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重要的，需要避免。当已知相关类的数据注释可用时，可以使用它来指导嵌入，以避免加剧类分离的扭曲。本文介绍的监督映射方法，称为ClassNeRV，提出了一个原始的应力函数，该函数考虑了类注释，并根据假邻居和错过的邻居来评估嵌入质量。ClassNeRV共享来自随机邻居嵌入(SNE)的一系列方法的理论框架。与之前的方法相比，我们的方法有一个关键的优势:在文献中，监督方法经常以扭曲数据邻居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了更好的结构保存。实验表明，ClassNeRV可以同时保持邻居结构和类分离，优于9种最先进的替代方案。

1介绍

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相似性度量。l DR可以利用课堂信息进行监督。因此，监督方法从数据的相对位置(也被非监督方法使用)和从类标签计算映射。DR技术I 2 3可用于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。当映射有标签的数据时，有两个相互矛盾的目标:

分类是典型的监督DR技术:强调类别分离，并在嵌入空间中使用分类精度进行测量。

探索性数据分析是典型的无监督DR技术，它在不了解类信息的情况下运行:数据邻域结构被优先化并被测量为原始空间和嵌入空间中数据相似度之间的差异。

这些目标来源于视觉分析任务34]。除非类和数据邻域结构在数据和嵌入空间中都很好地匹配，否则它们是矛盾的: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的，而高维数据的低维嵌入会带来不可避免的失真3:假邻居是内嵌点中的邻点，但不在数据中;漏邻居是数据中的邻点，但不在内嵌点中。

在这项工作中，我们提出了ClassNeRV，一种有监督的DR技术来完成探索性分析目标，同时考虑到类信息。我们的解决方案在原理上类似于早期的ClassiMap 5]基于距离的投影，但我们的方法来自与NeRV 6Z]， SNE[]和tSNE]相同的基于良好的概率框架。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结构为代价强制类分离，例如S-Isomap [10]， S-NeRV [I]和S-UMAP[12]。图1说明了ClassNeRV的基本特征。

我们的贡献有两个方面:我们提出了ClassNeRV，它利用类信息来确保在将高维标记数据嵌入到低维空间时更好地保存类。它的应力函数，来自无监督的NeRV 6Z]，引导优化，使邻域结构不可避免的扭曲被放置在它们对类结构的危害较小的地方。通过强调对阶级之间的错误邻居和阶级内部的错误邻居的惩罚，可以避免有害的扭曲。我们还从标准的可信度和连续性质量指标[I3]中导出了两个新的类感知质量指标，以特别解释影响类保存的扭曲。

2. Related Work

无人监督的嵌入。以前已经提出了许多线性或非线性算法，包括主成分分析(PCA) 14]，自组织映射(SOM)[15]，等距特征映射(Isomap) 16]，数据驱动的高维缩放(DD-HDS)[17]，局部仿射多维投影(LAMP)[18]和均匀流形逼近和投影(UMAP)[12]。在各种各样的技术中，邻域嵌入(NE)技术在保持邻域结构和计算时间方面是有效的。他们的概率框架也提供了一个理论背景，以解释获得的地图在一个邻域检索任务[Z]。NE方法计算，对于每一对点i, j，点j到点i的邻域的概率隶属度，有时称为相似度。这些隶属度分别在数据空间和嵌入空间中计算。通过最小化这两个空间之间的隶属概率的差异来获得映射。这些方法包括随机邻域嵌入(SNE) 8、t-分布式SNE (t-SNE) 9、Jensen Shannon嵌入(JSE) 19和邻域检索可视化(NeRV) 6Z]。SNE和t-SNE的区别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NeRV都扩展了SNE，以控制错误邻居和遗漏邻居之间的平衡。NeRV和JSE的这种可调性使它们最适合引入监督。

监督嵌入。监督映射方法主要侧重于类分离，在映射前或映射中修改数据邻居结构。其中许多方法通常减少类之间的数据相似性，并增加类内部的数据相似性，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，如监督局部线性嵌入SLLE及其变体20 2122]，监督异构映射(S-Isomap) 023，监督NeRV (S-NeRV) I或监督UMAP S-UMAP 2)，或在相似点上操作，如半监督版本的重尾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，如线性判别分析(LDA)[25]及其内核化变体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]局部适应邻域的大小，以保留基于类的分布。最后，ClassiMap优化了一个类似于局部多维标度(LMDS)[33]的应力函数，但在惩罚错误和遗漏邻居时，通过考虑类别来支持对标记数据的探索性分析。然而，基于距离的LMDS和ClassiMap在高维[34]中对范数集中现象都很敏感，而nev等NE技术使用移不变隶属度[35]来缓解这种影响。ClassNeRV从NeRV应力函数中派生出与ClassiMap相同的原理，以获得两者的好处。

3.2 ClassNeRV应力函数

我们的目标是实现既保留邻域结构又保留类的嵌入。我们建议嵌入不应该人为地将同一类中的点分开，也不应该人为地把不同类别的的点聚集起来。因此，我们需要分别惩罚更多类内的错误邻居和类间的错误邻居。为了控制这些基于阶级的扭曲，我们进一步将公式2)中的散度项划分为类内关系和类间关系。它给出了另外两个不同的基于类的权衡参数T和T(都在[0,1]中)，并导致ClassNeRV应力函数:

注意，为了保证应力函数的正性，用Bregman (B)散度代替了式(2)的KL散度。的确，KL散度仅定义为概率分布(和为1)，而在CClassNeRy的四项中，隶属度仅限于集合Sf或Sf和小于1。因此，KL散度不能直接适用于这样的隶属度(否则就不满足散度的已知性质，即非负性和不可分辨的同一性)，而Bregman散度可以直接适用于这样的隶属度。布雷格曼散度术语DB(B, B)和DB(B，)分别惩罚类内缺失邻居和假邻居，而术语DB(Bf, Bf)和DB(Bf, B)分别惩罚类间缺失邻居和假邻居。

参数T并通过加权这些项来定义ClassNeRV映射行为。TE控制类内错误邻居和错过邻居的惩罚平衡，而Tf控制类之间类似的平衡。因此，如果TE > Tf，即它的应力函数惩罚更多的类内缺失邻居和类间虚假邻居，则ClassNeRV是监督的。T-差值越大，监督水平越高，TE= 1和=0时最大。然而，使用TE<将有利于类内丢失的邻居，以及类之间的虚假邻居，鼓励同类分裂和明显的类重叠，从而破坏类的保存。如果TE =Tf，然后缩减到原始的NeRV，则ClassNeRV是无监督的，类内和类间的Bregman发散之和等于对应的KL发散与b..-B..项消去了

我们将TE和T重新参数化为T\*=(E+T)/2和E =(TE-T)/2。r\* E[0.1]控制了错误邻居和错过邻居惩罚之间的平均权衡(如NeRV中的T。e e[0,0.5]控制着监督的水平(监督越多，值越大)。逆变换为:T=7\*+ε和τ=T\*-ε。

给定T\*和e，通过使式(3)中的应力CClassNeRy相对于嵌入点的坐标最小化来获得ClassNeRV嵌入。采用多尺度优化方法[37]和准牛顿BFGS算法[38]进行优化。当前实现的复杂度为O(N2)(其中N是数据点的数量)，但是基于树的加速技术3940可以将其降低到O(Nlog N)。

3.3探索性分析监督技术的质量指标

为了评估邻里结构的保存，我们采用了可信度和连续性措施[13]，这是在探索性分析中评估无监督嵌入的标准3。这两个测量方法分别量化了给定邻域大小x的假邻域(Trustworthiness T)和缺失邻域(Continuity C)的平均水平，如下所示:

其中pij和rij分别是每个点j在数据空间和嵌入空间中每个点i的邻域中的秩，Fi(k)和Mi()是false和定义归一化系数Tmax(k)和Cmax(k)，使T和C的范围从0(理论最差映射)到1(理想映射)。经组合分析，当K < N/2时，Tmax(K)和Cmax(K)均等于kN(2N-3K-1)/2N(N-κ)(N-κ-1)/2 ifκ > N/2

为了评估类型保持性，我们还推出了两个新的衡量标准:限制在阶级关系T之间的可信度和限制在阶级关系CE内的连续性。这些类感知指示器被定义为:

请注意，这些类感知指标只考虑了部分无监督指标所考虑的失真，因此T >T和CE >C。因此，只要任何剩余的扭曲不影响类别间的可信度和类别内的连续性，它们就可以达到1。

监督降维中的另一个标准质量指标是在嵌入空间[3]中遗漏k-最近邻(k-NN)分类器的精度。当使用嵌入来进行分类时，这个指标是有意义的，但它只关注类别，不能考虑图3所示的邻域结构保留。补充材料给出了从该指标推导出的k-NN增益[32]的广泛结果。

4实验

4.1目标、数据、技术

我们在一个3D玩具数据集(Globe)和两个真正的高维数据集(Isolet 5和Digits)上说明了ClassNeRV与其他无监督和有监督DR技术相比的主要特征。Globe数据集(章节4.2)包含在三维欧氏空间R3中随机分布在单位球表面的512个数据(图2ab。这两类(蓝点和红点)对应于赤道上划分的两个半球。这些数据不能在没有失真的情况下嵌入到平面中，因此最终的映射取决于邻域(T\*)和class()惩罚之间的权衡集(参见章节3.2)。Isolet 5数据集4243 (Section4.3)包含1559个口语字母的英语发音录音，平均分布在26个类中，由617个特征描述。数字数据集44,43 (Section4.4)包含3 823张手写数字图像。8 x 8像素的图像(64D数据点)被分为10类(每个数字一个)。考虑真实类标签和随机生成标签来评估对错误标签的鲁棒性。为了简化图6中地图的可读性，考虑了500个样本的随机子集

我们将ClassNeRV与无监督PCA 4]， Isomap 16, UMAP I2]， tSNE和NeRV 6 Z，以及与有监督NCA [29]， S-Isomap [10, ClassiMap [S]和S-UMAP[12]进行比较。PCA, Isomap, NCA, tSNE, UMAP和S-UMAP的实现来自scikit-learn(版本0.22.1) 45和umap-learn(版本0.3.10)46个Python库。S-Isomap, ClassiMap, NeRV和ClassNeRV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tSNE初始化，我们使用了PCA而不是随机(因此所有方法都受益于光谱初始化)。为了与tSNE默认值p=30的可比性，我们在多尺度优化中的最终复杂度被设置为Globe的p= 32和Isolet的p=30。我们对所有技术和数据集都使用欧几里得距离。定性结果给出了散点图表示的嵌入与圆形框架遵循标准准则[348]。定量结果使用3.3节中描述的无监督和类感知指标计算。补充材料中显示了随机方法的其他运行，显示了类似的结果。

4.2两个半球的例子

图2显示了在(T\*， e)参数空间中Globe数据的ClassNeRV嵌入。图A, B, C对应原始NeRV，即没有任何监督(E = 0, TE = 7f)。在地图A上，错误的邻居是最受惩罚的(=7= 0)，允许一些错过的邻居，因此球体沿着子午线撕裂并展开。在地图C上，错过的邻居是最惩罚的(TE==1)，对应于原来的SNE映射，让假邻居，这样球体就被压扁了，混合了红色和蓝色的类。图B对应的NeRV与错过和错误邻居的平衡混合(T==0.5)。添加一些监督(e = 0.25)，地图D会惩罚中等程度的类内缺失邻居(TE =0.5)和强烈的全类虚假邻居(T\* = 0.25)，鼓励比地图a更少的阶级撕裂。相反，地图e会惩罚中等程度的类间虚假邻居(7 =0.5)和强烈的全类缺失(T\* = 0.75)，鼓励比地图C更多的阶级分离，但比地图D更多的类内虚假邻居。map F对应最大监督级别(= 0.5,TE = 1, Tf = 0)，惩罚类间虚假邻居最多(类重叠最少)和类内遗漏邻居最多(类内凝聚力最强)。为了获得ClassNeRV每个应力子项(式(3))的单独影响的详细概念，在补充材料中可获得消融研究。

图2b定量地支持这些定性观察结果。顶部热图显示了相同{T，}-参数空间中的结构保存指标:当T\*增加时，错误邻居(蓝色T)的数量增加，而错过邻居(红色C)的数量减少。底部热图显示了该空间中的类别保存指标:当监管级别e增加时，类别重叠(蓝色为T)和类别分裂(红色为CE)减少。分类热图还显示了更高的类别精度(灰色为10-NN)，与更高的类别内聚和更大的ε所鼓励的更低的类别重叠相一致。

图3和图4显示了ClassNeRV最有监督版本(=0.5)与其他技术的定性和定量比较。有监督嵌入的定性分析表明ClassNeRV(图2a，映射D, E和F)， ClassiMap(图3c)和NCA(图3b)映射都保留了两个类的实际三维线性可分性和邻接性。然而，S-Isomap(图3a)生成了一个错误的类重叠，而S-UMAP(图3d)错误地将类分开

。图4(上一行)显示ClassNeRV(-)达到了与tSNE(-)或UMAP(-)无监督技术(I)相似的结构保存水平，总体上具有更高的可信度(较少的假邻居)。它还更好地保留了类(II)，正如预期的那样对抗无监督技术。对于监督技术，ClassNeRV(-)比NCA(-)、ClassiMap(-)和S-Isomap(-)获得了更好的结构(III)和类(IV)保存。对于S-UMAP(-)，类独立指标(III)类似于ClassNeRV，除了更大的邻域大小K，这是由于S-UMAP倾向于过度分离实际上在Globe数据中相邻的类(图3d)。

4.3 Isolet数据集

在617D数据空间中计算的10-NN混淆矩阵(图5a)显示了所有准确率低于90%或与至少10%的另一个类混淆的类别(完整的混淆矩阵在补充材料中给出)。几个类部分重叠可能是由于相似的声音以音标表示，如字母F [ef]和S [es] (FS组)，字母C [si:]和Z [zi:] (CZ组)，字母A [er]， L [el]， M [em]和N [en] (ALMN组)，或字母B [bi:]， D [di:]， E [i:]， G [dzi:]， P [pi:]， T [ti:]和V [vi:] (BDEGPTV组)。图5显示了用于定性比较的嵌入。NCA(图Sd)倾向于人为地重叠类，例如QV和JP。相反，S-UMAP和S-Isomap倾向于过度分离类，完全忽略了混淆矩阵所显示的FS, CZ, ALMN, BDEGPTV组中字母的实际重叠。ClassiMap (Figure5e)更好地保留了BDEGPTV组，但未能表示其他组，如MN或CZ。NeRV设法在无监督的情况下保留组，这表明这些组的类实际上在数据空间中重叠。最后，ClassNeRV(图Sb)提供了更可靠的类表示，通过将MN、CZ、FS和BDEGPTV组中的字母相邻放在地图中，更好地显示出它们之间更强的混淆，同时保持其他字母在数据空间中的实际位置。因此，ClassNeRV映射可以更可靠地帮助领域专家发现，在这个特征空间中，字母强烈地按元音(BDEGPTV、FS或MN)分组，辅音的次要作用。

10-NN分类分数(图5)随着在这些地图上观察到的类别分离而增加。然而，由于S-Isomap在不考虑实际类在数据空间中的混淆的情况下，通过人为分离所有类达到最大值100%，因此这个分类指标无法评估类结构的保存情况。因此，它不能用于评估嵌入是否足够可靠，以支持对标记数据的探索性分析。

通过对ClassNeRV(-)与无监督技术在图4a(下一行)的定量比较可以看出，它保留了邻域结构，以及NeRV(-)和UMAP(-)，但略差于tSNE (-) (I)，对于小κ (II)，它比所有的无监督技术都更好地保留了类。图4b(底部行)表明，ClassNeRV(-)对Isolet数据的邻域结构的保护远远好于其他监督方法(III)。NCA(-)生成更多的假邻居(T <C)，而S-Isomap(-)和S-UMAP(-)生成更多的缺失邻居，与嵌入一致(图5p除了NCA(-)以外的所有技术在类保存方面都取得了类似的好结果(IV)，在小尺度和高尺度κ对ClassNeRV具有优势。

4.4手写数字

图6展示了NeRV和ClassNeRV应用于具有真标签和随机标签的数字数据集的映射。随机标签的情况允许研究监督方法在结构和类去关联时的行为。分离这些随机类会扭曲邻域，这可能被视为过度分离。我们可以从图6a和6b中观察到，对于数字数据集，依赖于真实类信息的ClassNeRV比NeRV能更好地保存类。相反，带有随机标签的ClassNeRV(图6d)不会导致随机类的明显过度分离。此外，当使用真实标签甚至随机标签时，ClassNeRV显示出比其他监督方法更好的结构和真实类的保存(图7)。

6更广泛的影响

这项工作提出了用于探索性数据分析的降维技术的改进。降维旨在支持数据科学家分析多维数据，也可用于将代表物理对象或人的高维数据可视化在二维地图中，以便公众根据相应数据的相似性获得主要对象/人组的概述。它对数据所代表的对象/人的性质是不可知论的。更好地理解大规模数据集的趋势和变化可以提高社会了解重要现象的能力。然而，降维可以产生这些对象/人的有偏见的表示，要么是由于数据本身的固有偏见(过多或不足表示的对象/人的类别，缺失或不相关的特征人为地聚集或分离(类别)对象/人)，要么是由于不可避免的投影偏差，称为扭曲，3人为地聚集在实际分离(类别)对象/人的2D表示中，或者在2D表示中人为地分离实际上相似的物体/人(类)。提出的ClassNeRV方法正是为了减少第二种类型的偏差。