arXiv:2301.08379v1 [cs.LG] 20 Jan 2023

异步训练的分布式拓扑映射

Abbas Siddiqui [1] **Dionysios Georgiadis** [2]

摘要

拓扑特征映射是数据的低维表示，能够保留空间依赖关系。当前训练此类映射的方法（例如自组织映射 - 自组织映射、生成式拓扑映射）需要集中控制和同步执行，这限制了可扩展性。我们提出一种算法，利用N 自主单元通过分布式异步训练生成特征映射。单元自主性通过在时间和空间上进行稀疏交互来实现，具体结合了分布式启发式搜索和级联驱动的权重更新方案，该方案由两条规则支配：单元 i）在接收到样本或邻居的权重向量时进行适应；ii）在适应达到预先设定的次数后，将其权重向量广播给邻居。因此，一次向量更新就可能触发一连串的适应。我们将这种连锁反应映射到统计力学模型，从而能够对级联的统计特性进行参数化。我们使用 MNIST，实证研究启发式搜索准确率和级联参数对映射质量的影响。我们还提供实证证据表明，算法复杂度随系统规模增长至多呈线性关系 N。研究发现，所提出的方法在多个数据集上的分类任务中，性能与类似方法相当。

1. 引言

拓扑映射是对高维数据进行低维离散表征。这些映射由神经元（又称单元）组成，并通过预先定义的规则晶格相互连接。每个神经元 j 都包含对数据 w_j的一种表征。训练此类映射时，目标有两个：i）每个神经元的表征 w_j 应近似数据分布的某个区域；ii）相邻神经元应包含相似的表征（遵循映射的拓扑）。此类映射已广泛应用于数据分析（主要用于聚类、函数逼近和降维；更多示例请参见（Kohonen，2013；Kohonen 等人，2001））），通常通过自组织映射算法（SOM）进行训练。

机器学习算法对可扩展性的需求日益增长（这源于数据可访问性的提高（Qiu 等人，2016））。然而，自组织映射需要集中式训练，这会损害其可扩展性。集中式训练的必要性源于最佳匹配单元（BMU）搜索（这是一个集中式过程，需要将样本与每个单元的表示进行比较 w_j）以及单元适配过程（在此过程中，中心实体掌握映射中所有单元的信息）。为应对可扩展性挑战，人们提出了许多变通方案，包括并行执行、专用硬件加速（使用图形处理器（Liu 等人，2018；Wang 等人，2017）或现场可编程门阵列（Girau & Torres-Huitzil，2018）），以及启发式方法（Kohonen 等人，2000）。然而，至关重要的是，自组织映射的分布式实现（Sarazin 等人，2014；Liu 等人，2018）依赖迭代式 MapReduce框架（Spark、HADOOP）。这类框架要求同步执行，因此容易受到网络延迟或慢速工作节点（即拖后腿的节点）的影响。异步训练方法因不受这些限制而受到关注（Mnih 等人，2016；Li 等人，2019；Fernando 等人，2017）。

为了解决可扩展性问题，我们提出一种通过异步执行来训练拓扑映射的算法。我们将映射单元转换为自主智能体，从而实现这一目标。每个智能体仅掌握特征映射的有限信息（少数几个邻居），并与同级节点进行稀疏交互。搜索和调整过程均为异步执行，且只需利用每个单元的预定义邻居的局部知识。

训练过程对每个样本 s_i都包含两个步骤。首先，通过启发式搜索找到一个良好匹配单元（GMU），然后 GMU 调整自身 w_j 以适应该样本——这可能会引发适应级联（如下文所述）。为了有效探索，搜索依赖于单元之间的非局部连接，这些连接称为远程连接。此外，搜索还会使用一组局部连接，这些连接位于

[1]德国巴伐利亚 [2]Aspect Capital，英国伦敦。联系：Abbas Siddiqui <ab- bas.siddiqui@gmail.com>、Dionysios Georgiadis <dionys- ios.georgiadis@aspectcapital.com>。

单元之间，称为近邻连接，用于贪心利用。适应级联的可能性使映射能够在训练过程中保留其拓扑。级联由一条简单规则实现：每当一个单元调整自身时，都可以利用其近邻连接，进而可能触发其他单元也进行调整。因此，一次调整可能触发另一次调整，后者又可能触发下一次调整，如多米诺骨牌般依次发生。

我们认为，前一段所述的方案可以作为一种通用方法，以异步方式训练保持拓扑结构的特征映射。我们给出了该方案的一个简单实现，并提供了经验证据来支持这一观点。我们的主要贡献在于展示了一种高度可扩展的拓扑映射异步训练算法。

为了初步研究这一方案并证明这一主张的可行性，我们给出了启发式搜索、级联规则以及创建近程和远程连接规则的基本实现。在第 2 节中，我们使用 MNIST 来探索所提出算法的行为，并据此得到一组建议的超参数。为了证明所提出算法的超参数配置具有适用性和稳健性，我们在第 4 节中使用该算法对不同数据集（包括 MNIST）进行分类。我们还能够在所提出的超参数配置下推导出该算法的计算复杂度。我们将所得分类性能与文献中报告的自组织映射（SOM）结果进行比较，发现两种方法的表现相当。3

2. 算法

每个 N 单元在两个空间中都有一个位置：单元空间和样本空间。在单元空间中，单元按规则格点排列，$\{0, ..\sqrt{N}\}^2$。单元在样本空间中的 j 位置由权重向量 w_j 给出。作为级联机制的一部分，每个单元都有一个计数器 c_j（初始化为 0），所有单元共用一个级联阈值 θ。通过训练，我们旨在获得权重w_j，使得：**i)** 两个空间（样本空间和单元空间）中单元的拓扑结构相似，并且**ii)** 权重能够准确近似样本空间中样本的分布。

连接根据单元在单元空间中j，k 的曼哈顿距离绘制 D_{jk}。

- **近邻连接**用于启发式搜索和适应。如果 $D_{jk} \leq 1$，则绘制这些连接，从而形成方形晶格。通过近邻连接与单元 j 相连的单元是 j的近邻，记为 $\mathcal{N}_j$。

- **远程连接**仅用于启发式搜索。它们通过以概率[1]$\sim D_{jk}^{-1}$ 将每个单元 j与固定数量 ϕ 的同级节点相连，以概率方式绘制这些连接。通过远程连接与单元 j 相连的单元称为远邻，记作 $\mathcal{F}_j$。关于 ϕ 选择方式的详细信息，请参阅相应章节。

映射训练包含两个过程：针对 GMU 的分布式启发式搜索，以及级联驱动的自适应。自适应仅依赖近邻连接，而启发式搜索同时使用近邻连接和远程连接。

2.1. 分布式启发式搜索

搜索旨在确定 GMU j^*，其方式类似接力赛：样本在各单元之间传递，以寻找能使以下目标最小化的 j^*：

$$q_{ij^*} = |w_{j^*} - s_i|. \tag{1}$$

搜索分为两个阶段。从一个随机单元 j（我们也将其设为 j^*）开始，执行以下步骤：

1. **随机探索**，通过随机均匀地选取一个远邻或 j，更新持有样本的单元的索引 j。样本被发送到该单元；如果$q_{ij} < q_{ij^*}$，则更新 GMU：$j^* \leftarrow j$。2. 重复步骤 1，共迭代 e 次（其中 e 是用户定义参数）。然后通过贪心搜索优化当前已知的最佳解 j^*。3. **贪心利用** 将样本与 j^*的近邻和远邻进行比较，并确定 k^*：

$$k^* \leftarrow \underset{k \in \mathcal{N}_j}{\text{argmin}}\ q_{ki} \tag{2}$$

4. 如果得到 $q_{ik^*} < q_{ij^*}$，则更新 GMU 索引 $j^* \leftarrow k^*$ 并重复步骤 3。否则，搜索终止，j^* 即为 GMU。

探索迭代次数 e（见步骤 2）可用于调整搜索的准确性。

为了量化此次搜索的性能，我们可以检查被识别为 GMU 的单元是否确实是最佳匹配单元（BMU），也就是说，是否为问题 $argmin_j\ q_j(i)$ 的全局最优解。如果搜索得到的 GMU 与 BMU 不一致，我们就称此次搜索失败。在接下来的小节中，我们通过计算训练后期搜索失败的比例来量化搜索性能。我们将这一数值称为搜索误差，记作 F。

[1]已知，这类连接概率参数化方式（基于距离的幂律）在涉及空间图的某些启发式搜索问题中表现最佳（Kleinberg, 2000），但未来的研究中可能会考虑不同的方案。

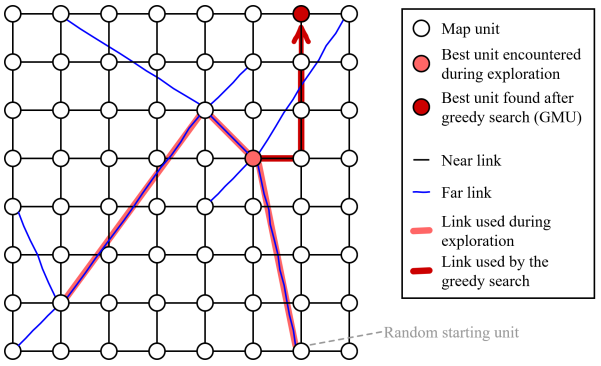

图 *1*。所提出的启发式搜索的可视化。单元与近邻和远邻相连（黑线和蓝线）。每个样本沿着连接跳转，尝试找到自己的最佳匹配单元。搜索分为两个阶段：沿远程连接进行随机探索（粉色高亮），以及沿近邻连接进行贪心搜索（红色高亮）。远程连接有助于有效探索（详见第 2.1 节）。图中仅展示了少部分远程连接。

由搜索不准确引发的问题，以及对 e 如何影响不同映射大小下的搜索准确率 N 的实证研究，可见下一节（参见小节3.1）。此外，搜索算法以伪代码形式呈现在算法 1中，并在图 1中进行了可视化。

请注意， $\mathcal{F}_j$ 中的远距离连接可使单元空间中的搜索在贪心搜索之前（以及期间）实现长距离跳转，从而帮助算法跳出局部极小值。充分增加每个单元的远程连接数量ϕ，可以让随机搜索过程迅速扩散至整个映射，在 $\mathcal{O}(\log(N))$ 次迭代内（Martel 和 Nguyen, 2004），从而确保搜索具有相对于映射大小的可扩展性。这种效应在图论中被称为小世界效应（Milgram, 1967），只需相对少量的 ϕ （Barri`ere 等人，2001）即可实现。此搜索算法足以用于演示，但未来也可以考虑更复杂的方案。

2.2. 级联驱动适应

确定 GMU j^* 后，即可进行适应。令 l_s为用户定义的常数学习率，权重向量按如下方式更新：

$$w_{j^*} \leftarrow w_{j^*} + l_s(s_i - w_{j^*}) \tag{3}$$

此外，GMU 的级联计数器会以概率 p_i 增加（通过应用 $c_{j^*} \leftarrow c_{j^*}+1$）。 p_i （称为级联概率）的具体参数化方式将在本节后文讨论。随后，级联适应可能会按照以下规则进行：

1. **触发：**如果级联计数器更新后得到 $c_j > \theta$，则称该单元触发：它会重置 $c_j \leftarrow 0$，并向所有近邻广播w_j。

2. **级联适应：** 单元 j 接收到权重向量 w_k后，会调整其自身的权重向量：

$$w_j \leftarrow w_j + l_c(i)(w_j - w_k) \tag{4}$$

其中， i 是最后一个样本的索引， $l_c(i)$ 是取决于训练索引 i的学习率。

3. **驱动：** 每次对 w_j 进行适应后，我们以 p_i的概率应用 $c_j \leftarrow c_j + 1$。

级联适应过程也以伪代码形式描述于算法1中。与样本 i 相关的级联事件规模通过级联规模 a_i 来衡量——其计算方法是在所有触发停止后，统计触发事件的数量。在以下分析中，我们通常会提及分数级联规模，其定义为 $A_i = a_i/N$。

每次触发事件都会使一个单元吸引样本空间中的近邻。这会提高映射的拓扑顺序，但也可能降低权重向量与样本的相似度。权重之间的吸引力可通过学习率 $l_c(i)$ 和 p_i来控制。为使映射在全局范围内有序，我们在训练初期允许大规模级联（规模为 $\mathcal{O}(N)$），并采用较高的 l_c。随后，在训练过程中逐步降低 l_c， p_i ，使权重能够更好地逼近样本。

级联学习率如下：

$$l_c(i) = \left(1 + tanh\left(\frac{c_o - i/i_{max}}{c_s}\right)\right) \Big/ 2 \tag{5}$$

其中， i_{max} 是训练迭代的总次数，c_o、 c_s 是用户定义常数。随着训练的进行，公式（5）使 $l_c(i)$沿平滑递减的斜率变化，同时确保 $l_c(i) \in (0,1)\ \forall i$。偏移参数 c_o 控制 l_c 在训练过程中何时达到0.5；例如，设定 $c_o = 0$ 或 $c_o = 1$ 时，分别会得到 $l_c(0) = 0.5$ 和$l_c(i_{max}) = 0.5$。参数 c_s 控制下降斜率： $c_s = 0$ 会使其瞬间从1降至0，而 $c_s \to \inf$则会使 l_c保持恒定。

操纵级联规模为了控制宏观级联行为，我们借助统计物理学中的方法对 p_i 进行参数化。 a_i 与 p_i 之间的统计行为（即概率分布 $\mathbb{P}(a_i = C; p_i)$）的依赖关系，可以在以下四个假设下进行解析理解：

- 在整个训练过程中，样本均匀地分布在单元空间中，

算法 1用于异步训练拓扑特征映射的所提方法伪代码——训练拓扑特征映射。

e：探索次数
l_s：样本学习率
θ：级联阈值
i_max：训练样本总数

函数 训练映射()：连接近邻() 连接远邻()**对于** *i*：*1* 到 *i* 最大值 **执行**样本 = 获取随机样本() bmu = 启发式搜索(样本) 调整权重(bmu, 样本, l s) 增加颗粒(bmu,i)**如果** 获取颗粒数*(bmu)* >= θ **那么**级联(bmu, i)**结束结束**

异步函数 启发式搜索($sample$)**：**最佳匹配单元 =获取随机单元() 随机邻居 =最佳匹配单元**对于** *1* 到 *e***执行**随机邻居 = 获取随机远邻(随机邻居) 临时值 = 获取距离(随机邻居, 样本)**如果**临时值< 最小距离 **则**最佳匹配单元 =随机邻居 最小距离 =临时值**结束 结束**临时值 = 获取最近邻居(最佳匹配单元, 样本)**当**临时值 < 最小距离 **执行**最佳匹配单元 = 邻居 最小距离 = 临时值 邻居 = 获取最近邻居(最佳匹配单元, 样本) 临时值=获取距离(邻居 , 样本)**结束 返回** 最佳匹配单元

异步函数级联 ($unit,\ i$)**：**清空颗粒(unit)**对于***unit***的**每个近邻 **执行**调整权重(neighbor, 获取权重(unit), getCas- cadingLearningRate(i))**如果***rnd* < 获取级联概率*(i)***则**添加颗粒(unit)**结束如果** 获取颗粒数*(neighbor)* >= θ **则**级联(neighbor, i)**结束 返回**

- 在训练过程中 p_i 的值缓慢变化（称为绝热近似），
- 展示映射与样本之间经过的时间足以让级联停止，
- 近邻连接的数量等于阈值 $\theta = |\mathcal{N}_j|$。

基于这些假设——并且对于 $p_i = 1$ ——级联可以精确映射为 BTW 沙堆模型（Bak 等人，1988）。对于 $p_i < 0$，级联可以映射为一种非保守变体，该变体基于成熟的沙堆模型，耗散以概率 $1-p_i$ 发生（Vespignani 等人，1998；Malcai 等人，2006）。在这一（耗散）变体中，已知级联规模服从幂律分布，并在特征分数大小 $\bar{\chi}$ 处呈指数截断。在耗散沙堆模型中，通常有 $\bar{\chi} \sim d^{-s}$，其中 d 为耗散率，s 为临界指数，具体取决于控制级联的规则（Vespignani 等人，1998；Malcai 等人，2006）。在我们的情形中，$s = 1$ （Vespignani 等人，1998），耗散由 $d \sim 1 - p_i$ 给出。因此，我们可以通过调整 p_i 来改变特征级联规模 $\bar{a}_i$。

为了研究算法在不同映射大小 N下的行为，有必要以一种能够产生尺度不变动力学的方式对 p_i 进行参数化：也就是说，级联 a_i/N 的相对大小与 N无关。对于足够大的 N，我们可以通过以下参数化实现尺度不变性：

$$p_i = \left(1 - \frac{1}{\sqrt{c_m N}}\right)\left(1 - \frac{i}{i_{max}}\right)^{\frac{c_d}{N}} \tag{6}$$

这一关系使 $\bar{a}_i$ 的初始值较高，随后随时间推移而减小。常数 $1/N << c_m < 1$ 控制训练早期的特征大小 $\bar{a}_i$，而 $c_d \in (0, \infty)$控制 $\bar{a}_i$ 随时间缩小的速率。上述行为与系统规模 N无关。

3. 实验分析

本节探讨模型超参数对生成映射质量的影响。我们的分析重点关注控制所提出算法的两个新颖组成部分的超参数：级联适应机制和分布式启发式搜索。

默认配置 除非另有说明，本节的所有实验均使用 MNIST 数据集和一个包含 900 个单元的映射，配置如下：远程连接数量 ϕ 设为 20（以确保网络具有稠密连接）。学习率参数设为 $l_s = 0.05$、$c_o = 0.5$ 和 $c_s = 0.5$。级联参数设为 $c_m = 0.1$ 和

$c_d = 100$。探索迭代次数设为 $e = 3N$。训练迭代次数与单元数量成正比（这也是与之密切相关的自组织映射算法的常见做法（Kohonen 等人，2000））。调整训练轮次数，使训练样本数量为$i_{max} \approx 600N$。

衡量映射质量 映射质量可以通过两种方式，使用两种不同且广为认可的指标进行评估。量化误差用于衡量权重向量对样本空间中样本分布的描述程度，而拓扑误差用于衡量映射的拓扑顺序（局部尺度上）（Li 等人，1993）。映射的量化误差和拓扑误差分别记为 Q 和 T。

3.1. 启发式搜索准确率

为了探索利用式搜索迭代次数参数 e 对最终映射的影响，我们使用递增的搜索迭代次数训练映射： $e \in \{0.01N, 0.05N, 0.1N, 0.2N, 0.3N \dots 5N\}$ 并比较得到的搜索准确率。为了量化搜索准确率，我们关注最后 1000 次训练迭代，并计算 GMU 与 BMU 不一致的次数所占的比例。该比例记为 F。

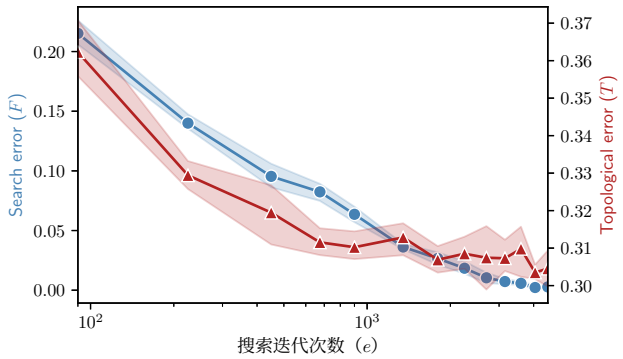

图 2。增加利用式搜索迭代次数 e 会渐近地降低搜索误差（蓝线）和拓扑误差（红线）。提高搜索准确率对映射质量的影响会逐渐减弱。置信区间以 1 std为置信水平，并基于5次运行计算。实验配置请参见第3 节第二段。

该实验的结果如图2所示，图中表明，在所考虑的 e范围内，启发式搜索的准确率呈指数增长。该图还描绘了映射的拓扑误差随 e变化的情况，表明进一步提高搜索准确率对拓扑质量的改善逐渐减小。基于这些结果，我们将在其余所有实验中将超参数$e = 3N$ 设为（因为 $3N$ 是最小的值。在搜索准确率 e 方面，结果为 $> 99\%$）。对于 $e = 3N$，关于搜索相对于 N的可扩展性将在后续实验中确立（见小节3.3），该实验表明，随着N 增加，映射质量也会提高。更详尽的分析请参见附录。

3.2. 级联适应

级联适应机制依赖于两个超参数，其中一个与平均级联规模有关
一个是在训练早期，另一个与级联规模在整个训练过程中下降的速率有关（分别对应 c_m， c_d ）。

为了评估这两个参数的影响，我们在以下参数构成的稀疏网格上训练映射：(c_m， c_d)，其中 $c_m \in \{0.01, 0.050.1 \dots 1\}$ 且 $c_d \in \{10, 10^2, 10^3, 10^4\}$。所得映射的拓扑误差和量化误差如图4 所示，涵盖两种映射大小。图 4 表明，量化误差和拓扑误差对 c_m并不敏感。因此，我们可以选择较小的 c_m，从而减小级联规模并降低计算量。然而， c_m 必须是$>> 1/N$ （见公式（6）），因此在其余实验中，我们折中选择 $c_m = 0.1$。

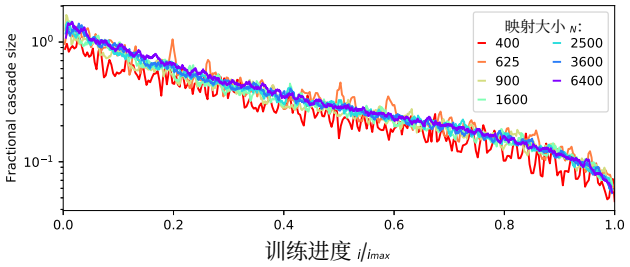

图 3. 增大映射大小 N 不会显著影响分数级联规模 A_i。我们跟踪训练期间 1000 分位数的上界A_i （采用宽度为 $i_{max}/100$ 的滚动窗口），并逐渐增大映射大小 N。跟踪到的轨迹收敛到同一条时间线上，表明 N 不会影响分数级联规模。实验配置见第 3 节第二段。

为了更仔细地考察另一项级联超参数 c_d 对算法性能的影响，我们使用逐渐增大的 c_d训练映射，并测量其量化误差和拓扑误差。图5展示了结果，表明 c_d 对映射质量的影响并不单一：增大 c_d 可能会降低量化误差，但代价是拓扑误差增加。因此， c_d 的合适取值可能取决于生成的映射的预期用途。我们分别选择 $c_d = 100$ 和 $c_d = 1000$ ，以突出较低的拓扑误差和量化误差。

级联由级联概率 p_i控制；如第2.2小节所述，该概率经过参数化，使级联行为能够自然地适应系统规模。具体而言，级联的分数规模 A_i 由 c_m控制， c_d 并与映射的绝对规模 N脱钩。我们通过训练规模不断增大的映射

$$N \in \{100, 225, 400, 625, 900, 1600, 2500, 3600, 6400\}$$

并比较由此产生的级联活动，对这一说法进行实证验证。为了比较不同规模映射的级联活动，我们 i）取宽度为 $i_{max}/100$的滚动窗口；ii）计算位于最高 A_i 分位数1000中的分数级联规模的均值。

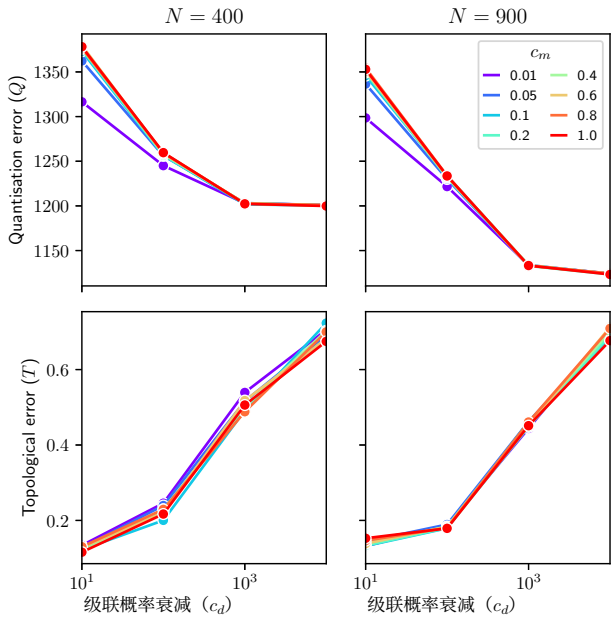

图 *4.* 对级联参数进行稀疏网格搜索 c_d， c_m结果表明， c_m 对映射的量化误差和拓扑误差影响极小（Q 和 T）。针对两种映射大小 N：400，900 （分别对应左列和右列）。实验配置详见第3 节第二段。

该分析结果见图3。图中所有映射大小对应的曲线重合，表明使用所提出的p_i参数化方式时，分数级联规模 A_i 不会显著依赖于系统规模——对于 $N > 400$。

3.3 可扩展性

为研究该算法的可扩展性，我们训练规模不断增大的映射 $N \in \{100, 225, 400, 625, 900, 1600, 2500, 3600, 6400\}$，针对 $c_d = 100$， $c_m = 0.1$， $e = 3N$。我们评估每个训练所得映射的质量（即量化误差、拓扑

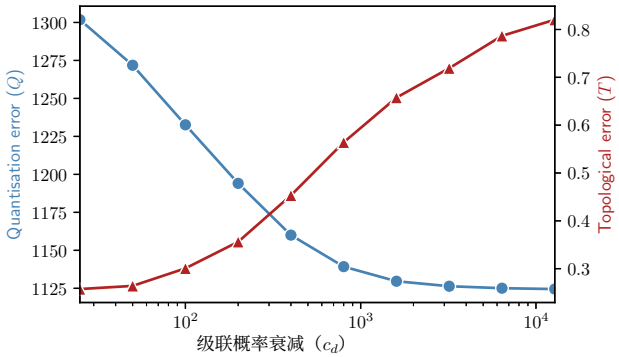

图*5.* 增大级联衰减参数 c_d 对映射质量的影响是复杂的：量化误差 Q 降低，但代价是拓扑误差增加 T。实验配置见第 3 节第二段。

误差、搜索准确率（定义见 2.1 节）。结果如图6所示，表明该算法在性能方面具有可扩展性：在所考察的 N 范围内，量化误差和拓扑误差均呈指数级下降。这一发现表明，超参数值对系统规模并不敏感，这对参数调优很有参考价值（也就是说，适用于小型映射的配置也可用于更大型的映射）。我们认为，这一便利特性得益于尺度不变的级联参数化，以及启发式搜索的可扩展性。

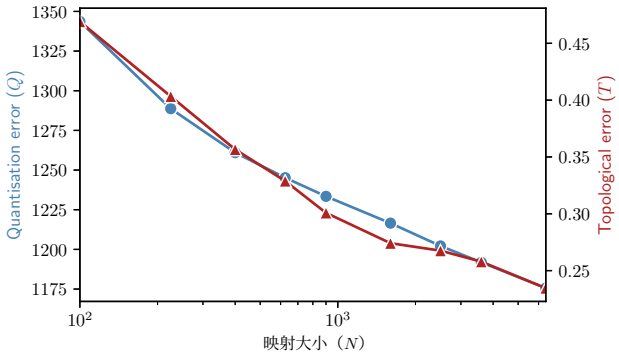

图*6.* 增大映射大小 N 会降低拓扑误差和量化误差（Q 和 T），这表明随着系统规模增大，映射质量也会提高——前提是较大的 N。实验配置见第 3 节第二段。

3.4 适用性

在本小节中，我们展示了所提出的算法在多种数据集上的表现可与文献中最接近的变体（自组织映射）相媲美。通过这一展示，我们还验证了所提出的超参数

数据集	类别	特征	样本
时尚-MNIST	10	784	59999 10000
字母	26	16	15000 5000
MNIST	10	784	59999 10000
卫星图像	6	36	4435 2000

表 *1.* 用于评估算法分类性能和数据集敏感性的数据集说明。数据集来源见（Melka& Mariage，2017）

能够适用于范围广泛的不同数据集。尽管如此，我们要指出，本研究的目标并非超越最先进水平，而是展示一种以异步方式训练特征映射的新方法。优化算法性能（通过进一步校准超参数或增加 N）可作为未来研究的课题。

为了更轻松地将算法的结果与文献进行比较，我们使用了一个包含 34×34 个单元的映射——由此得到$N = 1156$。我们还将 $c_d = 1000$，从而获得较低的量化误差（见小节3.2）。

分类 本节中，我们将所介绍的异步训练特征映射（AFM）用于分类任务。为简洁起见，并便于将我们的工作与（Melka& Mariage，2017）中提出的结果进行有意义的比较，我们刻意采用了一种简单的分类方案——与（Melka & Mariage，2017）中使用的方案几乎完全相同。

1. 每个映射单元 j 都会被分配一个类别 y_j，该类别代表与该单元最相似的样本类别。这些类别在训练完成后分配给各单元。

2. 令 Y_i 表示样本 i的标签。我们首先找到与 w_j最相似的样本：

$$i^*(j) = \underset{i}{\operatorname{argmin}} |w_j - s_i| \tag{7}$$

然后，将样本 i^*的类别 $y_j \leftarrow Y_{i^*}$赋予单元 j。

3. 最后，要预测某个样本的标签，我们找到该样本的最佳匹配单元，然后返回该最佳匹配单元的类别。

表2 通过呈现 5 次运行的均值和标准差，总结了该算法的分类性能。此外，该表还列出了包含 1200 个单元的自组织映射的性能指标——这些指标曾在（Melka & Mariage，2017）中给出。由于性能差异性能差异相对较小，因此我们可以认为，所介绍的算法与规模相当的自组织映射性能相当。

图 7表明，随着系统规模增大，分类精确率也会提高。

最后，表 3 使我们能够评估启发式搜索算法和级联机制对数据的结构。具体而言，表格3 列出了每个样本之后发生的权重更新操作数量，并在整个样本周期内取平均值。报告的数量之间的最大相对两两距离约为 1。5%，这表明所有模拟中的级联强度相近。表格 3还列出了每个映射在训练接近结束时的搜索准确率，并按数据集取平均值（搜索准确率的定义见小节 2.1）。不同数据集的搜索误差有所不同，但总体上仍然较低，最高为2。74%（MNIST）。不同数据集之间搜索准确率的差异，可能归因于数据集维度的不同（因为维度较高的样本空间可能更难搜索）。

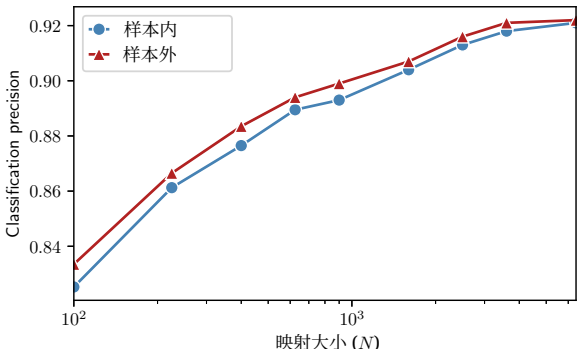

图7。用于分类时，所提出映射的精确率会随映射大小增加 N（样本内和样本外均如此）。所使用的分类方案见小节 3.4。实验配置参见第 3 节第二段。

3.5. 计算复杂度

先前的实验表明，在满足以下条件的超参数配置下，所提出的算法能够实现与自组织映射相当的性能：$e \sim N$，$p_i \leq 1$，$i_{max} \sim N$）。这一发现使我们能够推导出该算法复杂度的上界，如下所示。

对于每个 $i_{max} \sim \mathcal{O}(N)$ 训练样本，我们进行一次包含 $e \sim \mathcal{O}(N)$ 次迭代的利用式搜索，随后进行 g_i 次迭代的贪心搜索。由于贪心搜索期间不会重复访问任何单元，因此它是 $g_i \leq N$，所以 $g_i \sim \mathcal{O}(N)$。最后，搜索结束后，我们继续进行级联。对于 $p_i = 1$，级联规模达到最大值

		AFM		稀疏自组织映射	
数据集		精确率	召回率	精确率	召回率
时尚-MNIST	test	78.0 ± 0.6	76.8 ± 0.6		
	训练	78.0 ± 0.3	76.8 ± 0.4		
字母	test	$\mathbf{83.3 \pm 0.2}$	$\mathbf{82.8 \pm 0.3}$	81.5 ± 0.5	81.1 ± 0.5
	训练	$\mathbf{86.0 \pm 0.2}$	$\mathbf{85.7 \pm 0.3}$	83.7 ± 0.3	83.7 ± 0.3
MNIST	test	93.3 ± 0.1	93.2 ± 0.2	$\mathbf{93.4 \pm 0.2}$	$\mathbf{93.4 \pm 0.2}$
	训练	93.1 ± 0.2	93.1 ± 0.2	$\mathbf{93.5 \pm 0.2}$	$\mathbf{93.5 \pm 0.2}$
SatImage	test	$\mathbf{87.8 \pm 0.4}$	$\mathbf{87.6 \pm 0.6}$	87.6 ± 0.3	85.3 ± 0.4
	训练	91.7 ± 0.2	91.5 ± 0.2	$\mathbf{92.3 \pm 0.4}$	$\mathbf{92.4 \pm 0.3}$

表2。 将所提出的特征映射（AFM）的分类性能与自组织映射（SOM）的分类性能进行比较——采用文献（Melka & Mariage，2017）中的数值。计算了 AFM 进行五次运行时，在多个数值下召回率和精确率的平均值及标准差。结果与 SOM 的性能相当。我们使用 34×34 一个 AFM，采用 $c_d = 1000$ ——其余超参数值见第 3 节第二段。

（Vespignani 等人，1998）——所提出的模型可以精确映射到统计力学中的沙堆模型（Bak 等人，1988）。众所周知，沙堆模型的级联规模按 N（Bak 等人，1988）的规律缩放。由于在所提出的算法中 $p_i \leq 1$，沙堆模型的线性缩放是我们模型的一个上界，$\bar{a}_i \sim N$（上划线 $\bar{}$ 表示上界）。

数据集	最大分数级联	权重数量更新次数 / 样本	搜索误差
时尚-MNIST	0.75 ± 0.11	3.16 ± 0.01	$2.46\% \pm 0.26\%$
字母	0.76 ± 0.15	3.18 ± 0.01	$1.62\% \pm 0.28\%$
MNIST	0.87 ± 0.11	3.18 ± 0.01	$2.74\% \pm 0.74\%$
SatImage	0.78 ± 0.17	3.21 ± 0.02	$1.62\% \pm 0.56\%$

表3. 不同数据集的最大分数级联规模 $\max A_i$、平均级联规模 $\text{avg}(a_i)$ 和平均搜索误差 F。不同数据集之间的数值具有可比性，这表明算法行为不受数据集具体特征的影响。

根据上一段内容，我们可以推导出所提出的超参数参数化（$e \sim N$，$p_i \leq 1$，$i_{max} \sim N$）的计算复杂度：

$$\mathcal{O}(i_{max}\,(e + \bar{g}_i + \bar{a}_i)) = \mathcal{O}\big(N(N+N+N)\big) = \mathcal{O}(N^2) \tag{8}$$

因此，所提出的参数化的计算复杂度为 N^2，与文献中最接近的变体（自组织映射（Kohonen 等人，2000））属于同一类别。

4. 讨论

目前，所有最先进的拓扑映射（及其变体）的训练算法都依赖于密集连接的神经元。尽管稀疏拓扑结构能够带来可扩展性优势，但情况依然如此（Mocanu 等人，2018）。为了提升拓扑映射的计算性能，最先进水平依赖于分布式 MapReduce 平台（Sarazin 等人，2014；Liu 等人，2018）。然而，这类解决方案需要同步和中心实体，因此容易受到通信信道限制（即延迟、带宽和可靠性）以及慢速工作节点的影响（Mnih 等人，2016；Li 等人，2019；Fernando 等人，2017）。

通过启用训练过程的异步执行，可以克服这些限制。总体而言，启用算法异步执行所面临的挑战，在于确保算法各个部分之间的依赖程度足够低，也就是实现松散耦合。为此，我们提出一种拓扑映射训练算法，将所有计算分配给由自主单元组成的群体。每个单元只需了解少数预定义邻居，并与之进行低频交互，从而实现松散耦合。

我们通过实证研究该方案在 MNIST 上的表现，并在此过程中给出一套具体的超参数配置。随后，我们展示了所提出方法在分类任务中的潜力；结果发现，该算法的表现可与自组织映射相媲美。此外，我们证明，所给出的超参数配置并不强烈依赖于映射大小或数据结构——参见第3.3 和3.4小节。最后，我们在所给定的超参数配置下，得出了计算复杂度的解析上界3.5 ，并发现该上界与该领域中同步训练方法（自组织映射）的计算复杂度相同。

未来的工作可以着重于改进所提出算法的各个组成部分，例如采用更有效的分布式搜索方法。另一种选择是考虑其他级联机制，使其能够接受更快的外部驱动，从而提高数据吞吐量（即可以同时在不同单元上处理更多样本）。其他工作也可以着眼于工业应用，这些应用可能会受益于可处理的映射大小上限的提升。

参考文献

Bak, P., Tang, C., and Wiesenfeld, K. 自组织临界性。*Physical Review A*，38(1):364，1988。

Barri`ere, L., Fraigniaud, P., Kranakis, E., 和 Krizanc, D. 在具有长程连接的网络中进行高效路由。载于分布式计算国际研讨会，第 270–284 页。Springer，2001 年。

Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., Pritzel, A., 和 Wierstra, D. Pathnet：在超神经网络中进化通道梯度下降。*arXiv* 预印本 *arXiv:1701.08734*，2017 年。

Girau, B. 和 Torres-Huitzil, C. 自组织映射的容错性。神经计算与应用，2018 年 10 月。ISSN 1433-3058。doi：10.1007/s00521-018-3769-6。URLhttps://doi.org/10.1007/s00521-018-3769-6。

Kleinberg, J. M. 小世界中的导航。*Nature*, 406 (6798):845–845, 2000.

Kohonen, T. 自组织映射要义。神经网络, 37:52–65, 2013.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. 大规模文档集合的自组织。*IEEE TransactionsonNeuralNetworks*, 11(3):574–585, 2000年5月。ISSN 1045-9227。doi: 10.1109/72.846729.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. 大规模文档集合的自组织。*IEEE TransactionsonNeuralNetworks*, 11(3):574–585, 2000年5月。ISSN 1045-9227。doi: 10.1109/72.846729.

Kohonen, T., Schroeder, M. R. 和 Huang, T. S.（编）。自组织映射。Springer-Verlag，柏林，海德堡，第3版，2001年。ISBN 3540679219。

Li, A., Spyra, O., Perel, S., Dalibard, V., J aderberg, M., Gu, C., Budden, D., Harley, T. 和 Gupta, P. 人口训练的通用框架。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第1791–1799页，2019年。

Li, X., Gasteiger, J. 和 Zupan, J. 关于自组织特征映射中的拓扑失真。生物控制论，70(2)：189–198，1993年。

Liu, Y., Sun, J., Yao, Q., Wang, S., Zheng, K. 和 Liu, Y. 基于 MPI/CUDA 的可扩展异构并行自组织映射。载于 Zhu, J. 和 Takeuchi, I.（编），第十届亚洲机器学习会议论文集，第 机器学习研究论文集第95卷，第264–279页。PMLR，2018年11月14日至16日。网址：http://proceedings.mlr.press/v95/liu18b.html。

Malcai, O.、Shilo, Y. 和 Biham, O. 具有普适指数的耗散沙堆模型。《物理评论*E*》，73（5）：056125，2006年。

Martel, C. 和 Nguyen, V. 对Kleinberg（及其他）小世界模型的分析。载于第二十三届*ACM*分布式计算原理研讨会论文集，第179–188页，2004年。

Melka, J. 和 Mariage, J.-J. 面向稀疏输入数据的自组织映射高效实现。载于 *IJCCI*，第 54–63 页，2017 年。

Milgram, S. 小世界问题。《今日心理学》，2 (1)：60–67，1967 年。

Mnih, V.、Badia, A. P.、Mirza, M.、Graves, A.、Lillicrap, T.、Harley, T.、Silver, D. 和 Kavukcuoglu, K. 深度强化学习的异步方法。载于国际机器学习会议，第 1928–1937 页，2016 年。

Mocanu, D. C.、Mocanu, E.、Stone, P.、Nguyen, P. H.、Gibescu, M. 和 Liotta, A. 受网络科学启发，采用自适应稀疏连接实现人工神经网络的可扩展训练。《style id='1'>《自然·通讯》，9 (1)：1–12，2018 年。

Qiu, J., Wu, Q., Ding, G., Xu, Y., 和 Feng, S. 大数据处理机器学习综述。《*EURASIP*信号处理进展期刊》，2016(1):67，2016。

Sarazin, T., Azzag, H., 和 Lebbah, M. 使用 Spark-MapReduce 进行自组织映射聚类。载于*2014*年 *IEEE* 国际并行与分布式处理研讨会专题讨论会，第 1727–1734 页，2014 年 5 月。doi: 10.1109/IPDPSW.2014.192。

Vespignani, A., Dickman, R., Mu˜noz, M. A., 和Zapperi, S. 沙堆中的驱动、守恒与吸收态。《物理评论快报》，81(25):5676，1998。

Wang, H.、Zhang, N. 和 Cr´eput, J.-C. 一种用于大规模欧几里得旅行商问题的大规模并行神经网络方法。*Neurocomputing*，240:137–151，2017。

A. 附录

我们通过实验评估在所提出的超参数化方案（见第3节“默认配置”段落）下启发式搜索（见第2.1小节）的可扩展性。具体而言，我们使用 MNIST 数据集，针对不断增大的映射（$N \in \{100, 225, 400, 625, 900, 1600, 2500, 3600\}$）计算搜索误差（定义见第2.1小节），以及拓扑误差和量化误差（定义见第3节）。对于每个 N值，我们模拟10个映射。

实验结果如图8所示，其中展示了平均搜索误差、拓扑误差和量化误差，以及各自的误差带。最左侧的子图表明，随着 N增加，搜索误差保持在同一水平，这意味着所提出的搜索方法相对于系统规模具有可扩展性（针对给定的超参数设置和数据集）。此外，中间和最右侧的子图显示，量化误差和拓扑误差均随 N增加而降低。这些观察结果使我们得出结论：所提出的搜索方案相对于规模具有可扩展性——也就是说，随着 N 增加，映射的拓扑质量和近似质量都会提高。

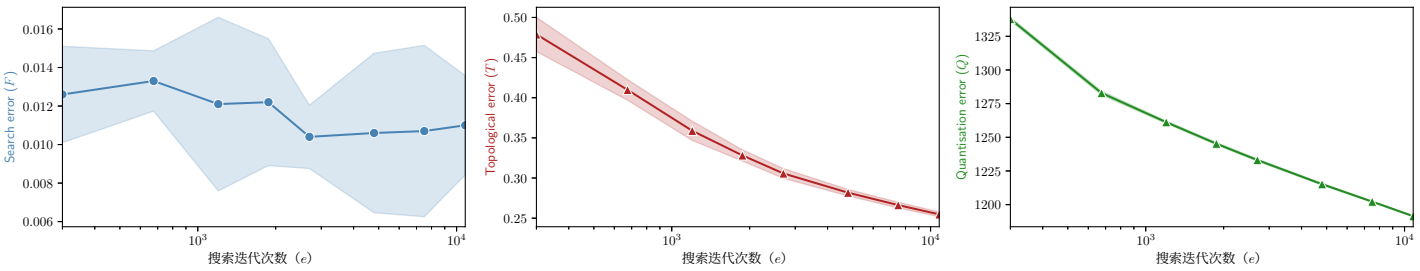

图 *8.* 所提出的启发式搜索方案的效果对映射大小 N 不敏感，因此随着 N 增加，映射质量也会提高。左侧子图显示搜索误差保持在相近水平，而中间和右侧子图显示拓扑误差和量化误差均随着 N 增加而降低。误差带表示一个标准差；在最右侧（绿色）子图中，误差带无法从视觉上辨认。每个 N 值均训练了 10 个映射，搜索迭代次数设为 $e = 3N$。实验设置的其余细节见第 3 节“默认配置”段落。