个时间序列的距离，而现有的其他方法则考虑形状和曲线。在k-shape中，在时间序列等移动和缩放期间，距离的度量保持不变。时间序列分配以及聚类质心在每次迭代中基于归一化互相关进行计算和更新。

基于shapelet的监督特征学习

shapelets selection

《Time series shapelets: A new primitive for data mining》、*《Searching and mining trillions of time series subsequences under dynamic time warping》*,两篇论文选择最佳的短时间序列段，引入了精确的类标签预测方法。它的核心概念就是利用给定的类别标签对所有**候选shapelet**（即candidate shapelet给定训练时间序列的分段）的可预测性进行**评分**。在研究时间序列shapelet的开创性工作中，提出了通过建立一个决策树分类器来递归搜索含有信息的shapelet，该分类器以由信息增益决定的距离为特征。Mood中位数，F-Stat 和Kruskall-Wallis也用于shapelet选择。

时间序列数据集通常包含大量候选子序列，这些子序列需要蛮力法来选择shapelet，但是这样的代价是它们往往会导致不理想的运行时间。因此，许多研究人员提出了各种提高计算速度的方法。其中有人提出了一种新颖的实现的方式是对信息增益使用启发式熵修剪（use entropy pruning on the information gain heuristic），并且尽早放弃距离计算。还有一些人则修剪搜索空间并重新利用（以前的）计算（re-use computations）。还有人通过在SAX表现方式（*Fast shapelets: A scalable algorithm for discovering time series shapelet*）中搜索潜在有趣的候选者或使用频率很低的Shapelet来对候选者进行修剪。

Shapelets learning

要想学习到有价值地shapelet，除了穷尽的搜索shapelet外，最近的研究提出了学习最佳shapelet的新观点。通过将shapelet表示为可通过回归技术学习的参数，而不是简单地在一组训练数据中搜索候选片段池，相对于其他基于shapelet的分类方法，此新方法在统计学上有显着改进。 回归学习方法能够学习并获得任意的shapelet。 分析人员不再局限于一组有限的候选shapelet。

与这些以前的研究流不同，论文提出的USSL方法通过引入无监督的Shapelet学习将未标记的时间序列与自动学习的Shapelet聚类，从而将无监督的特征选择与Shapelet学习结合在一起。

Unsupervised Shapelet Learning

USSL整体框架如下：

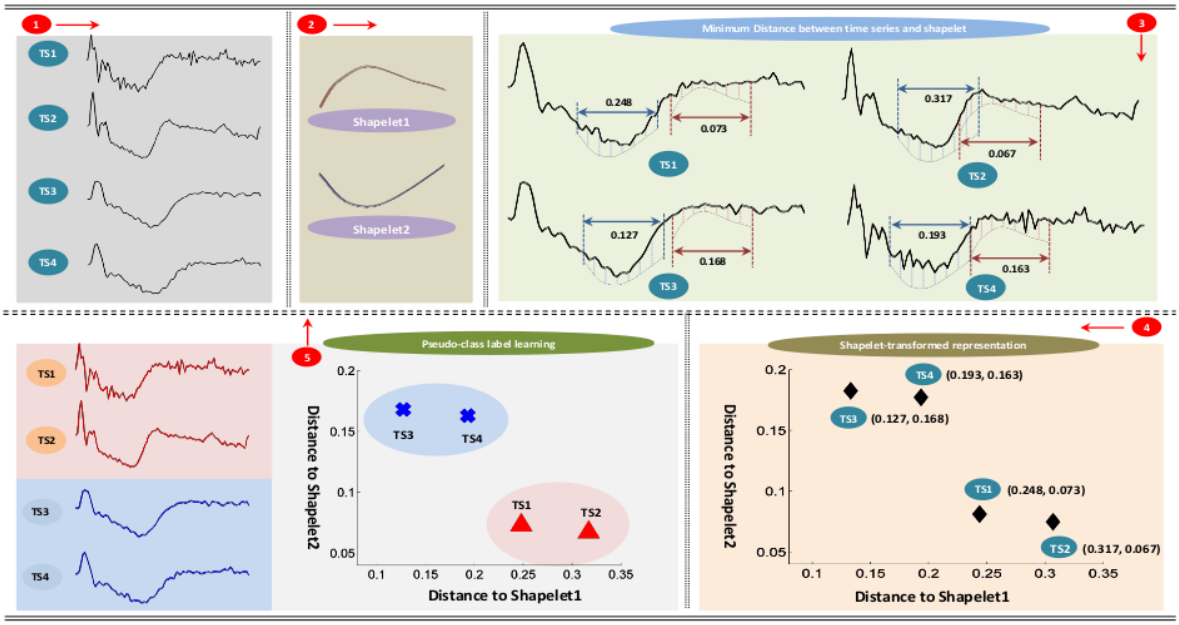


Fig. 2. The framework of the proposed Unsupervised Salient Subsequence Learning (USSL) model. From the original time series(①), we first learn shapelets using a shapelet similarity minimization principle(②); then, the distances of the learned shapelets and time series are calculated(③); the original data are mapped into the shapelet-based space(④). In shapelet-space, USSL learns the pseudo-class labels and a pseudo classifier using spectral analysis and regularized least-squares minimization(⑤); then, we update the shapelet with the new learned pseudo-class labels and the pseudo classifier(②); we then repeat this process until convergence. Finally, the USSL ensures the optimal shapelets and the pseudo-class labels.

> **【Fig.2解释】**无监督显着子序列学习（USSL）模型的框架。从原始时间序列（图1），我们首先使用shapelet相似度最小化原理（图2）学习shapelet。 然后，计算学习到的shapelet和时间序列的距离（图3）； 原始数据被映射到基于shapelet的空间上（图4）。 在shapelet空间中，USSL使用频谱分析和正则化最小二乘最小化（图5）学习伪类标签和伪分类器。 然后，使用新的学习到的伪类标签和伪分类器（图2）更新shapelet； 然后，我们重复此过程，直到收敛为止。 最后，USSL确保了最佳的shapelet和伪类标签。

整个USSL模型主要有以下几个部分：

1、Shapelet-transformed Representation

2、伪标签

3、频谱分析

4、最小二乘法

5、无监督的显著子序列学习

此外，论文还运用了坐标下降法：当要优化的函数含有多个自变量的时候，每次只更新一个自变量、固定其他自变量。

Shapelet-transformed Representation

参考论文：A Shapelet Transform for Time Series Classification

Shapelet Transform是为了将长的时间序列减少到长度更短得多的Shapelet 空间矢量。 时间序列有序性的Shapelet Transform很好地保留了形状信息。

在进行Shapelet Transform之前，我们需要生成一组Shapelets候选对象。

生成候选集合

假设有一个长度为m的时间序列T，S是T的一个长度为$l(l <= m)$的时间子序列。S是由T中连续l个点组成的。任意一个长度为m的时间序列包含了了$(m-l)+1$个长度为l的不同子序列。我们把时间序列T_i的所有长度为l的子序列表示成$W_{i,l}$，而长度为l的所有子序列的集合，数据集T为：

$$W_l = \{W_{1,l}, \ldots, W_{n,l}\}$$

数据集T的所有候选Shapelet的集合为:

$$W_l = \{W_{min}, W_{min+1}, \ldots, W_{max}\}$$

其中min≥3，max≤m。 请注意，$W_{i,l}$非常大，具有$O(m^3)$个候选形状。

Shapelet Transform

Shapelet Transform的过程如下：

Consider a time series set $\mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ and a set of shapelets $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$. The shapelet-transformed matrix is represented as $\mathbf{X} \in \mathbb{R}^{k \times n}$, where each element $\mathbf{X}_{(\mathbf{s}_i, \mathbf{t}_j)}$, simplified as $\mathbf{X}_{(ij)}$, represents the distance between time series $\mathbf{t}_j$ and shapelet $\mathbf{s}_i$.
$\mathbf{X}_{(ij)}$ can be calculated as

$$\mathbf{X}_{(ij)} = \min_{g=1,\ldots,\bar{q}} \frac{1}{l_i} \sum_{h=1}^{l_i} (\mathbf{t}_{j(g+h-1)} - \mathbf{s}_{i(h)})^2,$$

其中：

where
$\bar{q} = q_j - l_i + 1$ denotes the number of segments of
l_i-length in
$\mathbf{t}_j$;
l_i represents the length of shapelet
$\mathbf{s}_i$;
q_j is the length of
$\mathbf{t}_j$.
$\mathbf{X}_{(ij)}$ is a function with respect to shapelet set
$\mathcal{S}$, i.e.,
$\mathbf{X(S)}_{(ij)}$, by omitting the variable
$\mathbf{S}$ for simplicity.

Pseudo-Class Labels

没有标签的训练样本是无监督学习的最大挑战。 因此，我们在完成shapelet 转换表现形式后，需要引入伪类标签。 假设时间序列数据集属于c类，这意味着伪类矩阵Y∈Rc×n包含c个伪标签。 属于第i类的时间序列t_i的概率由Y_{ij}表示。 并且，如果对于∀ī≠i，存在$Y_{ij} > Y_{avg_i,j}$，则t_i属于群集i。

Spectral Analysis

论文还运用到了频谱分析。频谱分析已在无监督学习中广泛采用，指导原则是，密切相关的示例可能具有相同的伪类标签。

Shapelet Similarity Minimization

最后，鉴于我们希望学习具有多种形状的shapelets，如果该模型输出相似的shapelets，则会受到惩罚。

similar shapelets. Assume the shapelet similarity matrix as
$\mathbf{H} \in \mathbb{R}^{k \times k}$, and the similarity between two shapelets
$\mathbf{s}_i$ and
$\mathbf{s}_j$ is represented by the element
$\mathbf{H}_{(\mathbf{s}_i, \mathbf{s}_j)}$. We use
$\mathbf{H}_{(ij)}$ to denote
$\mathbf{H}_{(\mathbf{s}_i, \mathbf{s}_j)}$ for simplicity, and we have

$$\mathbf{H}_{(ij)} = e^{-\frac{\|d_{ij}\|^2}{\sigma^2}},$$

Unsupervised Salient Subsequence Learning

下面等式给出了形式化的无监督显着子序列学习模型。这是一个包含三个变量的联合优化问题：伪分类器。

$\mathbf{W}$; the pseudo-class labels
$\mathbf{Y}$; and the candidate shapelets
$\mathbf{S}$.

$$\min_{\mathbf{W},\mathbf{S},\mathbf{Y}} \frac{1}{2} tr(\mathbf{Y}\mathbf{L}_G(\mathbf{S})\mathbf{Y}^\top) + \frac{\lambda_1}{2}\|\mathbf{H}(\mathbf{S})\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{W}^T\mathbf{X}(\mathbf{S}) - \mathbf{Y}\|_F^2 + \frac{\lambda_3}{2}\|\mathbf{W}\|_F^2.$$

第一部分是谱正则化项，以保持时间序列之间的局部结构。第二部分通过最小化shapelets之间的相似性来使其多样化。第三部分则为学习最佳伪类标签和伪分类器的最小二乘正则化。

Experiment

Dataset

我们采用了UCR档案中的36个时间序列基准数据集。 每个数据集包含40到7,164个序列。 数据集内序列的长度相等； 数据集之间的序列长度在80到637之间变化。这些数据集都带有注释，并且每个序列仅属于一个类。 每个原始数据集中包含一个训练集和一个测试集。

链接：www.cs.ucr.edu/~eamonn/time_series_data/

Conclusion

论文通过整合shapelet学习，频谱分析，伪类标签和最小二乘最小化的优势，本文提出了一种用于时间序列的无监督显着子序列学习的新优化模型，称为USSL。 所提出的模型可以自动学习显著shapelet，以帮助时间序列聚类。 实际数据集和合成数据集的结果均表明，与最新的无监督时间序列学习方法相比，USSL可以学习有意义的无监督小形函数，并具有良好的性能。

我认为通过将时间序列异常检测问题转换为对shapelet的分析是一个很好的思路。我们可以搜索shapelet然后拿来作为分类依据的一个方法，因为实际时间序列很多有存在特征明显的shape，比如说心电图数据一次正常心跳简化一下就是前后两个小的峰中间加一个高峰，那么如果缺了一块的shape可能就是作为鉴别异常心跳的依据。