浙江大学

ZJUI 学院 2022 年暑期科研项目总结材料

中文论文题目： 融合图结构的时间序列建模

英文论文题目： Modelling and Classifying Time Series with Shapelet Transition Graphs

成员姓名：陈梓源（3200112407）

陈志榕（3200112414）

指导教师：杨　洋

指导教师所在学院：计算机科学与技术学院

暑期研究起止日期：2022 年 6 月 15 日 ~ 2022 年 8 月 15 日

摘要

本次暑期科研中，项目组成员研究融合图结构对时间序列进行建模的神经网络，围绕现有的Time2Graph模型“用图结构捕捉Shapelets（有代表性的子序列）迁移过程”的核心思想，探索加速Shapelets提取的算法。Time2Graph是浙江大学杨洋副教授团队在2020年提出的时间序列异常检测模型，其主要创新点是将时间序列分段与Shapelets对应，构建“状态转移图”并根据其表征进行分类，但模型的时间复杂度过大。

项目组设计了AQOURSNet（自编码无监督的强表征Shapelets提取）网络结构：对于给定的时间序列数据集，以所有子序列作为Shapelets候选，用TS2Vec（无监督时序表征网络）计算时序表征向量，用K-Means或K-Medians聚类产生Shapelets，用DTW（动态时间规整）描述子序列与Shapelets的相似程度（或距离），利用正则化距离构建的邻接矩阵建立图结构，用GAT（图注意力网络）学习图表征向量，依此对时间序列进行分类并尝试进行实现异常序列的提取。

代码、文档、数据集、参考文献等资料可参考项目组的GitHub代码仓库。

Abstract

In this summer research program, the project team researched on modelling time series from the perspective of graphs, with a focus on accelerating shapelet extraction in the current model of Time2Graph. Time2Graph is a model put forward by the Yang Yang Lab in Zhejiang University in 2020. The core idea is to match the subsequences with shapelets and perform classification on the constructed "transition graph," yet the model is time-consuming.

The project team designed the AQOURSNet (Autoencoded Quantification of Unsupervised Representative Shapelets) in response to this issue. For a given time series database, the model sets all subsequences as shapelet candidates and calculates their embeddings using TS2Vec. The K-Means or K-Medians cluster centers of these features are picked as Shapelets. The DTW (Dynamic Time Warping) distances between subsequences and shapelets form an adjacency matrix, which gives a graphical portrait for each time series. A GAT (Graph Attention Network) is then trained to embed and classify these graphs and, therefore, the series.

Code, docs, dataset, and references are available from the team's GitHub repository.

1 研究背景

过去几十年中，时间序列分类是一个热门的科研话题。基于Shapelets的时间序列分类方法引起众多学者的兴趣：Shapelets是最能表示整段时间序列的子序列，也即整段时间序列中信息含量最大的一段子序列，可以用于高效区分不同类的序列。2009年，Lexiang Ye和Eamonn Keogh提出了Shapelets的概念(Ye and Keogh, 2009)，十余年来产生了不少就如何提取Shapelets并基于此进行分类的研究。

在学术界，Shapelets的提取方法有两大流派。“选择法”的代表有Shapelets预处理(Lines, 2012)(Hills et al., 2014)、随机扩张变换(Guillaume, 2022)等，其核心思想为将所有时间序列的任意长度子序列作为Shapelets候选，用这些候选分别对时间序列进行分类，分类效果最佳的即为选定的Shapelets。这种算法直观易实现，但由于候选子序列的数目过于庞大，会消耗大量的计算资源，且算法复杂度过高。

“生成法”的代表有多层感知机(Grabocka, 2014)、深度网络(Wang, 2017)(Fawaz, 2019)、生成对抗网络(Wang, 2019)(Ma, 2020)、无监督学习(Li, 2021)等，主要思想是先随机初始化Shapelets，通过不断优化分类任务的效果，将Shapelets调整成最具有代表性的子序列。神经网络可以学习到深层表征，但经常有一部分生成的Shapelets与时间序列样本相似度过低，不具有可解释性。因此，项目组以寻找既能提高算法效率、又具备较高的解释性的Shapelets提取算法，作为暑期科研工作的重心。

Shapelets提取后，如何用其对时间序列进行分类也是研究热点之一。传统方法是直接利用Shapelets和时序的欧氏距离或动态时间规整（Dynamic Time Warping）距离进行分类。2020年杨洋副教授团队发布的Time2Graph模型(Cheng, 2020)结合“状态转移图”的思想，以Shapelets作为节点、Shapelets之间的转移概率作为边权，提出用图结构来描绘时间序列，后用图神经网络学习表征并分类。时间序列与图结构相融合的模型，创新性地将一维的时间序列同复杂的拓扑结构联系起来，具有更强的普适性与表征能力。同时，在异常序列检测中，异常样本往往由于样本量过小而被忽视，但Time2Graph模型可以捕捉异常序列具有的反常Shapelets迁移路径，增强对异常序列的关注。

论文阅读笔记（算法演化）　　论文阅读笔记（神经网络）

2 模型算法

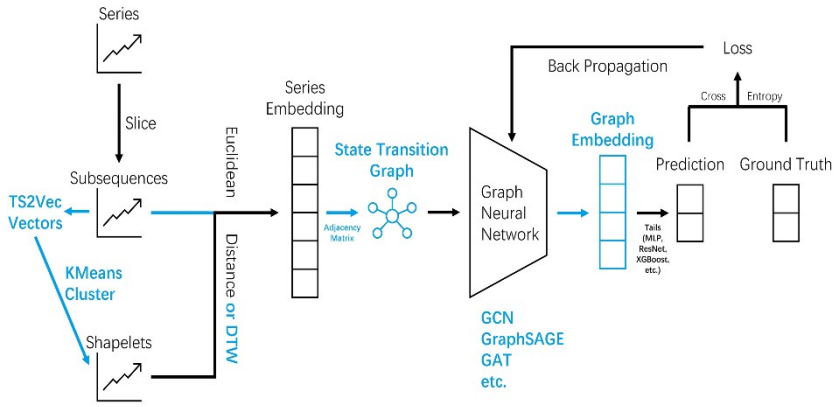

图1. AQOURSNet: Autoencoded Quantification of Unsupervised Representative Shapelets（自编码无监督的强表征Shapelets提取网络）结构图 | 蓝色为在文献基础上新增的结构

2.1 术语

时间序列/时序/Series：排列顺序表示测量值在时间上的顺序的数组

数据集/Dataset：多个时间序列的集合，一般来自对多个对象的同时测量

子序列/子序/Subsequences：特定时间序列的连续一部分，一般为固定长度

代表性子序列/Shapelets：最能表征数据集中时间序列并为其分类的子序列

距离/Distance：数值与子序列和代表性子序列之间相似程度呈负相关的量

图/Graph：以代表性子序列为顶点、以其互相迁移的概率为边权建构的图

2.2 Shapelets提取（代码）

首先选定子序列长度（超参数），对于数据集中的每个时间序列，用滑动窗口将其拆分为子序列，作为Shapelets候选值。一个自然的想法是，不同类别的时序，其子序的形状应当区别较大，故理想的Shapelets彼此应尽可能不相似，否则会出现冗余。

K-Means聚类是经典的无监督聚类算法，用于将高维空间中的点按距离划分类别。在选定聚类数量（超参数）后，以子序列之间的欧氏距离作为距离判据，采用每一个簇的中心点作为Shapelets，理论上便可实现“最大化彼此距离”的目标。

K-Means的时间复杂度是O(n)，运行效率相当高；但在实操过程中，这一算法仍有几处可优化。其一，各簇的中心点由簇中各样本点平均算得，经常过于平滑以致于损失局部信息，故转而采取各簇离中心点最近的样本点作为Shapelets。其二，提取出的Shapelets有时由同一时序移动窗口而得，形状有部分重叠，解决方法或是以动态时间

规整（DTW）作为距离判据，缩短形状重叠的子序之间的距离，鼓励其被聚类进同一簇；或是以TS2Vec (Kazemi, 2019)学习出的子序表征作为样本点，但耗时较多。

2.3 时序表征（代码）

对于每个时间序列，时序表征旨在获得每个子序到每个Shapelet的距离，作为该子序的表征向量。形状重叠问题同样存在时序表征阶段：在（第二次）利用移动窗口计算Shapelets与子序之间的距离时，存在形状重叠的Shapelets会产生相近的距离值，最终无法解决冗余问题。DTW同样可以鼓励选取多样化的Shapelets。

这里就“多样化”作详细解释。前文所提Shapelets提取的“选择法”与“生成法”两条路径，其本质是对Shapelets不同特性的追求。选择法的判据是分类效果，也就是选取分类效果最好（Distinctive）的Shapelets，这固然可以满足分类需求，但要求在Shapelets提取阶段就不断参考数据集标签、计算损失函数并进行反向传播，而无监督提取方法可以避免这部分的计算量，同时具有更好的可推广性。

与选择法形成对比的是，生成法重视的是表征力，也即暂时以非对比的视角，选取最能代表当前时序特征（Representative）的Shapelets，这同时也可以在一定程度上解决神经网络带来的不可解释性问题。文献(Ma, 2020)(Li, 2021)指出，多样化（Diverse）的Shapelets可以为每个时序计算出特点鲜明的表征，故项目组一直致力于利用DTW、TS2Vec等算法对子序或距离进行预处理，缓解形状重叠带来的冗余。

2.4 图建构（代码）

仍是对于每个时间序列，每对相邻子序的两个表征向量可以确定一次“Shapelets迁移”，考虑所有的子序确定的所有迁移，可由此构建状态转移图，用于下一步表征。一个Shapelet在转移图中体现为一个节点，两个Shapelets之间的转移概率体现为边权；“转移概率”的计算方法较复杂，此处从略。“状态转移图”或是借鉴了有限状态机中的提法，将时序与图论相结合，以动态视角分析时序演变，确实是新颖的想法。

由此构建的转移图理论上属于完全图，但由于当今部分图机器学习算法会放大过小的边权，模型中可以设置临界值（超参数）以裁去部分过弱的连接，默认保留30%的边。这种剪裁带来的影响是两面性的：一方面，异常时序会在相邻时间段中出现反常的状态转移，剪裁去部分正常边有利于放大异常；另一方面，这种反常转移出现的概率经常偏低，有时甚至在瞬间产生，因此有不小的风险裁去异常边，反而损失了最宝贵的表征。能否捕捉到微小的异常，也与子序长度密切相关，因此计算加速有时会损失精度。

2.5 图表征（网络结构、训练）

AQOURSNet中的图机器学习算法采用GAT（图注意力机制）(Veličković, 2017)学习状态转移图的表征，结合(Wang, 2017)提及的支持向量机、多层感知机与残差神经网络作为预测尾，将表征向量转换为预测值。值得注意的是，残差网络中原先存在卷积层，但卷积只适合直接运用于原始时间序列，Shapelets的排列顺序是随机的，时序表征向量的相邻值并不具有空间上的实际联系，故模型中将卷积层更换为全连接层。

由于暑期科研时间有限，项目组还有不少令人兴奋的想法未能如期实现，如原版Time2Graph中使用的XGBoost预测尾及其与PyTorch神经网络的联合训练、数据并行与多GPU分布式计算、用于异常检测的正负样本不平衡取样及其调优，等等。

3 实验数据

我们使用时序分类的经典UCR数据集(Dau, 2019)测试网络，总体达到SOTA平均水平。ShapeNet基线数据来自(Li, 2021)，Time2Graph基线数据来自(Cheng, 2021)。

数据集		AQOURSNet					基线	
		Shapelet数量	子序列数量	图表征向量维度	训练准确率	测试准确率	ResNet	BSPCOVER
不使用TS2Vec	Earthquakes	60	10	128	88.82	75.54	71.20	**81.68**
	Strawberry	60	10	256	92.50	**96.49**	89.57	93.24
	WormsTwoClass	40	10	128	100.00	70.13	**74.70**	74.59
	BeetleFly	60	20	64	100.00	80.00	85.00	**90.00**
	Coffee	60	10	256	100.00	**100.00**	**100.00**	**100.00**
	DistalPhalanx	60	40	256	78.83	76.81	77.10	**83.17**
	ECG200	80	20	64	100.00	**94.00**	87.40	92.00
	ECGFiveDays	40	40	128	100.00	95.01	97.50	**100.00**
使用TS2Vec	Earthquakes	80	40	128	81.99	74.82	71.20	**81.68**
	Strawberry	80	20	128	64.27	64.32	89.57	**93.24**
	WormsTwoClass	80	40	128	88.95	58.44	**74.70**	74.59
	BeetleFly	80	20	128	100.00	65.00	85.00	**90.00**
	Coffee	80	20	128	50.00	53.57	**100.00**	**100.00**
	DistalPhalanx	80	40	128	79.83	74.28	77.10	**83.17**
	ECG200	80	40	128	100.00	**92.00**	87.40	**92.00**
	ECGFiveDays	80	40	128	100.00	97.10	97.50	**100.00**

表1. AQOURSNet与ResNet基线(Fawaz, 2019)、BSPCOVER基线(Li, 2021)分类能力比较

数据集	AQOURSNet				Time2Graph基线		
	Shapelet 数量	子序列 数量	训练 准确率	测试 准确率	Time2Graph	Time2Graph+ Static	Time2Graph+
Earthquakes	40	20	95.96	75.54	**79.14**	76.98	77.70
Strawberry	50	20	98.69	96.49	**96.76**	95.95	96.49
WormsTwoClass	60	20	98.90	70.13	**72.73**	70.13	71.43

表2. AQOURSNet与Time2Graph基线(Cheng, 2020)(Cheng, 2021)分类能力比较

Time2Graph已被杭州电网应用于检测偷电用户，属于异常检测的经典案例。当异常样本比例很低（1.47%）时，传统的预测准确率不再有说服力，故项目组为训练报告提供了F1分数的支持。值得注意的是，单纯对负样本增强取样并不会显著提升AQOURSNet的分类效果，仍无法精确识别异常样本，或与前文所言“无法捕捉快速变化”相关。

结论

Time2Graph模型创新性地融合图结构与时序分析，以“状态转移”的思想，将时间序列刻画为Shapelets的迁移。本次暑期科研中，项目组设计的AQOURSNet旨在以无监督方法（K-Means）加速提取Shapelets，以数据预处理（TS2Vec）增强普适性，利用图神经网络（GAT）学习时序表征获得的邻接矩阵，并就形状重叠问题予以回应（DTW）。

该网络虽然在UCR数据集的二分类问题上基本达到SOTA平均水平，但在正负样本数量相差过于悬殊的情况下会失效，故高效异常检测算法仍有待进一步研究。尽管如此，项目组成员在暑期科研中仍收获颇丰：在读懂Time2Graph两篇论文的前提下，从零开始成功复现模型，锻炼编程能力；在原有网络基础上活跃地提出了数项改进措施，并充分融入实验室氛围，探索兴趣方向的同时，也为日后深造筑牢基础。

附件：与导师汇报工作用 PPT

Works Cited

Cheng, Ziqiang, et al. "Time2Graph: Revisiting Time Series Modeling with Dynamic Shapelets." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, 2020, pp. 3617-3624., https://doi.org/10.1609/aaai.v34i04.5769.

Cheng, Ziqiang, et al. "Time2graph+: Bridging Time Series and Graph Representation Learning via Multiple Attentions." *IEEE Transactions on Knowledge and Data Engineering*, 2021. https://doi.org/10.1109/tkde.2021.3094908.

Dau, Hoang Anh, et al. "The UCR Time Series Archive." *ArXiv.org*, 2019. https://arxiv.org/abs/1810.07758.

Fawaz, Hassan Ismail, et al. "Deep Learning for Time Series Classification: A Review." *Data Mining and Knowledge Discovery*, vol. 33, no. 4, 2019, pp. 917-963. https://doi.org/10.1007/s10618-019-00619-1.

Grabocka, Josif, et al. "Learning Time-Series Shapelets." *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 392-401. https://doi.org/10.1145/2623330.2623613.

Guillaume, Antoine, et al. "Random Dilated Shapelet Transform: A New Approach for Time Series Shapelets." *Pattern Recognition and Artificial Intelligence*, 2022, pp. 653-664. https://doi.org/10.1007/978-3-031-09037-0_53.

Hills, Jon, et al. "Classification of Time Series by Shapelet Transformation." *Data Mining and Knowledge Discovery*, vol. 28, no. 4, 2014, pp. 851-881. https://doi.org/10.1007/s10618-013-0322-1.

Kazemi, Seyed Mehran, et al. "Time2Vec: Learning a Vector Representation of Time." *ArXiv.org*, 2019. https://arxiv.org/abs/1907.05321.

Li, Guozhong, et al. "Efficient Shapelet Discovery for Time Series Classification (Extended Abstract)." *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 2336-2337. https://doi.org/10.1109/icde51399.2021.00254.

Li, Guozhong, et al. "ShapeNet: A Shapelet-Neural Network Approach for Multivariate Time Series Classification." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 8375-8383. https://doi.org/10.1609/aaai.v35i9.17018.

Lines, Jason, et al. “A Shapelet Transform for Time Series Classification.” *KDD ’12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 289-297. https://doi.org/10.1145/2339530.2339579.

Ma, Qianli, et al. “Adversarial Dynamic Shapelet Networks.” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, 2020, pp. 5069-5076. https://doi.org/10.1609/aaai.v34i04.5948.

Ma, Qianli, et al. “Triple-Shapelet Networks for Time Series Classification.” *2019 IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 1246-1251. https://doi.org/10.1109/icdm.2019.00155.

Medico, Roberto, et al. “Learning Multivariate Shapelets with Multi-Layer Neural Networks for Interpretable Time-Series Classification.” *Advances in Data Analysis and Classification*, vol. 15, no. 4, 2021, pp. 911-936. https://doi.org/10.1007/s11634-021-00437-8.

Rakthanmanon, Thanawin, and Eamonn Keogh. “Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets.” *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013, pp. 668-676. https://doi.org/10.1137/1.9781611972832.74.

Veličković, Petar, et al. “Graph Attention Networks.” *ArXiv.org*, 2018. https://arxiv.org/abs/1710.10903.

Wang, Yichang, et al. “Learning Interpretable Shapelets for Time Series Classification through Adversarial Regularization.” *ArXiv.org*, 2019. https://arxiv.org/abs/1906.00917.

Wang, Zhiguang, et al. “Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline.” *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1578-1585. https://doi.org/10.1109/ijcnn.2017.7966039.

Ye, Lexiang, and Eamonn Keogh. “Time Series Shapelets.” *KDD ’09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 947-956, 2009. https://doi.org/10.1145/1557019.1557122.