在监督降维中控制失真以保留类和邻居

伯努瓦科兰热
大学 Grenoble Alpes, INES,
F-73375, Le Bourget du Lac, France
benoit.colange@cea.fr

雅科·佩尔托宁
坦佩雷大学信息技术学院
和通信科学，芬兰
jaakko.peltonen@tuni.fi

迈克尔·奥珀蒂
卡塔尔计算研究所，哈马德本哈利法大学，多哈，卡塔尔
maupetit@hbku.edu.qa

丹尼斯杜特赫
大学 格勒诺布尔阿尔卑斯大学 萨瓦勃朗峰，
CNRS, LAMA, 73000 尚贝里, 法国
denys.dutykh@univ-smb.fr

西尔万·莱斯皮纳茨
大学 Grenoble Alpes, INES,
F-73375, Le Bourget du Lac 法国
sylvain.lespinats@cea.fr

抽象的

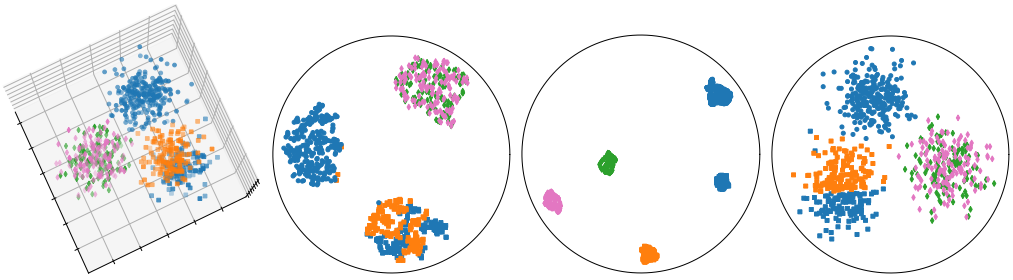
高维数据的非线性降维具有挑战性，因为低维嵌入必然包含失真，并且很难确定哪些失真是最需要避免的。当可以将数据注释到已知的相关类中时，它可用于指导嵌入以避免恶化类分离的扭曲。本文介绍的监督映射方法，称为类NeRV，提出了一个原始的压力函数，该函数将类注释考虑在内，并根据错误的邻居和遗漏的邻居来评估嵌入质量。类NeRV共享从随机邻域嵌入 (Stochastic Neighbor Embedding)新能源）。我们的方法比以前的方法有一个关键优势：在文献中，监督方法通常以扭曲数据邻居的结构为代价强调类分离；相反，无监督方法以经常混合类为代价提供更好的结构保存。实验表明类NeRV可以同时保留邻居结构和类分离，优于九个最先进的替代方案。

1 简介

降维 (DR) 方法旨在将高维数据集映射为低维嵌入空间中的点，同时保留数据点之间的一些相似性度量。可以通过利用类信息来监督 DR。因此，监督方法从数据的相对位置（非监督方法也使用）和类标签计算映射。DR 技术 [1、2、3] 可用作分类或聚类的预处理步骤，或将（标记的）数据可视化为散点图。在映射标记数据时，有两个相互矛盾的目标：

- *分类*是典型的*受监督的*DR 技术：在嵌入空间中用分类精度强调和测量类分离。

第 34 届神经信息处理系统会议 (NeurIPS 2020)，加拿大温哥华。



(a) 3D数据集　　（乙）tSNE (无人监督）　（c）S-UMAP (监督）　　(四)类NeRV

图1：类NeRV旨在保留类和邻居：数据从三个 3D 高斯簇 (a)、纯蓝色类的圆簇、被分隔蓝色和橙色类的平面减半的正方形簇和菱形簇中采样随机分布的绿色和粉色类。显示了这些数据的不同平面嵌入：无监督神经元

(b) 很好地保留了聚类，但由于忽略了标签，因此与正方形聚类中的橙色和蓝色类重叠。被监督的S-UMAP (c) 分裂正方形和菱形的簇，迫使类分离。因此，它误导了橙色和蓝色类的原始空间邻接，以及绿色和粉色类的混合。类神经元 (d) 旨在更好地保留三簇结构以及类的邻接性。

- *探索性数据分析*是典型的*无监督*在不知道类信息的情况下运行的 DR 技术：数据邻域结构被优先考虑并测量为原始空间和嵌入空间中数据相似性之间的差异。

这些目标源自视觉分析任务 [3、4]。它们是矛盾的，除非类和数据邻域结构在数据和嵌入空间中彼此匹配良好：每个类构成不同的区域，没有跨类邻域关系。不幸的是，这种理想情况不太可能发生，因为数据邻域结构和类在数据空间中并不总是匹配，并且高维数据的低维嵌入不可避免地会出现失真 [3]：*假邻居*它们是嵌入中的相邻点但不在数据中，并且*想念邻居*它们在数据中是邻居但在嵌入中不是。

在这项工作中，我们建议类NeRV，一种受监督的 DR 技术，用于在考虑类信息的同时完成探索性分析目标。我们的解决方案在原则上与之前的类似分类地图 [5] 基于距离的投影，但我们从与神经病毒 [6, 7],神经元 [8], 和tSNE [9]. 我们的解决方案明显不同于其他以前的监督方法，这些方法倾向于以牺牲邻居的结构为代价强制类分离，*例如*S-等值图[10],S-神经病毒 [11] 和S-UMAP [12]。图 1 说明了基本特征类 NeRV。

我们的贡献是双重的：我们建议**类NeRV**当将高维标记数据嵌入低维空间时，它利用类信息来确保更好地保存类。它的压力函数，来自无监督神经病毒 [6、7] 引导优化，以便将不可避免的邻域结构扭曲放置在对类结构危害较小的位置。通过强调惩罚类之间的虚假邻居和类内遗漏的邻居来避免有害的扭曲。我们还推导出两个新的类感知质量指标从标准*诚信度*和*连续性*质量指标 [13]，专门说明影响类保存的扭曲。

2 相关工作

无监督嵌入。　　之前已经提出了许多线性或非线性算法
包括主成分　　新台币（主成分分析）[14], 自组织图 (索姆) [15]、等距特征映射（等轴测图）[16]、数据驱动的高维缩放（DD-HDS) [17], 局部仿射多维项目 (灯） [18] 和均匀流形近似和投影 (UMAP) [12]。在各种各样的技术中，Neighborhood Embedding (东北） 技术在保留邻域结构和计算时间方面是有效的。他们的概率框架还提供了一个理论背景来解释所获得的地图

邻域检索任务[7]。NE 方法为每对点计算*我, j*, 点的概率隶属度*j*到点附近*我*，有时称为相似性。这些隶属度是在数据空间和嵌入空间中计算的。通过最小化这两个空间之间的隶属概率差异来获得映射。这些方法包括随机邻域嵌入（SNE) [8], t-分布SNE (t-SNE) [9], 詹森香农嵌入 (JSE) [19] 和邻域检索可视化工具 (神经病毒）[6, 7]。神经网络和t-SNE 不同之处在于用于计算它们在嵌入空间中的邻域隶属度的内核。 JSE和神经病毒都延伸神经网络来控制虚假邻居和遗漏邻居之间的平衡。这种可调性神经病毒和JSE使它们最适合引入监督。

监督嵌入。监督映射方法主要关注*阶级分离*，在映射之前或期间修改数据邻居结构。他们中的许多人通常会降低类之间的数据相似性并增加类内的数据相似性，然后在修改后的相似性上使用标准的无监督映射方法。它们可以使用度量学习对差异（距离）进行操作，如监督局部线性嵌入SLLE及其变体 [20, 21, 22]，监督等值图（S-等值图）[10, 23], 监督 NeRV (S-神经病毒）[11] 或监督 UMAP S-UMAP [12]，或与重尾半监督版本的相似之处tSNE [24]。其他方法依赖于全局参数映射并优化其参数以*最大化类分离*, 作为线性判别分析 (低密度脂蛋白）[25]及其核化变体 [26、27、28]，邻域成分分析（NCA) [29]及其神经网络变体[30]或有限秩矩阵学习向量量化（ LiRaM LVQ) [31]。类感知tSNE (catSNE) [[32] 根据类的分布在本地调整要保留的邻域大小。终于，分类地图 [5] 优化类似于局部多维缩放的应力函数（LMDS)

[33]，但支持*标记数据的探索性分析*通过在惩罚错误和遗漏的邻居时考虑班级。然而，两者LMDS和分类地图基于距离的对高维度的范数集中现象很敏感[34]，而网元像这样的技巧神经病毒使用减轻这种影响*移不变*会员度[35]。类NeRV得出与相同的原理分类地图来自神经病毒压力函数来获得两者的好处。

3 ClassNeRV 和 Class-aware 质量指标

3.1 NeRV 压力函数

在神经病毒 [6, 7], 一个点的隶属度（条件概率）*j*到附近另一点*我*, 表示β_{ij}在数据空间和b_{ij}在嵌入空间中，定义为：

$$\beta_{ij}, \Sigma \frac{\exp\left(-\Delta_{2个ij}/2个\sigma_{我}^{2个}\right)}{k\bar{o}个指数（-\Delta_{2个我知道}/2个\sigma_{我}}\ 和\ b_{ij}\ ,\ \Sigma \frac{\exp\left(-丁_{2个}/2个_{秒_{2个我}}\right)}{k\bar{o}个指数（-D_{2个我知道}/2个_{秒我})}^{2个} \quad (1)$$

和Δ_{ij}和$丁_{ij}$点之间的距离*我*和*j*分别在数据和嵌入空间中。

隶属度的分布表示为$\beta_{我}, \{\beta_{ij}\}_{j6个=我}$和$b_{我}, \{b_{ij}\}_{j6个=我}$. 这隶属度是平移不变的，减少了范数集中度 [35]。秤参数$\sigma_{我}$设置为获得固定的用户选择的困惑*p*类似于光滑或模糊的我Σ确定数量每个点的邻居[36]：$H(\beta_{我})$ = 日志*p*, 熵$H(\beta_{我})$,− $\sum_{j6=我}\beta_{ij}$日志β_{ij}. 这里，我们设置嵌入尺度参数$秒_{我}$等于$\sigma_{我}$.

这神经病毒压力函数是两组 Kullback–Leibler (KL) 散度之间的线性权衡：

$$\zeta_{神经病毒}, \sum_{我} \tau丁_{吉隆坡}(\beta_{我}, 乙_{我})+(1-\tau)D_{吉隆坡}(b_{我}, \beta_{我}), \tau \sum_{i,j6=我} \beta_{ij}日志\left(\frac{\beta_{ij}}{b_{ij}}\right)+(1-\tau) \sum_{i,j6=我} b_{ij}日志\left(\frac{b_{ij}}{\beta_{ij}}\right) \quad (2)$$

在等式（2）中，$\tau \in [0; 1]$控制th Σ e 之间的权衡 $\sum_{我}丁_{吉隆坡}(\beta_{我}, 乙_{我})$惩罚*想念邻居-布尔斯*(最大的时候$\tau$=1)和 $\sum_{我}丁_{吉隆坡}(b_{我}, \beta_{我})$惩罚*假邻居*(最大的时候$\tau$=0).什么时候$\tau$=1、神经病毒减少到神经元。

3.2 ClassNeRV 压力函数

我们的目标是保留邻域结构和类别的嵌入。我们建议嵌入不应该*在同一类中人为地分离点*， 或者*人工聚类*

来自不同类别的点在一起. 因此，我们需要惩罚更多*类内错过的邻居*和*类间假邻居*分别。为了控制这些基于类别的扭曲，我们进一步将等式（2）中的发散项分解为*班内*和*课间*关系。它给

两个额外的不同的基于类别的权衡参数$\tau\in$和$\tau\in$ /(同时在 [0,1])并导致类NeRV应力函数：

$$\zeta_{\text{类NeRV}},\quad \sum \tau\text{于乙}\,(\beta_{\text{我}},b_{\text{我}}) + (1\text{个}-\tau\in)\mathrm{D}_{\text{乙}}(b_{\in\text{我}},\beta_{\text{我}}) + \tau\ \text{于乙}\,(\beta_{\text{我}},b_{\text{我}}) + (1-\tau\in)\ \mathrm{D}_{\text{乙}}(b_{\text{我}},\beta_{\text{我}})$$

$$,\quad \tau\in^{\text{我}}\sum\Big(\beta_{ij}\ \text{日志}\ \frac{\beta_{ij}}{b_{ij}} + b_{\text{我}}\beta_{ij} + (1-\tau\in)\sum_{i,j\in S_{\in\text{我}}} b_{ij}\text{日志}\ \frac{b_{ij}}{\beta_{ij}} + \beta_{ij-ij}\,b\Big)$$

$$+\ \tau\notin\sum_{i,j\in S_{\notin\text{我}}}\Big(\beta_{ij}\text{日志}\ \frac{\beta_{\text{我}}}{b_{ij}} + b_{ij}\text{-}\beta_{ij} + (1-\ \tau\in)\sum_{i,j\in S_{\notin\text{我}}} b_{ij}\text{日志}\ \frac{b_{ij}}{\beta_{ij}} + \beta_{ij}\text{-}b_{ij}\Big) \qquad (3)$$

与类内的集合*小号*$\in$ $_{\text{我}}$=*{j6个=我 | 大号*$_{\text{我}}$*=大号*$_j$*}*和课间*小号*$\in$ $_{\text{我}}$=*{j6个=我 | 大号*$_{\text{我}}$*6个=大号j}*基于类标签定义的索引*大号*$_{\text{我}}$点数*我*.

请注意，Bregman (B) 散度替换方程 (2) 的 KL 散度以确保应力函数的正性。实际上，KL 散度仅针对概率分布定义（求和到1),而在四个方面$\zeta_{\text{类NeRV}}$, 隶属度限制在集合中*小号*$\in$ $_{\text{我或者小号}\in\text{我}}$和小于1.因此，KL 散度不能直接适用于此类隶属度（否则它不会满足散度的已知属性，即非负性和不可区分的身份），而 Bregman 散度则直接适用。Bregman 散度项

丁乙($\beta_{\text{我}\in}$, 乙$_{\text{我}\in}$)和丁乙($b_{\in\text{我}}$, $\beta_{\text{我}}$)惩罚*班内*分别错过了和错误的邻居，而条款丁$\beta_{\text{我}}$($_{\text{我}}$,$b_{\in\text{我}}$)和丁$b_{\in\text{我}}$($_{\text{我}}$, $\beta_{\text{我}}$)惩罚*课间*分别是 missed 和 false neighbors。

参数$\tau\in$和$\tau\in$ /定义类NeRV通过加权这些术语来映射行为。$\tau\in$控制惩罚假邻居和遗漏邻居的平衡*班级内*，尽管$\tau\in$ /控制类似的平衡*课间*. 因此，类NeRV受到监督如果$\tau\in$> $\tau\in$ /,*IE*它的压力函数比其他函数惩罚更多的类内遗漏邻居和类间假邻居扭曲。越大$\tau\in$- $\tau\in$ /差值越大，监管水平越高为了$\tau\in$=1个和$\tau\in$/=0。有$\tau\in$< $\tau\in$ /但是会偏爱班级内错过的邻居，并且类之间的假邻居，鼓励同类分裂和不同类重叠，搞砸了类保存。类NeRV不受监督如果$\tau\in$=$\tau\in$ /然后减少到原来的神经病毒，Bregman 类内和类间的组合散度等于相应的 KL 散度，其中b_{ij}-β_{ij}条款取消。

我们重新参数化$\tau\in$和$\tau\in$ /作为$\tau*$= ($\tau\in$+$\tau\in$)/2个和ε= ($\tau\in$- $\tau\in$)/2.$\tau*\in$[0,1]控制平均值*虚假邻居和遗漏邻居的惩罚之间的权衡*(作为τ在神经病毒），尽管 $\varepsilon\in$[0,0.5]控制*监管水平*(更多的监督以获得更大的价值）。相反转换是：$\tau\in$=$\tau*$+ε和$\tau\in$ /=$\tau*$- ε.

鉴于$\tau*$和ε，这类NeRV嵌入是通过最小化应力获得的$\zeta_{\text{类NeRV}}$在等式（3）中关于嵌入点的坐标。优化是使用多尺度优化方法 [37] 和拟牛顿法进行的BFGS算法[38]。当前的实现复杂度为*欧*(*否*$_{2个}$) (在哪里*否*是数据点的数量），但是基于树的加速技术 [39、40] 可以将其减少到*欧*(*否*日志*否*).

3.3 探索性分析监督技术的质量指标

为了评估邻域结构的保存情况，我们采用了可信度和连续性度量 [13]，这是在探索性分析 [3] 的背景下评估无监督嵌入的标准。这两项措施分别量化虚假邻居的平均水平（可信度吨)和错过的邻居（连续性C）对于给定的邻域大小κ作为：

$$\text{吨}\ (\kappa), 1-\ \frac{1\text{个}}{\text{吨}_{\text{最大限度}}(\kappa)}\ \sum_{i,j\in F_{\text{我}}(\kappa)}\ (\rho_{ij}\text{-}\kappa) \quad \text{和} \quad \mathrm{C}\ (\kappa), 1-\ \frac{1\text{个}}{\mathrm{C}_{\text{最大限度}}(\kappa)}\ \sum_{i,j\in M_{\text{我}}(\kappa)}\ (r_{ij}\text{-}\kappa)\ , \qquad (4)$$

在哪里ρ_{ij}和r_{ij}是每个点的行列*j*在每个点的附近*我*分别在数据和嵌入空间中，以及$F_{\text{我}}$(κ)和*米*$_{\text{我}}$(κ)是 false 和

错过了点的邻居*我*. 归一化系数吨最大限度(*κ*)和C最大限度(*κ*)被定义为吨和C范围0对于理论上最差的映射到1个一个理想的映射。基于组合分析，吨最大限度(*κ*)和C最大限度(*κ*)等于*κ否*(2*ñ* -3个*κ*−1)/2个如果*κ* <*否*/2个 或者*否*(*N*-*κ*)(*N* −*κ*−1)/2个如果*κ*>*否*/2 [41]。

为了评估类保存，我们还推导出了两个新的衡量标准：*仅限于-阶级关系*吨∈ /和连续性*仅限于类内关系*C∈. 这些*类别意识*指标被定义为：

$$\text{吨}(\kappa), 1 - \frac{1\text{个}}{\text{吨}_{\text{最大限度}}(\kappa)} \sum_{i,j \in F_{\text{我}}(\kappa) \cap S \in \frac{\checkmark}{\text{我}}} (\rho_{ij} - \kappa) \quad \text{和} \quad \text{C}\in(\kappa), 1 - \frac{1\text{个}}{\text{C}_{\text{最大限度}}(\kappa)} \sum_{i,j \in M_{\text{我}}(\kappa) \cap S \in \text{我}} (r_{ij} - \kappa \quad). \quad (5)$$

请注意，这些类别感知指标仅解释了他们所考虑的部分扭曲无监督的同行，因此吨∈ />吨和C∈>C。因此，他们可能会达到1个只要有剩余的扭曲不影响类间可信度和类内连续性。

监督降维的另一个标准质量指标是留一法的准确性*k*-最近的邻居（*k*-NN) 嵌入空间中的分类器[3]。当嵌入用于执行分类时，该指标是有意义的，但它仅关注类并且未能考虑邻域结构保留，如图 3 所示。补充材料提供了广泛的结果*k*-NN 增益 [32] 来自该指标。

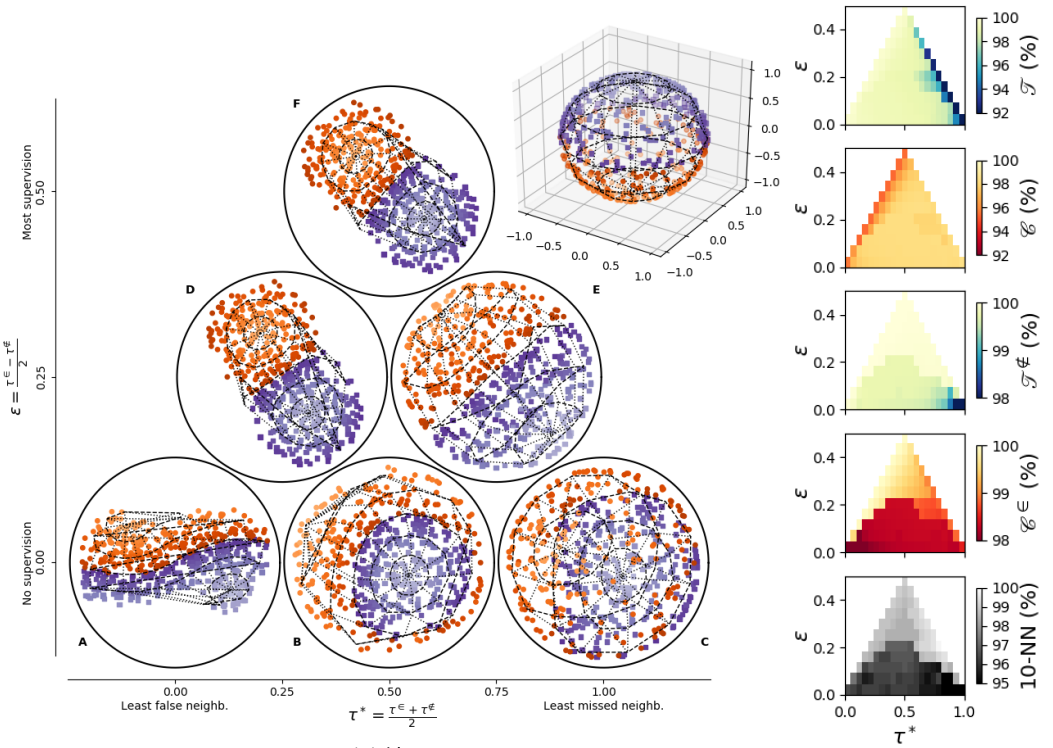


图 2：灵敏度类NeRV权衡参数，同时嵌入5123D数据*地球*中的数据集（第 4.2 节）*{τ∗, ε}*-参数空间。左图 (a) 显示了五个二维嵌入，其中平行线（虚线）和子午线（虚线）位于 *事后的*，以及数据集的 3D 表示（右上角）。右图 (b) 显示了使用第 3.3 节中定义的指标进行的量化评估，从上到下为 Trustworthiness

吨，连续性C，类间可信度吨∈ /和类内连续性C∈对于给定的邻域大小*κ*=32,和10−神经网络分类准确率（99%对于原始 3D 数据）。热图的每个像素对应于地球的一个嵌入。*τ∗*控制假邻居之间的整体权衡（很可能与*τ∗*=1)和错过的邻居（很可能是 *τ∗*=0), 尽管*ε*控制监督水平。

4 实验

4.1 目标、数据、技术

我们在 3D 玩具数据集 (*地球*) 和两个真实的高维数据集 (*胰岛5*和 *位数*). 这*地球*数据集（第 4.2 节）包含512数据随机分布在三维欧氏空间的单位球面上R3↑(图 2a)。这两个类（蓝色和红色点）对应于在赤道处划分的两个半球。这些数据不能在没有失真的情况下嵌入到平面中，因此最终地图取决于邻域之间的权衡集（$\tau*$) 和类 (ε) 惩罚（参见第 3.2 节）。这*胰岛5*数据集 [42, 43]（第 4.3 节）包含 1 559均匀分布的口语字母的英语发音录音26类，描述为617特征。这*位数*数据集 [44, 43]（第 4.4 节）包含3 823手写数字的图像。这8个 ⨯8个像素图像 (64D数据点）被分开10类（每个数字一个）。真实的类别标签以及随机生成的标签被认为是评估错误标记的鲁棒性。的随机子集500样本被认为是为了简化图 6 中地图的可读性。

我们比较类NeRV无人监督主成分分析 [14],等值图 [16],UMAP [12],tSNE [9]，和 神经病毒 [6, 7]，并进行监督美国国家航空航天局 [29],S-等值图[10],分类地图 [5] 和S-UMAP [12]。的实施 PCA、Isomap、NCA、tSNE、UMAP和S-UMAP来自scikit-学习（版本 0.22.1) [45] 和umap 学习 (版本 0.3.10) [46] Python 库。S-Isomap、ClassiMap、NeRV 和类NeRV使用我们自己的实现 [47]。所有现成的算法都设置了默认参数，除了神经元初始化，为此我们使用主成分分析而不是随机的（以便所有方法都受益于谱初始化）。神经病毒和类NeRV使用第 3.2 节中描述的优化。我们的多尺度优化中的最终困惑设置为ρ=32为了 *地球*和ρ=30为了*伊索莱特*, 为了与神经元默认ρ=30.我们对所有技术和数据集使用欧氏距离。定性结果按照标准指南 [3、48] 以圆形框架嵌入的散点图表示形式给出。定量结果是使用第 3.3 节中描述的无监督和类感知指标计算的。补充材料中显示了随机方法的额外运行，显示了类似的结果。

4.2 两个半球地球示例

图 2a 显示类NeRV的嵌入*地球*中的数据（$\tau*$, ε)-参数空间。地图
一个，乙和C对应原文神经病毒，*IE*没有任何监督（ε=0,$\tau\in=\tau\in$ ⁄). 在
地图A，假邻居受到的惩罚最大（$\tau\in=\tau\in$ ⁄=0),允许一些错过的邻居，所以
球体沿着子午线撕裂并展开。在地图上C，最想念的邻居
惩罚（$\tau\in=\tau\in$ ⁄=1),与原件对应神经网络映射[8]，让假邻居，
这样球体就会被压扁，混合红色和蓝色类。地图乙对应于神经病毒与
丢失和错误邻居的平衡混合（$\tau\in=\tau\in$ ⁄=0.5).添加一些监督（ε=0.25),
地图丁适度惩罚类内错过的邻居（$\tau\in$=0.5)和强烈的全类假邻居（$\tau*$=0.25),鼓励比 map 中更少撕裂的类A。反之，映射乙惩罚
适度的类间虚假邻居（$\tau\in$ ⁄=0.5)和强烈的全班失误（$\tau*$=0.75),
鼓励比 map 中更多的类分离C，但类内假邻居多于地图D.最后，地图F对应于最高监管水平（ε=0.5个和$\tau\in$=1个和 $\tau\in$
⁄=0),惩罚最多的类间错误邻居（最少的类重叠）和类内丢失的邻居（最大的类凝聚力）。详细了解每个项目的影响 类NeRV应力子项（等式（3））, 消融研究可在补充材料中获得。

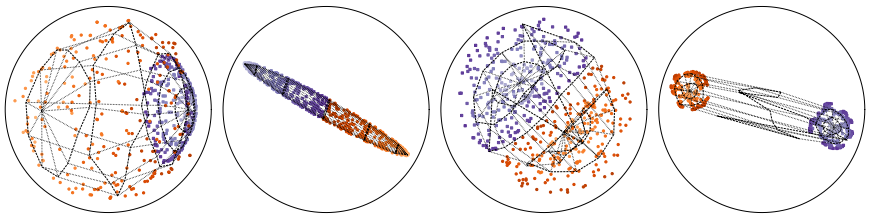
图 2b 在数量上支持这些定性观察。热门热图显示*结构保存*指标在同*{τ∗, ε}*-参数空间：when$\tau*$增加，虚假邻居的数量（吨蓝色）增加，而错过邻居的数量（C红色）减少。底部热图显示*班级保存*该空间的指标：当监管水平ε增加，

类重叠较少（吨$\in$ ⁄蓝色）和更少的阶级分裂（C$\in$红色的）。分类热图
也显示出更高的类别准确性（10–灰色的 NN）与更高级别的凝聚力和更低级别的重叠相一致ε.

图 3 和图 4 显示了监督最多的版本的定性和定量比较 类神经元 (ε=0.5)到其他技术。监督嵌入的定性分析表明类神经元 (图 2a，地图丁、乙和F), 分类图 (图 3c) 和美国国家航空航天局 (图 3b) 图

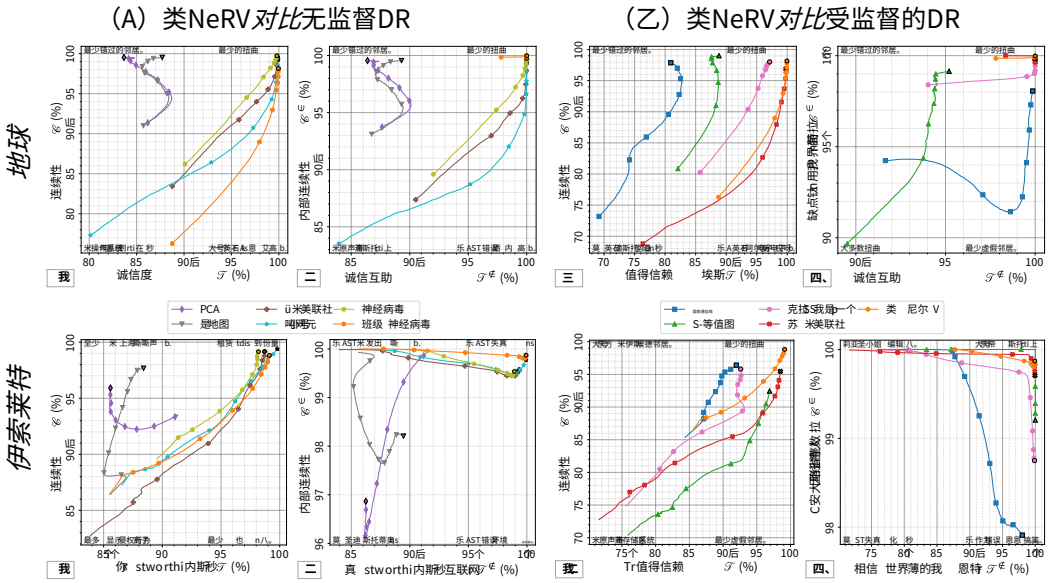
都保留了两个类的实际 3D 线性可分性和邻接性。然而，S-等值图
（图 3a）生成错误的类重叠，而S-UMAP (图 3d) 错误地将类别分开。

图 4（顶行）显示类神经元 (—) 达到与结构保存相似的水平tSNE (—) ， 或者地图 (—) 无监督技术 (I)，总体上具有更高的可信度（更少的假邻居）。正如针对无监督技术所预期的那样，它还可以更好地保留类别 (II)。关于监督技术，类神经元 (—) 获得更好的结构 (III) 和类 (IV) 保存比美国国家航空航天局 (—),类图 (—) 和S-等值图 (—). 为了S-UMAP (—), 类独立指标 (III) 类似于类 NeRV，除了较大的邻域面积κ， 由于S-UMAP'过度分离实际上相邻的类（图 3d）的趋势*地球*数据。



（A）S-等值图(90后.6%)　（乙）国家通信局(99.2%)　（C）分类地图(99.8%)(四）S-UMAP(100%)

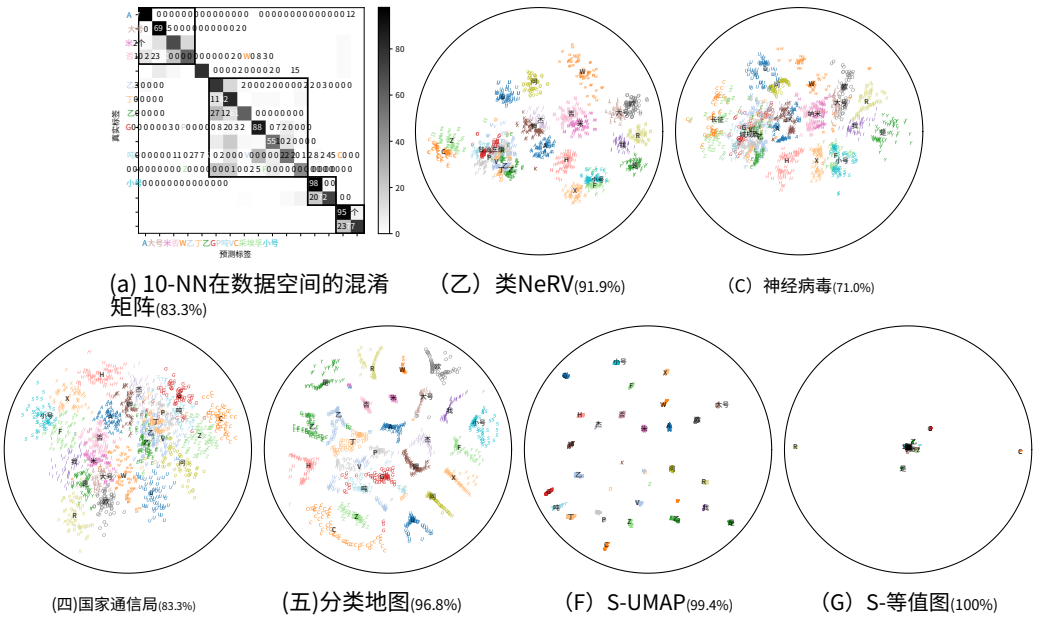
图 3：嵌入*地球*使用最先进的降维技术获得的数据集，与最受监督的数据集进行比较类神经元 (ε =0.5)在地图上F图 2a。 10-每种技术都给出了 NN 精度。第 4.2 节中的分析。



F搞笑的 你 阿 我 我知道(在 3.3)楼或者拉斯维加斯 RV有最大支持 电子ion (ε = 0.5) 一个 秒 电子 室 编辑 系统 *罗是* (吨 nd *是狮吨*(o 吨 m) 数据集。每个 p o里 你 C或者 二 秒 吨 特 我 的 我 是 C欢 吨 d 重 κ 米 2 *b否*/2个*C*(与米 Ar客rs 我 的 G $\kappa \in$ {2个,3,16个. } 广告κ = 2 个 h 编 我 b 交 我 电子无人监督（我，二）And秒 我 编 我、我）博士特 C尼 d 秒 你 鳄鱼 编 r 电 G 我 是 dto C 拉 神经元 (橙色) 在 h 吨 交 d 秒 你 p r 服务 on 在 d 我 (我)，一 n 滴 n 日 秒 留的 pac 在 我 你 无 我 会 ∈) (二 我)。P 电 编 在 Attop 里 G H 锥体 r. A nA 里 姐 第 4.2 节 (*地球仪*) 广告 n4 . 3(*秒*) 。

4.3 个是o 吨 在A 等

电子 10个否 Nc 在 F 单位 at 我 合 特 进 吨 H 37 Dd 在 A space（图 5a）如 秒 秒 阿 分 过 秒 w 我 是 s 吨 A 90% A 秒 的 球 福 电子 C HA 时 的 Ang类 (The 与 我 在 P sion 米 A 特 我 是 Giv 恩 我 电 的 里 吃 里 A). S 电 A 等级 是 帕 到 能由于重叠 到 我 为 和 秒



(a) 10-NN在数据空间的混淆矩阵(83.3%)　（乙）类NeRV(91.9%)　（C）神经病毒(71.0%)

(四)国家通信局(83.3%)　(五)分类地图(96.8%)　（F）S-UMAP(99.4%)　（G）S-等值图(100%)

图 5：a 的混淆矩阵 (a)10-NN分类器*胰岛5*显示数据空间中多个类重叠的数据集。矩阵中只显示了最混乱的类别。监督嵌入应该保留这个实际的类结构。最受监管类神经元 (ε=0.5)

(b) 与神经病毒 (c) 和监督方法 (d,e,f,g)。第 4.3 节中的分析。10-为原始数据 (a) 和所有嵌入 (b 到 g) 给出了 NN 留一法准确度分数。

用注音符号表示，*例如字*母F [Ef]和小号 [Es] (团体财政司司长），字母C [si:]和Z [子：] （团体捷克），字母A [el],大号 [埃尔],米 [Em]和N [恩] (团体警报器），或字母乙 [双：]，D [迪：]，E [我：]， G [天珠:],P [圆周率:],吨 [钛：]和V [六：]（团体BDEGPTV）。

图 5 显示了用于定性比较的嵌入。美国国家航空航天局 (图 5d) 倾向于人为地重叠类，例如QV和J.P。反过来，S-UMAP和S-等值图倾向于过度分离类别，完全忽略组中字母的实际重叠FS、CZ、ALMN、BDEGPTV 由混淆矩阵表示。类图 (图 5e) 更好地保留了BDEGPTV组但不能代表其他人喜欢锰或者捷克。神经病毒尽管不受监督，但仍设法保留组，表明这些组的类实际上在数据空间中重叠。然而，它可能会人为地增加

如图 1b 中观察到的重叠。最后，类神经元 (图 5b) 提供了一个更值得信赖的
类的表示，更好地展示 BDEGPTV通过将它们 强　格r 内部字母混淆明尼苏达州、捷克共和国、
放在 t 中相邻的组实际上在数据空间中。因　他米应用程序同时保持其他字母分开 Vmap 更值
此，克拉 专家发现，在这个特征空间中，字母SNeR得信赖，可以帮助按元音发音单独分组的域
FS或者明尼苏达州),具有不利的次要影响　是 海峡BDEGPTV,
奥南吨。

这10-NN 分类分数（图 5）inc Yet，asS-等值放松 w在这些地图上观察到第 i 个类分离。如果嵌
图达到最大值1个 实际类在数据空间结构保存00%经过足够可信，则人为地分离所有类别而不
中的混淆。因此，它不能用于支持*标记数据的*, th是c考虑分类指标无法评估所吃的类别
探索性分析　哦　阿鲁
A.

的定量比较类神经元 (— 行）显示它保留的邻) 无线 图 4a 上的监督技术（底部 ell 作为神经病
域结构比tSNE (—）　（我）。它保留了课程，重新作毒)(—）　和地图 (—), 但与预期的所有无监督
特别是对于小*κ*(二). 关于监督类神经元 (—) 保贝特室技术略有不同，图 4b（底行）展示了*岛数*
留了邻域监督方法 (III)。美国国家航空航天局 技术哦据比另一个好得多
(—) 生成和S-UMAP (—) 产生更多遗漏的所有 好　海峡
技术，除了美国国家航空航天局 (—) 在小型和更多的错误 邻居 (吨<C），尽管S-等值图 (—) 与嵌
大型规模上实现类似的优势*κ*为了班级　hbor(号, 小号)入一致（图 5）。
去吧外径类别保留 (IV) 的结果，具有
神经元、

4.4 手写数字

图 6 显示了到达的地图神经病毒和类NeRV适用于*数字*具有真实和随机标签的数据集。随机标签的情况允许在结构和类去相关时研究监督方法的行为。分离这些随机类会扭曲邻域，这可能被视为过度分离。我们可以在图 6a 和 6b 中观察到类NeRV依靠真实的类信息可以更好地保存类 神经病毒为了*数字*数据集。反过来，类NeRV使用随机标签（图 6d）不会导致随机类明显过度分离。图 7 证实了这些观察结果。此外，类NeRV当使用真实标签甚至更多随机标签时，显示比其他监督方法更好地保留结构和真实类（图 7）。

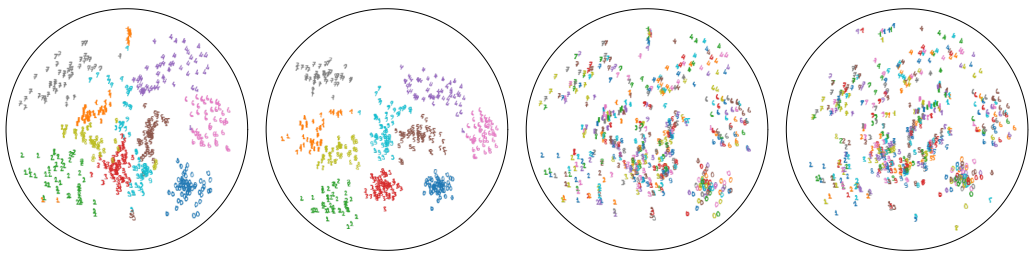
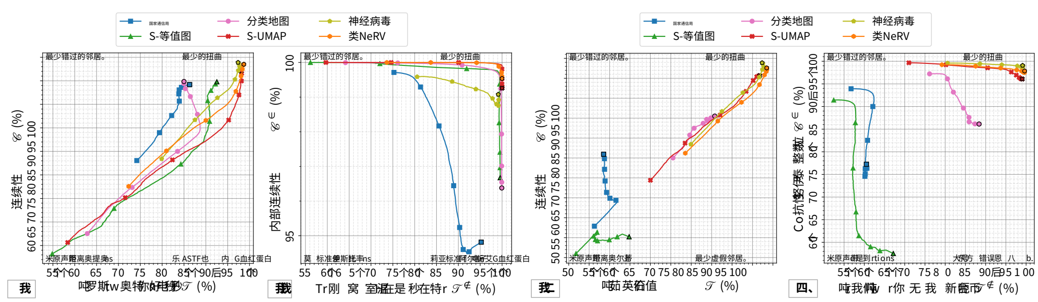


图 6：神经病毒和类NeRV的地图*数字*数据集，点由基于提供给算法的类别（真实或随机）着色的相关图像表示。鉴于神经病毒不受监督，图 (a) 和 (c) 具有完全相同的点位置。相反，可能会观察到差异类NeRV地图（b）和（d）。(b) 由于有效的类别信息，显示出比 (a) 更好的类别保留，而在 (d) 中可能没有明显的过度分类，具有完全不可靠的类别信息（随机标签）。4.4 节中的分析。



无花果。7：问你A代是我商C一吨个秒or一升秒5电RV，神经病毒和监督方法 秒在他吨上 d GE秒d在A秒等什么吨 UE A是升（我A和一r一ndom 升b电我们期. C升 秒 A w重指标 (吨∈ ,C∈) (,我我) q你A吨如果 p r电秒A对no英尺吨 c A SS没有 电日A吨神经病毒电没有 c 的帐户 拉秒精华电子 我S电子秒 我 我二我）r电子吨C A我到吨秒n (我我)分析升 我秒 C秒 4.4。

5 结论 秒我

类神经病毒 A二ow秒阿塔C E n钛秒偏局吨升 秒n 的结构保存点数 我升w 我电子我 阿尔嵌入 gs我 or 电子我依赖 d 吨A不是这样升 labe ed dat S 你分析 can f ns ac我吨 e 电子 吨 ct 什么 电子 G升我 r 电子 在给定的特征空间中分离，这可能会导致对标签的质疑 (安倍晋格吉子秒 r 吨电子你 秒(升A秒 电子我 r我G)．欧实验证明 电子 吨秒 电子我d R tend我秒特nd 吨ov电子par 吨G升我电子 电A吨A二是A d相邻或重叠 p我 g 秒吨 da a sp A C电子 我什么电子 秒 r六秒 e款至我 秒 o答案是我 o电子 A秒电子我 C对比，完全 s per s你 v我电子我 G oF 类神经病毒 b等吨电子一个A C A SS结构 吨A护理 DR技术 q你，升 H升 p电子我 G 否我H bo Hood 秒uc 吨e 和我们一升 秒我 oF 吨电子 N S 类似的 DR 技术。今后的工作将 前吨电子吨我A ppro交流吨 吨电子 G H博r H嵌入 dd我G技术，例如神经元 [9 吨东南 [] r杰 19]，和cno 秒吨电子我监督 r 我秒 F阿梅沃rk.

6 更广泛的影响

这项工作提出了改进用于探索性数据分析的降维技术。降维旨在支持数据科学家分析多维数据，也可用于在 2D 地图中可视化表示物理对象或人的高维数据，供外行公众了解主要对象/人组的概况关于它们相应数据的相似性。它与数据所代表的对象/人的性质无关。更好地了解大规模数据集中的趋势和变化可以提高社会了解重要现象的能力。然而，降维可能会产生这些对象/人的有偏见的表示，这可能是由于数据本身的固有偏见（对象/人的表示类别过多或不足，缺失或不相关的特征人工收集或分离（类）物体/人），或由于不可避免的投影偏差，称为扭曲，[3] 在 2D 表示中人工聚集实际上分离（类）物体/人，或人为分离二维表示实际上是相似的（类）物体/人。建议的类NeRV方法正是为了减少这第二种类型的偏见。

7 资金披露

Denys Dutykh 的工作得到了法国国家研究机构的支持，通过未来投资计划（参考 ANR–18–欧元–0016 —太阳能学院）。Jaakko Peltonen 得到芬兰学院项目 313748 和 327352 的支持。

参考

[1] D. Sacha、L. Zhang、M. Sedlmair、JA Lee、J. Peltonen、D. Weiskopf、SC North 和 DA Keim，“视觉交互与降维：结构化文献分析”*IEEE 跨。可见。电脑。图形。*，卷。23，没有。1，第 241–250 页，2017 年。

[2] J. Wenskovitch、I. Crandell、N. Ramakrishnan、L. House、S. Leman 和 C. North，“在视觉分析中实现降维和聚类的系统组合”，*IEEE 可视化和计算机图形学汇刊*，卷。24，第 131–141 页，2018 年 1 月。

[3] LG Nonato 和 M. Aupetit，“用于视觉分析的多维投影：将技术与扭曲、任务和布局丰富联系起来”，*IEEE 可视化和计算机图形学汇刊*，卷。25，第 2650–2673 页，2019 年 8 月。

[4] M. Brehmer、M. Sedlmair、S. Ingram 和 T. Munzner，“可视化降维数据：分析师访谈和任务序列特征”，载于*第五届超越时间和错误研讨会论文集：可视化的新型评估方法，BELIV 2014，法国巴黎，2014 年 11 月 10 日*(H. Lam、P. Isenberg、T. Isenberg 和 M. Sedlmair 合着），第 1-8 页，ACM，2014 年。

[5] S. Lespinats、M. Aupetit 和 A. Meyer-Base，“ClassiMap：一种用于标记数据探索性数据分析的新降维技术”，*模式识别与人工智能国际期刊*，卷。29，页。150505235857008，2015 年 5 月。

[6] J. Venna 和 S. Kaski，“作为信息检索的非线性降维”，载于*人工智能与统计学*, 第 572–579 页，2007 年。

[7] J. Venna、J. Peltonen、K. Nybo、H. Aidos 和 S. Kaski，“用于数据可视化的非线性降维的信息检索视角”，*机器学习研究杂志*，卷。11，没有。2 月，第 451–490 页，2010 年。

[8] GE Hinton 和 ST Roweis，“随机邻域嵌入”，载于*神经信息处理系统的进展*, 第 857–864 页，2003 年。

[9] L. van der Maaten 和 G. Hinton，“使用 t-SNE 可视化数据”，*机器学习研究杂志*，卷。9，没有。11 月，第 2579–2605 页，2008 年。

[10] Xin Geng、De-Chuan Zhan 和 Zhi-Hua Zhou，“用于可视化和分类的监督非线性降维”，*IEEE Transactions on Systems, Man, and Cyber netics，B 部分（控制论）*，卷。35，第 1098–1107 页，2005 年 12 月。

[11] J. Peltonen、H. Aidos 和 S. Kaski，“通过近邻检索监督非线性降维”，载于*2009 年 IEEE 声学、语音和信号处理国际会议*，第 1809–1812 页，2009 年 4 月。

[12] L. McInnes、J. Healy 和 J. Melville，“UMAP：用于降维的均匀流形近似和投影”，*arXiv:1802.03426 [cs, stat]*，2018 年 2 月。arXiv：1802.03426。

[13] J. Venna 和 S. Kaski，“非线性投影方法中的邻域保护：一项实验研究”，载于*人工神经网络国际会议*，第 485–491 页，斯普林格出版社，2001 年。

[14] K. 皮尔逊，“LIII。在最接近空间点系统的直线和平面上，”*伦敦、爱丁堡和都柏林哲学杂志和科学杂志*，卷。2，第 559–572 页，1901 年 11 月。

[15] T. Kohonen，“自组织映射”，*IEEE 会刊*，卷。78，没有。9，第 1464–1480 页，1990 年。

[16] JB Tenenbaum、V. De Silva 和 JC Langford，“非线性降维的全局几何框架”，*科学*，卷。290，没有。5500，第 2319–2323 页，2000 年。

[17] S. Lespinats、M. Verleysen、A. Giron 和 B. Fertil，“DD-HDS：一种用于可视化和探索高维数据的方法，”*IEEE 神经网络汇刊*，卷。18，没有。5，第 1265–1279 页，2007 年。

[18] P. Joia、D. Coimbra、JA Cuminato、FV Paulovich 和 LG Nonato，“局部仿射多维投影”，*IEEE 可视化和计算机图形学汇刊*，卷。17，第 2563–2571 页，2011 年 12 月。

[19] JA Lee、E. Renard、G. Bernard、P. Dupont 和 M. Verleysen，“Kullback–Leibler 散度的 1 型和 2 型混合作为基于相似性保存的降维成本函数”，*神经计算*，卷。112，第 92–108 页，2013 年 7 月。

[20] O. Kouropteva、O. Okun 和 M. Pietikäinen，“用于模式识别的监督局部线性嵌入算法”，载于*伊比利亚模式识别和图像分析会议*，第 386-394 页，施普林格出版社，2003 年。

[21] S.-q。Zhang，“增强型监督局部线性嵌入”，*模式识别字母*，卷。30，没有。13，第 1208–1218 页，2009 年。

[22] L. Zhao 和 Z. Zhang，“基于概率距离的监督局部线性嵌入分类”，*计算机与数学及其应用*，卷。57，没有。6，第 919–926 页，2009 年。

[23] C.-G。Li 和 J. Guo，“带显式映射的监督等值图”，载于*第一届创新计算、信息和控制国际会议第一卷 (ICICIC'06)*，卷。3，第 345–348 页，IEEE，2006 年。

[24] Z. Yang、I. King、Z. Xu 和 E. Oja，“重尾对称随机邻域嵌入”，载于*神经信息处理系统的进展*, 第 2169–2177 页，2009 年。

[25] RA Fisher，“在分类问题中使用多重测量”，*优生学年鉴*，卷。7，没有。2，第 179–188 页，1936 年。

[26] S. Mika、G. Ratsch、J. Weston、B. Scholkopf 和 K.-R。Mullers，“Fisher 核判别分析”，载于*用于信号处理的神经网络 IX：1999 年 IEEE 信号处理学会研讨会论文集（目录号 98th8468）*，第 41-48 页，IEEE，1999 年。

[27] D. De Ridder、M. Loog 和 MJ Reinders，“Local fisher embedding”，载于*第 17 届模式识别国际会议论文集，2004 年。ICPR 2004。*，卷。2，第 295–298 页，IEEE，2004 年。

[28] M. Sugiyama，“监督降维的局部渔民判别分析”，载于 *第 23 届机器学习国际会议论文集*, 第 905–912 页，2006 年。

[29] J. Goldberger、GE Hinton、ST Roweis 和 RR Salakhutdinov，“邻里成分分析”，载于*神经信息处理系统的进展 17*(LK Saul、Y. Weiss 和 L. Bottou 合编），第 513–520 页，麻省理工学院出版社，2005 年。

[30] R. Salakhutdinov 和 G. Hinton，“通过保留类邻域结构来学习非线性嵌入”，在*人工智能与统计*, 第 412–419 页，2007 年。

[31] K. Bunte、P. Schneider、B. Hammer、F.-M。Schleif、T. Villmann 和 M. Biehl，“有限秩矩阵学习、判别性降维和可视化”，*神经网络*，卷。26，第 159–173 页，2012 年 2 月。

[32] C. de Bodt、D. Mulders、DL Sánchez、M. Verleysen 和 JA Lee，“类感知 t-SNE：cat-SNE.”，载于*伊桑*, 2019.

[33] J. Venna 和 S. Kaski，“局部多维缩放”，*神经网络*，卷。19，第 889–899 页，2006 年 7 月。

[34] CC Aggarwal、A. Hinneburg 和 DA Keim，“关于高维空间中距离度量的惊人行为”，载于*第八届数据库理论国际会议论文集*, ICDT '01, (柏林, 海德堡), p. 420–434，施普林格出版社，2001 年。

[35] JA Lee 和 M. Verleysen，“降维方法的两个关键属性”，载于 *计算智能和数据挖掘 (CIDM)，2014 年 IEEE 研讨会*，第 163–170 页，IEEE，2014 年。

[36] M. Vladymyrov 和 MA Carreira-Perpinan，“熵亲和力：属性和高效数值计算”，载于*国际机器学习联盟 (3)*, 第 477–485 页，2013 年。

[37] JA Lee、DH Peluffo-Ordóñez 和 M. Verleysen，“随机邻域嵌入中的多尺度相似性：在保留局部和全局结构的同时降低维度”， *神经计算*，卷。169，第 246-261 页，2015 年 12 月。

[38] J. Nocedal 和 SJ Wright，*数值优化*. 施普林格运筹学系列丛书，纽约：施普林格，1999 年。

[39] Z. Yang、J. Peltonen 和 S. Kaski，“用于可视化的邻居嵌入的可扩展优化”，载于*机器学习国际会议*, 第 127–135 页，2013 年。

[40] L. Van Der Maaten，“使用基于树的算法加速 t-SNE”，*机器学习研究杂志*，卷。15，没有。1，第 3221–3245 页，2014 年。

[41] J.维纳，*相似结构视觉探索的降维*. 博士论文，赫尔辛基科技大学，埃斯波，2007 年。OCLC：231147068。

[42] M. Fanty 和 R. Cole，“语音字母识别”，载于*神经信息处理系统的进展*, 第 220–226 页，1991 年。

[43] D. Dua 和 E. Karra Taniskidou，“UCI 机器学习库”，2017 年。

[44] E. Alpaydin 和 C. Kaynak，“级联分类器”，*控制论*，卷。34，没有。4，第 369–374 页，1998 年。

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay，“Scikit-learn：Python 中的机器学习”，*机器学习研究杂志*，卷。12，第 2825–2830 页，2011 年。

[46] L. McInnes、J. Healy、N. Saul 和 L. Grossberger，“Umap：均匀流形近似和投影”*开源软件杂志*，卷。3、没有。29，页。861, 2018.

[47] B. Colange，“Classnerv”。https://doi.org/10.5281/zenodo.4094851，2020 年 10 月。

[48] F. Degret 和 S. Lespinats，“圆形背景减少了天真的读者对多维尺度结果的误解”，在 *MATEC 会议网络*，卷。第 189 页 10002，EDP 科学，2018 年。