arXiv:2302.03132v1 [cs.LG] 6 Feb 2023

通过时间序列的持久景观进行神经网络中的重要性归因

Aina Ferr`a [1, 2], Carles Casacuberta [1, 2], Oriol Pujol [1,2]

摘要

我们提出并实现了一种方法，利用通过拓扑数据分析获得的面积归一化持久景观矩阵，借助神经网络分析时间序列。我们在网络架构中加入了一个门控层，能够识别对分类任务最相关的景观层级，从而实现重要性归因。接下来，我们将已选的景观函数与原始时间序列中相应的临界点进行匹配。通过这种匹配，我们能够重建时间序列的大致形状，从而深入了解分类决策。我们使用来自心电信号数据集的输入数据对该技术进行了测试。

引言

利用拓扑数据分析（TDA）方法提升神经网络性能的做法十分普遍。在某些情况下，其目的是提供灵活的向量化表示 [4], ，或提高预测准确率或分类准确率 [10], ，又或通过向算法输入从数据中提取的拓扑信息来正则化学习算法 [6, 7, 15]。拓扑学也被用于缩减数据集的规模，同时避免训练准确率大幅下降 [17]。文献 [22]综述了深度学习中利用贝蒂数进行时间序列分析的拓扑数据分析方法。文献 [20], 研究了数据集经过训练良好的神经网络各层时其拓扑结构的变化追踪，而文献 [16]分析了神经元激活的拓扑结构。文献 [1, 8]讨论了如何借助持久性描述符评估泛化差距，而无需测试集。

Bubenik 在 [3] 中首次将景观用作持久性描述符。 [18] 中曾将景观用于深度学习，目的是通过向后续层添加输入数据的拓扑特征信息来提高可学习性，但并非用于可解释性。激活景观也曾被用作神经网络性能的拓扑摘要 [23], ，以及用于个性化心律失常分类 [24]。

本文运用拓扑数据分析来提升深度学习中分类结果的可解释性。具体而言，我们使用持久性景观，从数据中提取神经网络为执行分类任务而重点关注的特征信息。我们对数据进行预处理，使网络接收一系列景观层级，而非原始信号。持久性景观的层级结构使我们能够设计一种方法，找出信息量最大的层级。为此，我们在选定的架构中引入一个附加层，其任务是为数据集中给定信号的持久景观层级分配权重。随后，我们仅使用权重最高的景观层级再次运行网络。结果表明，所选景观层级集合（通常为 2 至 4 个）所得的分类准确率与前 10 个层级的集合相近。

为深度学习分类任务选取最相关的景观层级，使得仅使用所选景观函数便能部分重构给定数据成为可能。由此得到的简化数据版本揭示了数据信号中的哪些部分与网络的分类任务最为相关。这不仅提供了有关网络学习过程的信息，也揭示了数据所承载的最关键特征。

[1]巴塞罗那大学（UB）数学与计算机科学系，加泰罗尼亚议会大道585号，08007 巴塞罗那，西班牙， {aina.ferra.marcus, carles.casacuberta, oriol pujol} @ub.edu[2]本研究得到 MCIN/AEI/10.13039/501100011033 资助，资助编号为 PRE2020-094372（A. Ferr`a）；此外还得到项目 PID2019-105093GB-I00（A. Ferr`a、O. Pujol）和 PID2020-117971GB-C22（C. Casacuberta）的资助。

在心跳分析（第 3.2 节）中，我们确认，与使用原始数据时相比，输入来自已选景观层级的信号重构后，我们的神经网络仍能获得相近的准确率。这为分类结果提供了更有力的可信依据，证明网络在学习过程中并未关注人为伪造的细节。

我们的重构方法在一篇配套文章中有更精确的描述 [14], ，该文探讨了与本文相关的一些数学问题，并涉及拓扑数据分析中的逆问题，即如何从持久性摘要中恢复特定类型的数据 [2, 13, 21]。

作为我们方法的研究成果，我们发现，对于某个数据集，被选为最相关的景观层级数量与使用该数据集训练的神经网络的准确率呈负相关（见下图 6）。因此，对于深度学习分类器而言，难以确定哪些景观层级最为重要的数据集，往往对应着较低的分类性能。

第 1 节汇总了持久性景观的基本事实，第 2 节介绍了我们针对景观层级的归因算法。在第 3.1 小节中，我们使用 UCR时间序列分类档案中的九个数据集验证了我们的方法 [9], ，并在第 3.2 小节中使用该方法测试了 MIT-BIH 心律失常数据库[19]中心电信号的分类准确率。正如第 3.3 小节所讨论的，信号平移可能会导致神经网络的分类准确率下降，而持久性景观及其得出的结果在数据水平平移下保持不变。因此，在这类平移可能由不希望出现的影响所致时，使用景观进行分类具有优势。

1 次水平集的持久景观

时间序列数组可以看作一维连续分段线性函数，可对其应用持久同调来研究次水平集的演变。因此，我们考虑沿着 y 轴滑动参数 t ；对于定义在区间 $[a, b]$ 上的每个函数 f ，以及 t 的每个取值，我们计算相应次水平集 $L_t(f)$ 的连通分支数，其定义如下：

$$L_t(f) = \{x \in [a, b] \mid f(x) \leq t\}.$$

这与函数 f 的图像中位于高度 t 或以下部分的连通分支数一致。给定函数的所有次水平集构成一个持久模，其在 t 处的取值为向量空间 $H_0(L_t(f); \mathbb{R})$，其中 H_0 表示零维同调，系数取自实数域 $\mathbb{R}$ 。

关于持久模及其相关条形码和持久图的背景知识，请参阅 [12]。条形码 描绘了每个次水平集连通分支的存续时间：从它出现时的高度 $t = b$ （诞生），到它与其他某个连通分支合并时的高度 $t = d$ （消亡）。相应的持久性图中，对于每条从 b 开始并在 d 结束的条形码线段，都包含一个点（b, d）。出于实际考虑，舍弃描绘存续至无穷远的本质同调类的无限射线。

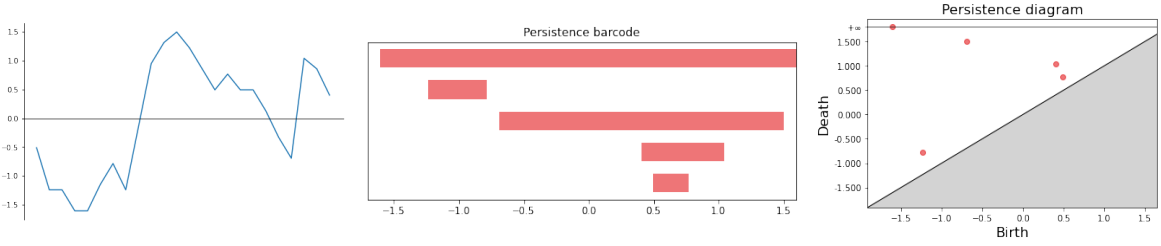


图 1：从左至右依次为分段线性函数、其关于次水平集的零维同调条形码，以及相应的持久性图。

条形码或持久性图并不适合用于深度学习。神经网络在处理数组形数据时表现最佳。本文使用景观 作为持久性摘要。

持久性景观在 [3] 中得到定义；对于信号的次水平集，它们通过一系列有限的、具有紧支撑的连续分段线性函数来表达连通分支的演变。从计算角度看，每个景观函数都可以表示为一组离散化数值，因此适合输入深度学习系统。

与一个持久性图相关联的景观函数序列定义如下。对于持久性图中的每个点（b，d），考虑相应的帐篷函数

$$\Lambda_{(b,d)}(t) = \max\{0, \min\{t-b, d-t\}\}.$$

接下来，为每个 $k \geq 1$ 定义一个分段线性函数 λ_k：$\mathbb{R} \to \mathbb{R}$

$$\lambda_k(t) = \mathrm{kmax}\{\Lambda_{(b,d)}(t)\},$$

其中，kmax 返回给定实数集合中第 k大的值，集合中的元素按重数计数；如果不存在第 k大的值，则返回零。因此，由于持久性图中的点数有限，对于所有足够大的 k，$\lambda_k = 0$ 均成立。前几个景观层级 λ_1，$\lambda_2 \ldots$ 描绘了持久性最强的拓扑特征，而最后几个层级则对应持久性较弱的现象。

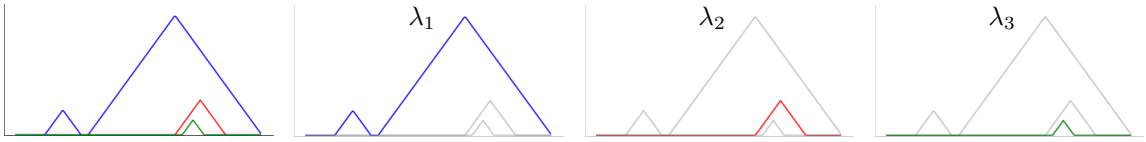


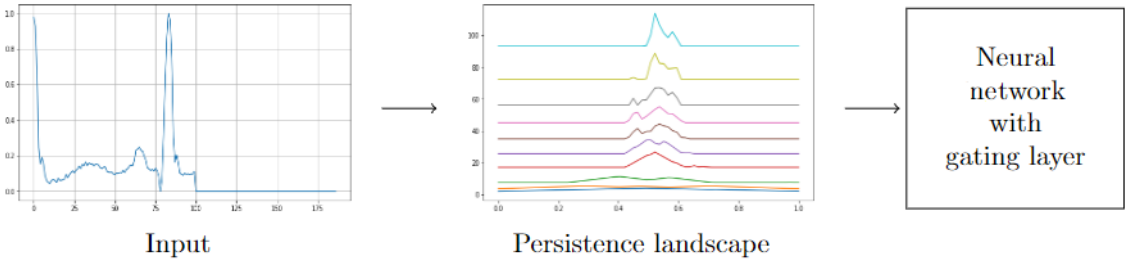
图2：持久景观的非零层级序列 λ_k （左）。

2 重要性归因

持久性景观可分层为层次序列，这一特点使得我们能够设计一种机制，为给定信号样本的景观层级进行重要性归因排序。在 [14] 中，介绍了一种确定性过程，用于从若干选定方向的方向持久景观重建信号。 [14] 中还展示了如何仅使用已选景观层级的一个子集，对给定信号进行部分重建；这正是本文关注的重点。将这一过程与机器学习相结合，为景观分配一系列权重后，我们能够大幅减少给定数据函数的临界点数量，同时几乎不损失分类准确率。

为此，我们将给定信号的次水平集持久性所对应的景观函数堆叠成一个矩阵，并将其输入神经网络。持久景观提供了一种便捷的表示方式，因为每个景观层级都对应输入信号振荡的不同区域。

由于我们的目标是输入一个深度学习模型，因此我们决定对每个景观函数下的面积进行归一化，使网络专注于其形态，而不是实际数值。图3展示了这一过程。



Fig图 3：通过持久景观提取信息并输入神经网络 .

不同信息层级的存在自然引出了一个问题：对于分类任务而言，哪些层级比其他层级更重要？为了实现这一想法，我们提出使用门控层：在整个架构中保持矩阵形状不变，并在应用全连接层之前，将每个景观层级 λ_k 乘以一个小于 1 的正可学习权重 w_k。由此，我们得到一组权重，用以指示每个景观层级对分类任务的影响程度。通常，网络应将前几个景观层级视为比后几个更重要，因为前几个层级包含有关最持久拓扑特征的信息。

通过对所有景观层级进行排序，我们能够确定网络在哪个信息阈值处停止学习。这主要有两方面的好处：首先，通过减少传递给网络的景观函数数量，我们能够减少用于训练系统的信息；其次，我们能够将重要性归因于原始数据中产生最相关景观层级的部分。

3 实验设置与结果

本节介绍我们使用具有固定架构和不同输入信号的神经网络进行实验所得到的结果。我们的主要目标是评估仅使用一组已选景观层级时的分类准确率变化，并将其与使用完整景观及原始数据的情况进行比较，同时确定每个数据库中最相关的景观特征。我们将该方法应用于九个性质迥异的数据库，以评估其稳健性。

数据。 我们将该方法应用于从 UCR时间序列分类档案 [9]中选取的数据集。数据集的选择标准如下：数据集最多包含五个不同类别，且样本总数除以类别数应大于或等于 500。采用这些标准是为了避免数据稀缺问题，以及类别不平衡或类别过多所造成的困难。表1汇总了各数据集的特征。

数据集	样本	长度	类别	不平衡
ECG5000	5000	140	5	Yes
FreezerRegularTrain	3000	301	2	No
HandOutlines	1370	2709	2	Yes
ItalyPowerDemand	1096	24	2	No
MoteStrain	1272	84	2	No
PhalangesOutlinesCorrect	2658	80	2	Yes
StarLightCurves	9236	1024	3	Yes
Wafer	7164	152	2	Yes
Yoga	3300	426	2	Yes

表1：各数据集特征摘要。对于每个数据集，我们列出了样本总数、每个样本的长度、类别数，以及该数据集是否不平衡。

方法。 为了避免数据集之间数值范围不同导致方法的准确率出现差异，我们已将输入函数标准化，使其取值介于 0 和 1 之间。此外，应用拓扑预处理时，我们对景观进行归一化，使每个景观函数下方的面积等于 1。这样做可以让神经网络研究景观的形状，而不只是考虑其实际数值。

本研究的主要目标是比较景观层级与仅经过标准化预处理的原始数据基线相比，在捕获信息方面的能力。此外，为评估已选景观层级是否足以进行分类，我们比较了将完整景观输入神经网络所得的结果，以及仅使用已选层级所得的结果。

神经网络的架构如下：三个卷积层与保行最大池化层相结合，后接两个全连接层（图4）。门控层用于选择和归因，仅在以景观层级作为输入时才会出现。在这种情况下，门控层位于最后一个最大池化层与第一个全连接层之间。实验采用5折交叉验证。训练集占每个数据集的80%。神经网络训练240轮，初始学习率为0.01，每100轮除以5。我们选择这一架构是为了使其较为通用，并未试图在原始数据或景观方法上实现尽可能高的准确率。我们的目的是评估该方法的有效性，同时避免因针对性选择最优架构而引入可能的特殊性。

至于性能指标，本文仅考虑准确率。

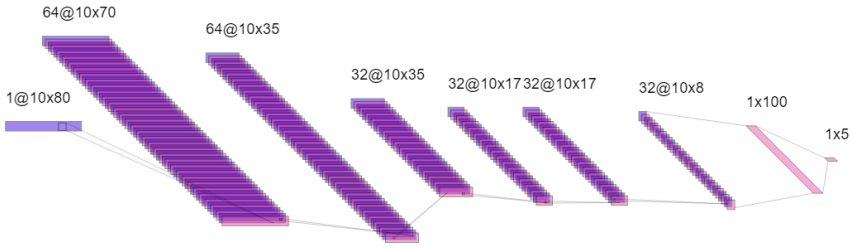


图4：本研究设计的神经网络架构。使用景观作为输入时，门控层位于第一个全连接层（粉色）之前。

3.1 方法验证

3.1.1 性能结果

我们对来自 [9] 的9个不同数据集进行了相同的实验，以验证结果的稳定性（表2）。对于每个数据集，我们都运行了一个神经网络（图4），并使用三种不同的输入：原始数据、由持久景观层级构成的序列，以及已选层级的子集。由于完整的非零景观层级序列长度不固定，我们选取了前10个层级 λ_1，……，λ_{10}，因为在大多数情况下，第10个层级已经为零；而固定使用更多景观层级会在训练过程中造成内存困难，却无法显著提高准确率。

随后，通过选择门控层提供的权重最高的层级，选出了数量较少的主要景观层级。根据数据集的不同，已选层级的数量为2至5个（图5）。关于如何选择合适的景观层级子集，详见§ 3.1.2。

表2展示了使用5折交叉验证进行的各项实验的平均准确率和标准差。表中列出了使用原始数据、未归一化景观、归一化景观以及所选归一化景观层级子集所获得的平均准确率结果。结果表明，景观能够取得足够高的分类准确率，尤其是在进行归一化后（第三列和第四列）。就此而言，在九个数据集中有四个数据集的景观准确率与使用原始数据所得的准确率在一个标准差范围内具有统计可比性。

在表2中，使用粗体突出显示了基于TDA的策略所得到的、彼此具有统计可比性的结果——包括取得极大准确率的方法。未归一化景观在大多数情况下始终会遗漏相关信息，导致准确率显著下降。同样值得注意的是，已选景观层级的表现与完整的（归一化）景观相近。这进一步支持了这样一种假设：数据中包含的大部分信息都可由少量景观层级捕获。

在 PhalangesOC 数据集中，归一化景观的表现甚至优于原始数据。正如讨论中所指出的，这可能是由于景观表示本身具有弹性形变不变性。

数据集	原始数据	未归一化	归一化	已选	No.
ECG5000	94.72 ± 0.7	$\mathbf{92.96 \pm 0.5}$	$\mathbf{93.12 \pm 0.4}$	$\mathbf{92.88 \pm 0.3}$	3
FreezerRT	99.53 ± 0.4	63.70 ± 4.2	$\mathbf{88.97 \pm 0.3}$	$\mathbf{88.70 \pm 1.6}$	2
HandOutlines	89.20 ± 1.7	75.77 ± 4.6	$\mathbf{85.26 \pm 2.0}$	81.90 ± 2.0	5
ItalyPowerD	97.73 ± 1.1	87.18 ± 3.4	$\mathbf{89.45 \pm 1.7}$	$\mathbf{90.36 \pm 1.3}$	2
MoteStrain	90.98 ± 1.1	71.84 ± 1.8	$\mathbf{77.33 \pm 2.8}$	$\mathbf{76.31 \pm 1.7}$	3
PhalangesOC	64.02 ± 2.1	63.95 ± 4.2	$\mathbf{68.95 \pm 0.9}$	$\mathbf{69.21 \pm 0.8}$	4
StarlightCurves	95.70 ± 0.4	89.82 ± 2.3	$\mathbf{94.92 \pm 0.1}$	$\mathbf{95.22 \pm 0.5}$	3
Wafer	99.65 ± 0.2	90.86 ± 0.8	$\mathbf{98.63 \pm 0.3}$	$\mathbf{98.79 \pm 0.4}$	3
Yoga	82.48 ± 1.6	64.21 ± 4.1	75.36 ± 1.1	$\mathbf{78.33 \pm 1.0}$	4

表2：针对九个信号数据集，神经网络（图4）运行五次后在测试集上得到的平均准确率（以百分比表示）及标准差。原始数据（第一列）所得的准确率，与前10个景观层级在未进行面积归一化（第二列）和进行面积归一化（第三列）时所得的准确率，以及最具信息量的景观层级（第四列）所得的准确率进行比较。最后一列显示了每种情况下选取的景观层级数量。基于TDA的策略中，在统计上具有可比性的准确率以粗体显示。

3.1.2 景观层级排序

我们方法的关键在于确定哪些景观层级包含最多与分类结果相关的信息。门控层将每个景观层级 λ_k（其中有$k = 1$，.....，10）乘以一个可学习权重 w_k with $0 \leq w_k \leq 1$。神经网络完成完整训练过程后，所得权重将用于评估每个景观层级的重要性。

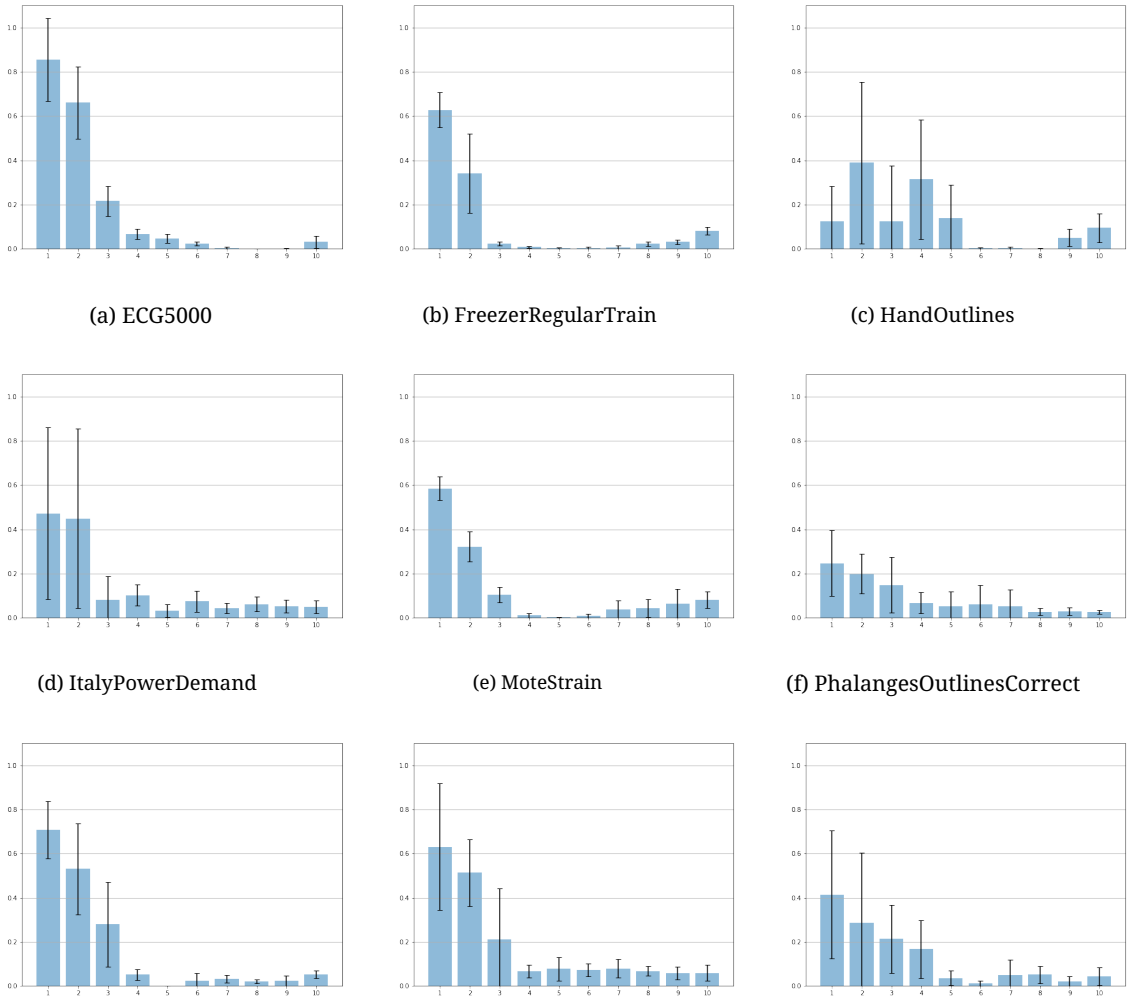
为确保结果具有显著性，我们进行了五次实验，并记录了每个景观层级的平均权重值和标准差，如图5所示。尽管没有明显的数值方法可以根据权重确定应视为重要的景观层级数量，但我们采用了以下标准。如果 $w_k < \frac{1}{2} w_{k-1}$ for some k，我们将 k 称为显著下降。如果 k 是最大的显著下降，且 $w_{k-1} > 0.1$，那么我们选择 λ_1，.....，λ_{k-1}作为最重要的景观层级。如果不存在满足 $w_{k-1} > 0.1$的显著下降，则选取最小的 k，使得 $w_1 + \cdots + w_{k-1} > w_k + \cdots + w_{10}$，并同时选择 λ_1，.....，λ_{k-1}。

除了极少数情况外，网络认为最初的景观层级更为重要。这些层级包含每个信号中持续时间最长的拓扑特征信息（次水平集的连通分支）。所有实验均使用了前10个层级。在某些情况下——即图5中的(d)和(f)——对于数据集中的所有样本，λ_k 与 $k > 6$ 对应的景观层级均为零。在这些情况下，门控层会为这些空层级分配较小但不一定为零的权重。

值得注意的是，末端景观层级（即本研究中的第10层）往往始终比紧邻其前的层级更为重要，但对于整个数据集都为零的情况除外。这表明末端景观层级可能包含判别信息，值得进一步研究。

图5显示，对于某些数据集，所有权重均低于0.4，具体来说是(c)和(f)，(i)的情况也大致如此。查看表2可知，这些数据集恰好是在神经网络使用原始数据训练后，测试集准确率低于90%的那些数据集。原始数据取得较高分类准确率的数据集，也正是重要景观层级数量最少的数据集。事实上，图6 显示，准确率与已选景观层级数量之间呈反向关系。

作为不利案例的示例，我们现在讨论使用 [9] 中的 FordA 和 TwoPatterns 数据集所获得的结果。这些数据集具有一个共同特性，即它们由波长不断变化的波状信号组成，而对其进行分类的关键信息在于波浪发生变化的位置，即 x-坐标。在其中一个数据集（FordA）中，原始数据很难



(a) ECG5000 (b) FreezerRegularTrain (c) HandOutlines

(d) ItalyPowerDemand (e) MoteStrain (f) PhalangesOutlinesCorrect

(g) StarLightCurves (h) Wafer (i) 瑜伽

图5：使用带门控层的神经网络（图4）运行五次后，九个数据集的前十个景观层级的平均权重和标准差。

进行分类，而在另一个数据集（TwoPatterns）中，原始数据则很容易分类。在这两种情况下，用持久性景观替代数据都会抹去对神经网络分类器而言相关的信息——因为在振幅保持不变的情况下，景观对波长变化具有不变性——因此，如果使用景观而非原始数据，我们会得到较低的准确率和显著的过拟合。

从图 7 可以看出，对于 FordA 数据集（即使使用原始数据，神经网络也难以进行分类），持久景观层级的权重都相近，且相关性较低。相比之下，在 TwoPatterns 的情况下，前几个景观层级呈现出明显的排序。因此，即使在不利的情况下，景观选择也能提供有关数据集的有意义信息，因为神经网络对原始数据的正确分类能力与我们的方法所发现的重要景观层级数量之间存在稳定的反向关系。总之，图 6 和图 7 表明，景观层级选择的结果与神经网络的表现相关。

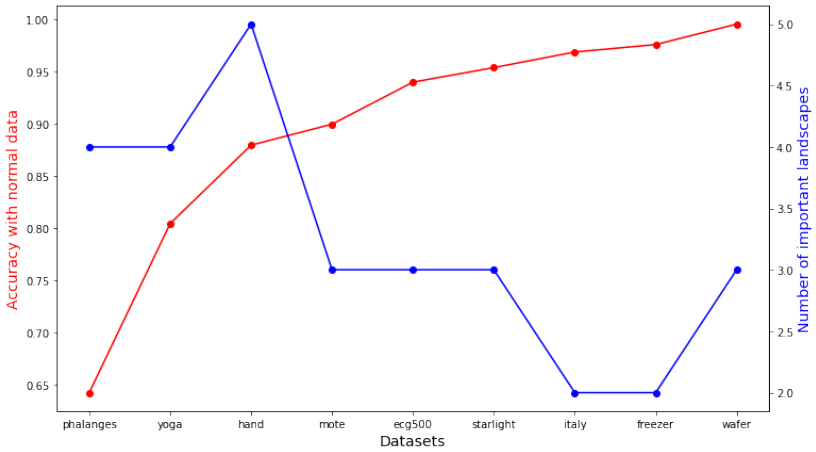
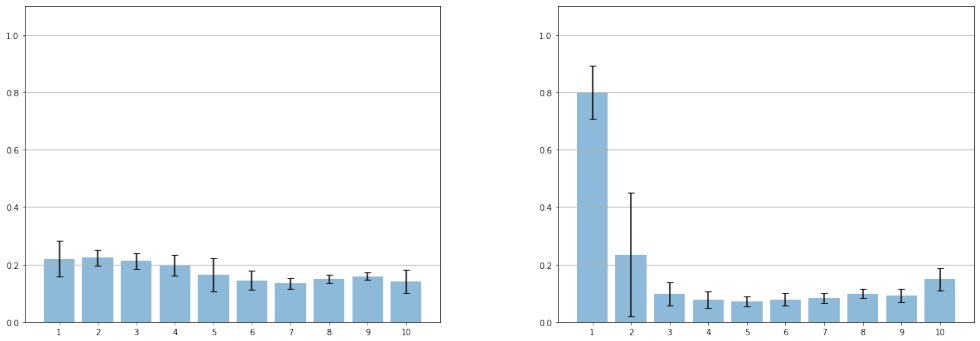


图 6：使用原始数据训练的神经网络准确率（红色）与被选为重要的景观层级数量（蓝色）呈反比关系。横轴上的数据集按准确率递增排序。



(a) FordA　　(b) TwoPatterns

图 7：两个数据集经过五次配备门控层的神经网络（图4）运行后，前十个景观层级的平均权重和标准差。

3.2 一个用例：心跳分析结果

作为一个应用案例，我们使用算法对 MIT-BIH 心律失常数据库 [19] 中的心电图信号（ECG）进行分类，以评估心律失常检测器。该数据集可从 [11] 获取，其中包含 48 段各半小时的心电图记录，记录来自 47 名受试者（25 名男性，年龄为 32 至 89 岁；22 名女性，年龄为 23 至 89 岁），采集时间为 1975 年至 1979 年。我们的样本包含 87,554 个心搏，分为五类：一类为正常心搏（ 82.77% ）；三类对应不同的心律失常类型，分别为室上性早搏（2.54%）、室性早搏（6.61%）以及室性与正常心搏融合（0.73%）；另有一类为无法识别的心搏（7.35%）。

表3展示了经过5折交叉验证后的平均准确率。将输入原始未处理信号的神经网络（图4）的分类准确率（98.41%），与采用相同架构、使用10层景观（94.55%）以及仅使用最重要的三个景观层级（94.00%）时的准确率进行了比较。由于表2表明归一化景观优于未归一化景观，因此对景观进行了面积归一化。选择三个层级是依据网络赋予的权重，如图8所示，其中 $k = 4$ 是最大的显著下降。

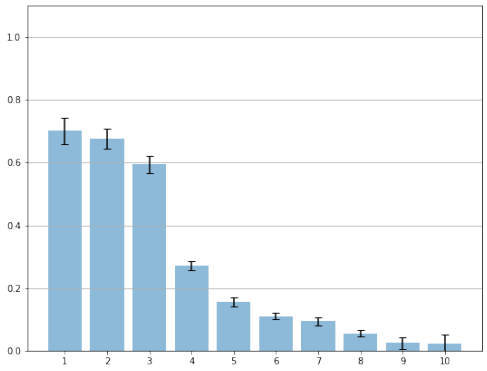


图 8：对 MIT-BIH 心律失常数据库中的一个样本进行五次神经网络运行后，前十个景观层级的平均权重和标准差（图4）。

接下来，我们在四个示例中使用了 [14, 第 3] 节详细介绍的部分重构技术，这些示例分别对应以下类别：(a)正常心搏，(b)室上性早搏，(c)室性早搏，以及(d)室性与正常心搏融合。每种情况下均使用三个景观层级进行近似。结果如图10所示。

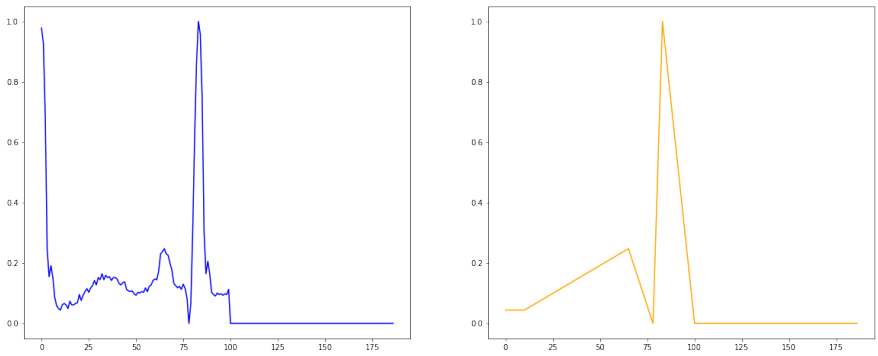
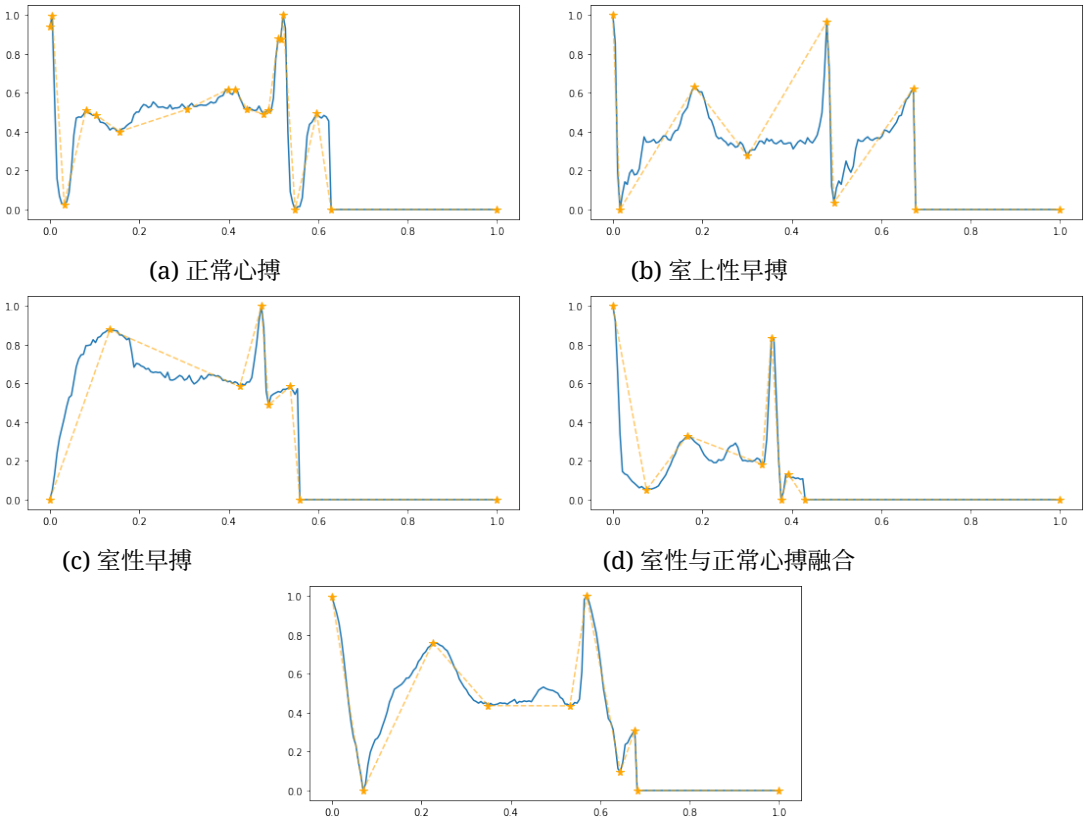


图9：心电信号函数（左）及其由一组已选景观层级近似重构的结果（右）。

每个景观函数 λ_k 都与给定信号f 临界点的 y 值列表配对，具体如 [14, 命题3.1]所述。因此，我们得到一个与 f相关联的临界点 y 值列表，这些临界点对应于已选景观层级的子集。为获得匹配的 x 值，我们将该列表中的值与 f 的所有临界点列表进行了比较，并按照所得 f 临界点的 x 坐标顺序将它们连接起来，绘制出一幅新的图像，如图9所示。该过程详见算法1、2和3。所得的简化图（图10）标出了根据我们实验中所用神经网络确定的关注点，这些点用于心电图样本的分类。因此，它们编码了网络为完成分类任务所重点关注的最相关信息。

	原始数据	**10 个层级**	**3 个层级**	**重构结果**
准确率	98.41 ± 0.09	94.55 ± 0.16	94.00 ± 0.13	97.04 ± 0.14

表3：输入未处理数据、具有十个景观层级的处理后数据、使用最显著的三个景观层级处理后的数据，以及通过三个景观层级近似重构的数据时，我们的神经网络（图4）的分类准确率（以百分比表示）。



(a) 正常心搏　(b) 室上性早搏

(c) 室性早搏　(d) 室性与正常心搏融合

(e) 无法分类的心搏

图10：使用四种心跳类型各自最重要的三个景观层级，对心电图图像进行部分重建。

随后，我们将波函数的简化重建结果（图9）输入网络，以检验重建方法提炼出的数据特征是否足以支持网络的分类任务。结果见表3，表明简化信号取得了与原始数据相近的准确率（97.04% 对 98.41%）。

3.3 平移不变性

持久性摘要不受信号水平平移的影响，因此，基于景观的分类任务的准确率不会因这类平移而改变。然而，平移可能会导致以原始数据为输入的神经网络的分类准确率下降。为展示这一现象，我们使用了与§ 3.2相同的心电图数据集，但通过在每个心搏信号的开头和结尾随机分配若干个零，对每个心搏进行了修改。因此，原始数据集中每个心搏均由长度为187的向量表示，而在我们的实验中，我们加入零，使其长度增加到374。

	原始数据	**双倍长度**
准确率	98.41 ± 0.09	96.77 ± 0.10

表4：我们的神经网络在未修改数据与修改后数据上的准确率（以百分比表示）。修改后的数据是在每个信号的开头和结尾插入零值片段，使信号长度加倍（第二列）。

使用与§3.2相同的神经网络，对平移后的心电图图形进行分类，并重复五次，所得准确率低于使用原始数据时的准确率（表4）。然而，平移不会改变图形的次水平集的连通分支的演化，因此，与平移后的图形相关联的景观与原始数据的景观相同。这表明，在实际情况下，若数据信号的平移在本质上并不显著，却仍可能被神经网络误解，那么使用持久性描述符可能会带来优势。

4 讨论

本文展示了拓扑数据分析在机器学习中的一种应用，具体而言，是在神经网络结果可解释性领域的应用。我们的方法能够从给定数据中提取出足够相关的部分信息，用于分类，同时不会显著降低准确率。我们使用景观来描述信号次水平集的持久性，因为景观具有层级结构，可借助神经网络中门控层分配的权重对各个层级的重要性进行排序。

我们证实，对于信号的准确分类而言，并不需要使用完整的持久景观：一旦确定了对网络最重要的景观层级子集，仅使用这些层级进行实验，所得准确率在统计上具有可比性（表2）。

本研究的一项创新是对景观函数进行归一化，使图像下方的面积始终等于一。此举旨在向神经网络输入形状而非幅值。如表2所示，使用归一化景观得到的准确率高于归一化前的结果。此外，在大多数情况下，归一化后的准确率标准差更低，这表明归一化提升了稳定性。

此外，我们的方法允许利用已选景观层级集合对给定信号进行部分重构，从而展示数据中的哪些特征对于所选架构的分类最为重要。一般而言，持久性描述符并非单射，无法用于恢复数据；只有在考虑一组方向持久图的某些情况下才可以做到这一点 [2, 5, 14, 21]。然而，我们当前的问题有所不同：我们不需要仅凭持久图就完整重构一个函数，而是要将持久性描述符归因于原始信息。因此，我们的重构任务包括将持久景观中的点与给定信号中相应的部分相匹配。实际上，信号临界点的 y 值由持久图中的点确定，如 [14, 第 1] 节所示。

由于我们对给定信号进行离散化，可能会出现数值精度方面的问题。在景观函数中搜索峰值时，信号中临界点对应的 y 值可以按照所选精度计算。将这些 y 值与原始函数进行比较时，我们不能期待差值为零，因为我们处理的是近似值。相反，我们必须采用低于某个阈值的值。为了避免选取与正确临界点不对应的 x 值，每当我们得到一个候选 x 值时，都会检查原始函数，确认它确实来自一个极小值或极大值。

我们的方法的一个特点是，信号次水平集的持久性图无法捕捉数据沿 x 轴的分布信息，只能捕捉其沿 y 轴的分布信息。因此，诸如景观之类的持久性摘要对于 x 轴上的平移或尺度变化具有不变性；更一般地说，对于水平弹性变形也具有不变性。当周期函数或几乎周期函数的波长对分类至关重要时，这可能会成为使用持久同调的一个缺点，例如第 3.1.2 小节中的 FordA 和 TwoPatterns 数据集所示。然而，如果沿 x 轴扩展或收缩会产生不良影响，这种不变性也可能成为优势，例如 [10] 中的心动过缓和心动过速，或第 3.3 小节中的实验。因此，使用持久性摘要有助于消除因在不具备物理意义的 x 坐标上发生偏移而产生的伪影；例如，将时间序列分割成较短的信号时，或信号中出现随机水平片段时。

无论使用持久性描述符而非原始数据对性能指标有何影响，通过使用景观，我们都能洞察用于分类给定数据的关键模式，从而使这一过程更加值得信赖。因此，我们的方法不仅能够提供有关网络学习过程重点的信息，还可用于探索并更好地理解数据集。

参考文献

[1] Rub´en Ballester 等。“使用拓扑数据分析预测神经网络中的泛化差距”。载于：（2022）。arXiv：2203.12330 [cs.LG]。[2] Robin Lynne Belton 等。“从持久图重建嵌入图”。载于：《计算几何：理论与应用》90（2020），第101658页。 DOI：10.1016/ j. comgeo.2020.101658。[3] Peter Bubenik。“使用持久景观进行统计拓扑数据分析”。载于：《机器学习研究期刊》16（2015年1月），第77–102页。[4] Mathieu Carri`ere 等。“PersLay：用于持久图的神经网络层与新的图拓扑特征”。载于：《第23届国际人工智能与统计学会议论文集》。第108卷。《机器学习研究论文集》。PMLR，2020年，第2786–2796页。 URL：https://proceedings.mlr.press/v108/carriere20a.html。[5] Fr´ed´eric Chazal 和 Steve Oudot。“迈向欧几里得空间中的基于持久性的重建”。载于：《计算几何年度研讨会论文集》（2008年1月）。DOI：10.1145/1377676.1377719。[6] Chao Chen 等。“通过持久同调为分类器引入拓扑正则化项”。载于：《第22届国际人工智能与统计学会议论文集》。第89卷。《机器学习研究论文集》。PMLR，2019年，第2573–2582页。 URL：https : / / proceedings.mlr.press/v89/chen19g.html。[7] James Clough 等。“使用持久同调的基于深度学习的图像分割拓扑损失函数”。载于：《IEEE模式分析与机器智能汇刊》PP（2020年9月）。 DOI：10.1109/TPAMI.2020.3013679。[8] Ciprian A. Corneanu、Sergio Escalera 和 Aleix M. Martinez。“无需测试集即可计算测试误差”。载于：《2020 IEEE/CVF计算机视觉与模式识别会议》。2020年，第2674–2682页。 DOI：10.1109/CVPR42600.2020.00275。[9] Hoang Anh Dau 等。UCR时间序列分类档案。https://www.cs.ucr. edu/~eamonn/time_series_data_2018/。2018年。[10] Meryll Dindin、Yuhei Umeda 和 Frederic Chazal。“通过模块化神经网络进行心律失常检测的拓扑数据分析”。载于：《人工智能进展》。Cham：Springer International Publishing，2020年，第177–188页。DOI：10.1007/978-3-030-47358-7_17。[11] 心电图心跳分类数据集。用于心跳分类的分段与预处理心电信号。 URL：https://www.kaggle.com/datasets/shayanfazeli/heartbeat。[12] Herbert Edelsbrunner 和 John Harer。“持久同调——综述”。载于：《当代数学》453（2008年），第1–2页。 DOI：dx.doi.org/10.1090/conm/453。[13] Brittany Terese Fasy 等。“从持久图重建任意维度单纯复形的第一个算法”。载于：（2022年1月）。arXiv：1912.12759v4 [cs.CG]。[14] Aina Ferr`a、Carles Casacuberta 和 Oriol Pujol。“从方向持久图重建单变量函数”。载于：（2022年）。arXiv：2203.01894 [math.AT]。[15] Rickard Br¨uel Gabrielsson 等。“用于机器学习的拓扑层”。载于：《第23届国际人工智能与统计学会议论文集》。第108卷。《机器学习研究论文集》。PMLR，2020年，第1553–1563页。 URL：https://proceedings.mlr.press/v108/gabrielsson20a.html。[16] Daniel Goldfarb。使用拓扑数据分析理解深度神经网络。2018年。arXiv：1811.00852 [cs.LG]。[17] Rocio Gonzalez-Diaz、Miguel A. Guti´errez-Naranjo 和 Eduardo Paluzo-Hidalgo。“用于减少神经网络训练资源的基于拓扑的代表性数据集”。载于：《神经计算与应用》34（2022年），第14397–14413页。 DOI：10.1007/s00521-022-07252-y。[18] Kwangho Kim 等。“PLLay：基于持久景观的高效拓扑层”。载于：《第34届神经信息处理系统大会》。2020年。

[19] George B. Moody 和 Roger G. Mark。“MIT-BIH 心律失常数据库的影响”。载于：*IEEE* 医学与生物学工程杂志 20.3（2001），第 45–50 页。 DOI：10.1109/ 51.932724。[20] Gregory Naitzat、Andrey Zhitnikov 和 Lek-Heng Lim。“深度神经网络的拓扑学”。载于：《机器学习研究期刊》21.184（2020），第 1–40 页。 URL：http://jmlr.org/ papers/v21/20-345.html。[21] Katharine Turner、Sayan Mukherjee 和 Doug M. Boyer。“用于形状与曲面建模的持久同调变换”。载于：*Information and Inference*：*IMA* 杂志 3.4（2014），第 310–344 页。 DOI：10.1093/imaiai/iau011。[22] Yuhei Umeda、Junji Kaneko 和 Hideyuki Kikuchi。“拓扑数据分析及其在时间序列数据分析中的应用”。载于：《富士通科学技术杂志》55（2019 年 1 月），第 65–71 页。[23] Matthew Wheeler、Jose Bouza 和 Peter Bubenik。“作为神经网络性能拓扑摘要的激活景观”。载于：*2021* 年 *IEEE* 国际大数据会议。IEEE，2021。 DOI：10.1109/bigdata52589.2021.9671368。[24] Yan Yan 等。“基于持久景观的个性化心律失常拓扑数据分析”。载于：预印本 （2019）。 DOI：10.20944/preprints201908.0320.v1。

算法 1： 获取临界点的 y 值

输入： 景观数据的 l 向量，景观数据的 t 坐标向量；
输出： 临界点 y 值的列表；

批评 $\leftarrow$ {};
对于 $i \leftarrow 0$ **到** $|l|$，**执行**
  如果 l_i 是一个起飞顶点 **则**
    临界点 $\leftarrow$ 临界点 $\cup \{t_i\}$;
  否则如果 l_i 是局部最小值或局部最大值 **那么**
    临界点 $\leftarrow$ 临界点 $\cup \{t_i\}$;
    临界点 $\leftarrow$ 2(临界点$_i$ $-$ 0.5临界点$_{i-1}$);
结束 i
返回临界点

算法 2:获取临界点的 x 值

输入： 临界点的 y 值列表 crits，$original$ 原始数据向量，x 向量的 x原始数据的坐标；
输出： 临界点的 x 值列表；

candcrits $\leftarrow$ {};
对 $crit \in$ 临界点 **执行**
  $compare \leftarrow (original - crit)^2$;
  对 $i \leftarrow 0$ **到** $|compare|$ **执行**
    如果 $compare_i < 10^{-4}$ **则**
      候选临界点 $\leftarrow$ 候选临界点 $\cup \{i\}$;
  end
结束
xcrits $\leftarrow$ {};
对于 $i \in$ candcrits **执行**
  如果 $original_i$ 是局部最小值或局部最大值 **则**
    xcrits $\leftarrow$ xcrits $\cup \{x_i\}$;
结束
返回xcrits

算法 3： 与一组景观层级相关的函数临界点子集

输入： 景观（景观层级列表），$original$ 原始数据向量；
输出： 给定函数的选定临界点列表；

临界点 $\leftarrow$ {};
对于 $land \in$ 景观 **执行**
  $candcrits \leftarrow$ {获取 y 值($land$)};
  临界点 $\leftarrow$ 临界点 $\cup$ {获取 x 值($candcrits, original$)};
结束
临界点 $\leftarrow$对临界点排序;
返回临界点