摘要非线性降维方法通常用于高维数据的可视化，尽管现有的方法已被设计用于其他相关任务，如流形学习。由于任务没有明确的定义，因此很难评估可视化的质量。我们对具体的可视化任务给出了严格的定义，从而产生了可量化的优度度量和新的可视化方法。任务是在可视化的情况下进行信息检索:根据显示的相似性找到相似的数据。信息检索的精确性和召回率之间的基本权衡也可以在可视化中量化。用户需要给出丢失相似点与检索不相似点的相对成本，之后可以测量总成本。然后，我们介绍了一种新的方法NeRV(邻居检索可视化器)，该方法通过最小化代价来生成最优的可视化。我们进一步推导了监督可视化的变体;类信息在计算相似关系时被严格考虑。我们的经验表明，在可视化任务中，无监督版本优于现有的无监督降维方法，而有监督版本优于现有的有监督方法。

关键词:信息检索，流形学习，多维标度，非线性降维，可视化

1. 简介

高维数据集可视化是非线性降维方法的传统应用之一。在高维数据中，例如实验数据，其中每个维度对应一个不同的测量变量，不同维度之间的依赖关系通常将数据点限制在一个多维度远低于数据空间维数的流形上。为流形学习设计了许多方法，即寻找和展开低维流形。自2000年以来，在流形学习方面出现了一个研究热潮，目前已有许多已知的方法，至少可以成功地展开某些类型的流形。一些成功的方法包括等线映射(Tenenbaum et al.， 2000)，局部线性嵌入（LLE;Roweis和Saul，2000），拉普拉斯特征图（LE;贝尔金和尼约吉，2002a），最大方差展开(MVU;温伯格和索尔，2006)。

研究表明，多种学习方法并不一定有利于信息可视化。当为了可视化目的将输出维数固定为2时，一些方法存在严重的困难(Venna和Kaski, 2007a)。这是很自然的，因为它们被设计来寻找一个流形，而不是压缩到一个更低的维度。

在本文中，我们讨论了具体的可视化任务，即将数据投影到二维显示器上的点上。注意，这个任务与流形学习不同，如果流形的固有维数大于2，流形不能在二维中完美地表示。由于表征必然是不完美的，定义和使用表征的良性度量是至关重要的。然而，尽管对提取流形的方法进行了大量的研究，但关于良好的二维表示应该是什么样的以及如何衡量其好坏的讨论却很少。在最近一项对2000-2006年69篇降维论文的调查中(Venna, 2007)，发现28篇(40%)的论文仅将玩具或真实数据集的可视化作为质量的证明。大多数比较定量的方法都是基于两种策略之一。第一个是测量所有成对距离的守恒或者说所有成对距离的顺序。这种方法的例子包括多维标度(MDS)类型的代价函数，如Sammon's cost和Stress，将输入空间中的距离与输出空间中的距离联系起来的方法，以及评估所有两两距离保存的各种相关度量。另一种常见的质量保证策略是对低维空间中的数据进行分类，并报告分类性能。

使用上述方法来衡量可视化性能的问题是，它们与可视化的联系是不清楚的，充其量是间接的。除非可视化的目的是帮助分类任务，否则投影的分类精度揭示其作为可视化的优点是不明显的。另一个广泛采用的原则是保持成对距离，这是一个定义明确的目标;如果分析师希望使用可视化来评估所选数据点对之间的距离，这是一个合理的目标，但我们认为这不是分析师使用可视化的典型方式，至少在分析的早期阶段，当时还没有形成关于数据的假设。包括我们在内的大多数方法都是基于核心的成对距离，但我们考虑了每个成对距离的上下文，从而产生了一种更自然的评估可视化性能的方法;由此产生的方法有一个自然和严格的解释，我们将在下面和下面的部分讨论。

本文严格研究了将高维数据集投影到二维平面上以实现相似关系可视化的具体信息可视化任务。该任务与信息检索任务有一个非常自然的映射，将在第2节中讨论。作为信息检索的概念明确地揭示了精确性和召回率之间的必要权衡，使真实的相似点可见，避免虚假的相似点。一旦将成本分配给每种错误类型，就可以精确地量化权衡，一旦定义了总成本，就可以优化它，这将在第3节中讨论。然后，我们表明所得到的方法，称为邻居检索可视化器NeRV，可以进一步扩展到有监督可视化，并且无监督和有监督方法都在经验上优于它们的替代方法。NeRV包括之前的随机邻居嵌入方法(SNE;Hinton和Roweis, 2002)作为一个特殊的情况下，权衡设置，只有召回是最大化的;从而为SNE提供了一种新的信息检索解释。

2. 可视化作为信息检索在本节中，我们正式定义了特定的可视化任务;这是可视化作为信息检索任务的一种新颖形式化。我们首先在第2.1节给出简化设置的定义，然后在第2.2节对其进行泛化。

2.1二元邻域关系相似度可视化

在下面我们首先定义具体的可视化任务和它的代价函数;然后，我们证明了成本函数与传统的信息检索指标精度和查全率有关。

2.1.1任务定义:相似度可视化

设(x)1是一组输入数据样本，设每个样本i有一个输入邻域Pi，由接近i的样本组成。通常，P;可能由所有的输入样本(除了i本身)组成，它们落在i的某个半径内，或者P;可能由与i最相似的固定数量的输入样本组成。在任何一种情况下，设r为集合P的大小;相似度可视化的目标是为输入数据生成低维输出坐标(yi)-1，用于可视化信息检索。给定任意样本i作为查询，在可视化信息检索中基于样本的可视化检索;得到的结果是一组接近y的样本Qi;在可视化;我们称之为Q;输出邻域。Q通常由所有输入样本j(除了i本身)组成，其可视化坐标为y;都在y的半径内;或者Qi可能由固定数量的输入样本组成，这些输入样本的输出坐标最接近y，。在这两种情况下，设k为集合Qi中的点数。气的点数可能与P的点数不同;例如，在可视化中，如果许多点被放置在y附近，那么检索y的某个半径内的所有点，与输入空间中邻居点的数量相比，可能会产生过多的检索点。图1说明了这个设置。剩下的问题是什么是好的可视化，也就是说，代价函数是什么。表示同时处于Qi和P中的样本数量;通过NTP，(真阳性)，样本在Q;但在Pi中不存在NFp、i(假阳性)和P中的样本;而不是Q;通过NMiss，(错过)。假设用户为每个误报分配了一个成本CFp，为每个误报分配了一个成本CMiss。对于查询i，对所有数据点求和，则为



2.1.2精密度与查全率的关系

相似度可视化的代价函数(1)与传统的信息检索、精度和查全率度量密切相关。如果我们允许CMiss是相关点总数r的函数，更具体地说，CMiss(ri) = CMiss/ri，并将每个检索点的代价取为

2.1.3讨论对于高维数据集，通常不可能在低维显示上显示数据内的所有相似关系;因此，所有线性或非线性降维方法都需要权衡它们的目标是在显示器上显示哪种类型的相似关系。公式(1)确定了两种错误的代价。图2用一个简单的例子说明了这种权衡(使用第3节中介绍的方法计算)，其中三维球面在二维中可视化。如果我们取一些在显示a中，这样的检索产生很少的误报但有很多漏报，而在显示B中，这样的检索产生很少的漏报但有很多误报。权衡也可以在两种可视化的(平均)精度-召回曲线中看到，其中检索点的数量是变化的，以产生曲线。可视化A达到了更高的精度值，但在达到高召回率之前，精度下降了很多。可视化B在曲线的左端具有较低的精度，但即使达到高召回率，精度也不会下降太多。

图2:假阳性和漏检之间权衡的演示。左上:从球体表面采样的三维数据集;为了清晰起见，只显示了前半球。字形形状(大小、延伸率和角度)显示每个点的三维坐标;在线版本中的颜色显示相同的信息。底部:数据集的两个嵌入。在嵌入A中，球体被切开并折叠。这种嵌入消除了假阳性，但也有一些遗漏，因为撕裂不同一侧的点最终彼此相距很远。相比之下，嵌入B通过简单地将球体压扁来最小化未命中的次数;这导致了大量的假阳性，因为球体相对两侧的点相互映射得很近。右上:对于两个投影，输入邻域大小r = 75的平均精度-平均召回曲线，作为输出邻域大小k的函数。嵌入A具有更好的精度(在曲线的左端产生更高的值)，而嵌入B具有更好的召回性(在曲线的右端产生更高的值)。

请注意，为了量化权衡，需要同时使用精度和召回率。这需要一个足够丰富的检索模型，检索点的数量可以与相关点的数量不同，这样精度和查全率就会得到不同的值。在信息检索中，众所周知，如果相关项和检索项(这里的点)的数量相等，精度和查全率也相等。最近的“局部连续性”准则(Chen和Buja, 2009年的公式9)只是在这个约束下的精度/召回率;因此，我们给出了一个新颖的信息检索解释作为一个附带结果。这样的标准是有用的，但它只给出了一个有限的可视化质量的观点，因为它对应于一个有限的检索模型，不能完全量化精度-召回的权衡。在本文中，我们将在可视化中使用固定半径的邻域(在第2.2节中更精确地定义)，这自然会产生不同数量的检索点和相关点。

本节中介绍的简单可视化设置是一种新颖的可视化公式，作为明确定义的起点非常有用。然而，在实际应用中，它有一个缺点:过于简单的二元固定大小邻域没有考虑相关性等级。该代价函数不惩罚违反原始相似性排序的相邻样本;代价函数以相同的代价惩罚所有邻域违规行为。接下来我们将介绍一个更实用的可视化设置。

2.2连续邻域关系的相似度可视化

通过在(i)输入和(ii)输出空间中定义概率邻域，以及(iii)用概率邻域替换二进制精度和召回度量，我们推广了简单二元邻域的情况。最后将表明，对于二元邻域，解释为在邻域集中成为邻域的恒定高概率和在其他地方成为邻域的恒定低概率，测量降低到标准精度和查全率。

2.2.1检索概率模型

我们从定义输出空间中的邻域开始，并通过定义邻域点上的概率分布来实现。这样的分布可以解释为一个关于给定可视化显示的用户如何进行检索的模型。

给定查询点在显示器上的位置，yi，假设用户每次选择一个点进行检查。用qi表示用户选择y的概率。如果我们可以定义这样的概率，它们将定义一个检索y的邻居的概率模型。

qjli的形式可以通过一些公理选择和一些任意选择来定义。由于qjji是每个i在j上的概率分布，它们必须是非负的并且总和为1 / j;因此，我们可以将它们表示为qji = exp(fi,j)/Σktiexp(-fik)，其中fij E r。fij应该是y之间距离(不相似度)的递增函数;和y;;我们进一步假设fi只依赖于y, y;而不是在其他点y上。剩下的就是选择的形式作用力。一般来说，不应该有任何理由偏爱任何特定的相邻点，因此形式不应该依赖于j。然而，它可以依赖于i;我们假设它有一个简单的二次形式fij = lly;-y;ll2/o?|ly;-y;l|是欧几里得距离，正乘数1/o?允许函数对每个i以单独的速率增长。这就产生了定义

2.2.2相关性概率模型

我们将输入数据样本的简单二元邻域扩展为概率邻域，如下所示。假设用户在原始数据空间中选择查询点i的邻居，她将选择概率为pili- the pjl的点j;为原始数据定义一个相关性的概率模型，等价于i周围的一个邻域:选择这个邻域的概率越高，它与i的相关性越大。我们定义概率pj;类似于qjli, as

其中d(-，-)为原始数据中合适的差值测度，x为原始数据中y表示的点;在可视化中。有些数据集可以直接提供d(-，-)的值;否则，分析人员可以选择适合数据特征向量的差值测量。在本文的后面，我们将使用简单的欧几里得距离和更复杂的距离测量，其中包含关于数据的额外信息。

已知d(-，-)的值，上述邻域pjli的定义可以由与qli-相同的理由驱动，即如果没有关于原始邻域的其他信息可用，则给定形式的pjli是一个很好的选择。其他选择也是可能的;特别是，如果数据直接包括邻居概率，它们可以简单地用作pji-同样，如果有更准确的用户行为模型，它们可以代替qji。Pjli和qji的形式不必相同。

对于每一点i，比例参数o，控制概率pil随距离下降的速度。这些参数可以通过先验知识确定，但是在没有先验知识的情况下，通过指定邻居的选择应该有多大的灵活性来设置o是合理的。也就是说，我们将o设置为一个值，使p.，分布的熵等于logk，其中k是由用户设置的相关邻居数量的粗略上限。我们在输入和输出空间中都使用相同的相对尺度o(公式2和3)。

2.2.3成本函数

剩下的任务是衡量在给定可视化的情况下，在输出空间中完成的检索与在输入空间中定义的真实相关性的匹配程度。以上两者都是根据分布来定义的，测量的自然候选是Kullback-Leibler散度，定义为

p;qi是特定点i的邻居分布，分别在输入空间和可视化中。对于上述定义的特定概率分布，KullbackLeibler散度与精度和召回率密切相关。具体而言，对于任意查询i, Kullback-Leibler散度D(pi,qi)是召回率的泛化，D(qi,pi)是精度的泛化;对于简单的“二元”邻域定义，Kullback-Leibler散度和精度-召回措施变得等效。证明在附录A中。

我们称D(gi,pi)为平滑精度，D(pi,qi)为平滑召回率。为了评估一个完整的可视化而不是单个查询，我们以标准方式定义了聚合度量:平均平滑精度定义为E;[D(qi, pi)]，平均平滑召回率定义为E;[D(pi,qi)]，其中E表示期望，对查询(数据点i)采用均值。

平均平滑精密度和召回类似于平均精密度和召回，因为我们通常不能同时达到两者的最佳。我们回到图2，它说明了一个三维球面的非线性投影的权衡。子图A是通过最大化平均平滑精度创建的;球体已经被切开并折叠，这最大限度地减少了假阳性的数量，但也会导致一些失误，因为位于点云相对边缘的一些点原本在球体上彼此靠近。子图B是通过最大化平均平滑召回率创建的;球体被压扁，这最大限度地减少了失误的数量，因为在原始数据中彼此接近的所有点在可视化中都彼此接近。然而，随后会出现大量的假阳性，因为球体的相反边已经相互映射在对方的顶部，因此许多在可视化中看起来彼此接近的点实际上最初彼此远离。

2.2.4易于解释的替代优度测度

平均平滑精度和召回率是对可视化性能的严格和动机良好的衡量标准，但对于人类分析师来说，它们有一个实际的缺点:误差没有上限，而且规模往往取决于数据集。这些度量对于比较相同数据的几个可视化非常有用，并且将被证明是有用的优化标准，但是我们还希望有易于解释普通数字的度量。我们通过引入基于平均秩的平滑精度和召回来解决这个问题:简单地替换p定义中的距离;还有秩的qii，所以最近邻居的概率是1，次近邻的概率是2，以此类推。这对错误施加了一个上限，因为最坏的情况是数据集中的排名在可视化中颠倒了。将误差除以其上界，我们得到的测量值位于区间[0,1]，与数据无关，因此更容易解释。缺点是用等级代替距离使得测量忽略了数据中的大部分邻域结构，因此我们建议使用基于平均等级的平滑精度和召回率作为更容易解释的补充，但对平均平滑精度和召回率的区分性较差，而不是替代。

3.邻域检索可视化器(NeRV)

在第2节中，我们将相似度可视化定义为一种信息检索任务。可视化的质量可以通过两个损失函数来衡量，平均平滑精度和召回率。这些措施推广直接精度和召回措施的非二元邻域。它们的另一个优点是，它们是函数的连续可微函数输出可视化坐标。然后，很容易使用这些度量作为可视化方法的优化标准。我们现在介绍一种优化视觉信息检索性能的可视化算法。我们称该算法为邻域检索可视化器(NeRV)。

如图2所示，精度和召回率通常不能同时最小化，用户必须选择哪个损失函数(平均平滑精度或召回率)更重要，通过为遗漏分配成本和为误报分配成本。一旦分配了这些成本，可视化任务就是将总成本最小化。在实践中，假阳性与失球的相对代价被作为参数λ给出。NeRV代价函数就变成了

例如，将λ设置为0.1表示用户认为精度误差(1-0.1)/0.1 =召回中类似误差的代价为9倍。

根据输出坐标y优化代价函数(4);对于每个数据点，我们使用标准的共轭梯度算法。每次迭代的计算复杂度为O(dn2)，其中n为数据点的数量，d为投影的维数。(在我们之前的会议论文中，需要一个粗略的近似算法来提高速度;这被证明是不必要的，O(dn?)复杂度不需要任何近似。)请注意，如果输入空间中的成对距离矩阵没有直接作为数据提供，它可以像往常一样从输入特征中计算出来;这是在算法开始时完成的一次性计算，需要O(Dn?)时间，其中D是输入维数。

一般来说，NeRV优化用户定义的成本，在平均光滑精度和平均光滑召回率之间形成权衡。如果在式(4)中设λ = 1，我们得到随机邻居嵌入的代价函数(SNE;见Hinton和Roweis, 2002)。因此，作为附带结果，我们得到了对SNE的新解释，它是一种使平均平滑召回最大化的方法。

3.0.5实用的优化建议

在计算输入数据的距离矩阵之后，我们缩放输入距离，使平均距离等于1。我们使用在单位正方形上的随机投影作为算法的起点。事实证明，即使是这种简单的选择也比其他选择带来了更好的结果;当然，还可以使用更智能的初始化，例如使用主成分分析投影数据。

为了加快收敛速度并避免局部极小值，我们应用进一步的初始化步骤:我们运行十轮共轭梯度(每轮两个共轭梯度步骤)，在每轮之后降低公式(2)和(3)中使用的邻域缩放参数o。最初，我们将o设置为输入数据直径的一半。我们线性地减少它们，使最终值使p分布的熵等于有效邻居数k，这是第2.2节中推荐的选择。这个初始化步骤与算法的其余部分在每次迭代中具有相同的复杂度O(dn2)。在这个初始化阶段之后，我们执行20个标准共轭梯度步骤。

4. 使用NeRV进行无监督可视化

将神经网络应用于无监督降维很容易。与任何无监督分析一样，分析人员首先为输入数据选择合适的无监督相似度或距离度量;对于向量值输入数据，这可以是标准欧几里得距离(我们将在这里使用)，也可以是领域知识建议的其他度量。一旦分析人员通过为a选择一个值来指定精度和召回率的相对重要性，NeRV算法就会根据给定的距离计算嵌入。在本节中，我们将进行广泛的实验，比较NeRV与其他降维方法在几个数据集(包括基准数据集和现实生活中的生物信息学数据集)的无监督可视化上的性能。在接下来的小节中，我们将描述比较方法和数据集，简要讨论实验方法，并给出结果。

4.1无监督可视化的比较方法

对于无监督可视化任务，我们比较了NeRV与以下无监督非线性降维方法的性能:主成分分析(PCA;:主成分分析(PCA;Hotelling, 1933)，度量多维尺度(MDS;见Borg和Groenen, 1997)，局部线性嵌入(LLE;Roweis和Saul, 2000)，拉普拉斯特征映射(LE;Belkin和Niyogi, 2002a)，基于hessian的局部线性嵌入(HLLE;Donoho和Grimes, 2003)，等线图(Tenenbaum等人，2000)，曲线成分分析(CCA;Demartines和Hérault, 1997)，曲线距离分析(CDA;Lee等人，2004)，最大方差展开(MVU;Weinberger和Saul, 2006)，标志最大方差展开(LMVU;Weinberger等人，2005)，以及我们之前的方法local MDS (LMDS;Venna和Kaski, 2006)。

主成分分析;Hotelling, 1933)发现线性投影最大限度地保留了数据中的方差。更技术地，投影方向可以通过求解输入数据点的协方差矩阵C的特征值和特征向量来求得。对应于两个或三个最大特征值的特征向量被收集到矩阵a中，数据点x;可以用y;=Ax;来投影，其中y是得到的x;的低维表示。PCA与线性多维缩放(线性MDS，也称为经典缩放)密切相关;Torgerson, 1952;Gower, 1966)，它试图找到保持平方距离的低维坐标。可以证明(Gower, 1966)，当所求解的维数相同且距离度量为欧几里得时，原始数据在PCA子空间的投影等于线性MDS发现的点的构型。这意味着PCA试图保持数据点之间的平方距离，而线性MDS找到的解决方案是原始数据的线性投影。

传统多维标度(MDS;(参见Borg and Groenen, 1997)存在于几种不同的变体中，但它们都有一个共同的目标:找到一种输出坐标的配置，以保留输入数据的成对距离矩阵。对比实验选用了最简单的非线性MDS方法——度量MDS;其代价函数(Kruskal, 1964)，称为原始应力，为

和x;在输入空间中，d(yi,y;)是它们对应的表示(位置)之间的距离y;输出空间中的y。这个代价函数相对于表示yi是最小化的。

Isomap (Tenenbaum et al.， 2000)是MDS的一个有趣变体，它同样找到与给定距离矩阵匹配的输出坐标的配置。不同之处在于Isomap不以简单的欧几里得距离计算成对的输入空间距离，而是沿着数据流形计算测地线距离(技术上，沿着连接所有k个最近邻居形成的图)。给定这些测地线距离，输出坐标由标准线性MDS找到。当找到这些输入距离的输出坐标时，原始数据中的流形结构就展开了;已经证明(Bernstein et al.， 2000)该算法渐近地能够恢复某些类型的流形。我们在实验中使用了http://isomap.stanford.edu上提供的isomap实现。

曲线分量分析(CCA;Demartines和Hérault, 1997)是MDS的一种变体，它试图只保留可视化中彼此靠近的点之间的距离。这是通过对MDS代价函数(5)中的每个项进行加权来实现的，该系数依赖于可视化中相应的成对距离。在我们使用的实现中，系数只是一个阶跃函数，如果距离低于预定阈值，它等于1，如果距离更大，它等于O。

曲线距离分析(CDA;Lee et al.， 2000,2004)是CCA的延伸。其思想是用测地线距离替换原始空间中的欧几里得距离，方法与等距地图算法相同。否则算法不变。

本地MDS (LMDS;Venna和Kaski, 2006)是我们早期的方法，它是CCA的扩展，重点关注具有可调成本函数权衡的局部邻近性。它可以被视为神经网络概念发展的第一步。

局部线性嵌入(LLE;Roweis and Saul, 2000)算法基于数据流形足够平滑且采样足够密集的假设，使得每个数据点都靠近流形上的局部线性子空间。LLE对整个数据流形进行局部线性近似:LLE首先估计每个数据点的局部坐标系统，通过计算线性系数，尽可能地从其k个最近的邻居中重建数据点。为了展开流形，LLE找到尽可能保留先前估计的局部坐标系的低维坐标。技术上，LLE首先最小化重构误差E(W)=Σ，Ix-Σ，Wijxll2相对于系数Wi,j，在约束条件下，如果i和j不是邻居Wij = 0， Σ，Wij = 1。给定权重，接下来通过对低维表示Y最小化E(Y) = Σ: lly:-Σ，Wi,iYll2来找到点的低维配置;每个数据点。

拉普拉斯特征映射(LE;参见Belkin和Niyogi, 2002a)使用图嵌入方法。形成一个无向k-近邻图，其中每个数据点是一个顶点。如果i的k个最近邻居中有jis，则点i和点j由一条权值Wij=1的边连接，否则边权值设为零;这种简单的加权方法在实践中表现良好(Belkin and Niyogi, 2002b)。为了找到图的低维嵌入，算法试图将图中连接的点尽可能地相互靠近，并且不关心其他点发生了什么。从技术上讲，对于低维点位置yi，它最小化Σlly:-yl2W， = y7Ly，其中L = D- w是拉普拉斯图，D是元素Dii =ΣWij的对角矩阵。然而，这个代价函数有一个不受欢迎的平凡解:将所有点放在同一个位置将使代价最小化。这可以通过添加suit来避免能力约束。在实践中，低维构型是通过求解广义特征值问题Ly = λDy (Belkin and Niyogi, 2002a)得到的。最小的特征值对应于平凡解，但特征向量对应于下一个最小的特征值给出拉普拉斯特征映射解。

基于hessian的局部线性嵌入(Hessian-based local linear embedding, HLLE;Donoho and Grimes, 2003)算法类似，但拉普拉斯L被Hessian H取代。最大方差展开算法(MVU;Weinberger和Saul, 2006)将降维问题表示为半定规划问题。展开折叠国旗的一种方法是把它的四个角拉开，但不要用力到撕破国旗。MVU将这一想法应用于投影流形:投影最大化方差(拉开距离)，同时保持邻居之间的距离(没有眼泪)。局部距离保持的约束可以用映射的Gram矩阵K表示。最大化映射的方差等价于在一组约束条件下最大化K的迹，这可以用半定规划来实现。

MVU的一个显著缺点是当数据点n很大时，为n x n个矩阵求解半定程序所需的时间。Landmark MVU (LMVU;Weinberger et al.， 2005)通过显著减小半定规划问题的大小来解决这个问题。像LLE一样，LMVU假设数据流形足够平滑且采样密集，因此它在局部近似线性。LMVU不像MVU那样直接嵌入所有数据点，而是随机选择m«n个输入作为所谓的地标。由于局部线性假设，其他数据点可以使用线性变换从标志近似重建。由此可见，格拉姆矩阵K可以用标志间内积的m × m子矩阵来近似表示。因此，我们只需要优化超过m × m个矩阵，一个小得多的半定程序。最近其他加速MVU的方法包括基于拉普拉斯图的矩阵分解(Weinberger et al.， 2007)。

除了上述比较方法外，最近在降维方面的其他工作包括最小体积嵌入(MVE;Shaw和Jebara, 2007)，它与MVU类似，但MVU最大化了Gram矩阵的整个迹(所有特征值的和)，MVE最大化了前几个特征值的和，最小化了其余的和，以便在降维后剩下的几个维度中保留最大的特征谱能量。在实际应用中，对所得准则的变分上界进行了优化。最近，van der Maaten等人(2009)在分类准确性和我们的旧标准可信度-连续性方面比较了一些无监督方法。

4.2无监督可视化数据集

我们使用了两个合成基准数据集和四个现实生活数据集进行实验。平面s曲线数据集是嵌入三维空间的s形二维曲面上的人工采样集。对于非线性降维方法来说，几乎完美的二维表示应该是可能的，因此这个数据集可以作为完整性检查。噪声s曲线数据集在其他方面与普通s曲线数据集相同，但每个数据点都添加了显著的球形正态分布噪声。结果是一团点，原来的s形很难用肉眼观察。

人脸数据集由40个不同的人的10张不同的人脸图像组成，总共有400张图像。对于给定的主题，图像在光线和面部表情方面有所不同。每张图像的大小是64 x 64像素，每个像素有256个灰度级。该数据集可在http://www.cs.toronto.edu/~roweis/data.html上下载。

小鼠基因表达数据集是来自不同小鼠组织的基因表达谱的集合(Su et al.， 2002)。在45个组织中测量了超过13000个小鼠基因的表达。我们使用了一种非常简单的过滤方法，类似于Su等人(2002)最初使用的方法，来选择用于可视化的基因。在45个组织(维度)中至少有一个明确表达的小鼠基因(Affymetrix芯片的平均差异，AD > 200)中，随机选取了1600个基因(点)样本。在此之后，每个组织中的方差归一化为统一。

基因表达纲要数据集是人类基因表达阵列的大型集合(http://dags.stanford.edu/cancer;Segal et al.， 2004)。由于目前所有方法的实现都不能容忍缺失数据，因此我们删除了缺失值的样本。首先，我们从300多个基因阵列中删除了缺失的基因。然后我们删除了仍然缺少值的数组。这样得到的数据集包含1278个点和1339个维度。

海水温度时间序列数据集(Liitiäinen和Lendasse, 2007)是几年来海水每周温度测量的时间序列。每个数据点是一个52周的时间窗口，为下一个数据点向前移动一周。总共有823个数据点和52个维度。

4.3无监督实验的方法

将NeRV的性能与4.1节中描述的11种无监督降维方法进行了比较，即主成分分析(PCA)、度量多维尺度(MDS)、局部线性嵌入(LLE)、拉普拉斯特征映射(LE)、基于hessian的局部线性嵌入(HLLE)、等线图、曲线分量分析(CCA)、曲线距离分析(CDA)、最大方差展开(MVU)、地标最大值方差展开 （LMVU） 和局部 MDS （LMDS）。LLE， LE， HLLE， MVU， LMVU 和 ISOMAP都是用开发者的代码计算出来的;MDS, CCA和CDA使用了我们的代码。

4.3.1优度测度

我们使用了四对性能测量来比较这些方法。第一对是平均平滑精度-平均平滑召回，即我们的可视化质量的新措施。输入邻域的规模固定为20个相关邻域(见章节2.2)。尽管我们认为，如第2节所述，平滑精度和平滑召回比精度和召回更复杂，但我们也绘制了标准的平均精度-平均召回曲线。通过将原始数据中一个点的20个最近的邻居固定为相关项的集合，然后将从可视化中检索到的邻居的数量在1到100之间变化，绘制每个数字的平均精度和召回率。我们的第三对测量是我们新测量的基于排名的变体，基于平均排名的平滑精度-基于平均排名的平滑召回。回想一下，我们介绍了基于排名的方法变量更容易解释，但不太有区别，替代平均平滑精度和平均平滑召回。输入邻域的规模再次固定为20个相关邻域。

我们的第四个衡量标准是可信度-连续性(Kaski et al.， 2003)。这些测量背后的直观动机与本文中一样，是在精确度和召回率之间进行权衡，但这些测量的定义更特别。当时，我们没有明确的联系信息检索，这使得NeRV特别有吸引力，我们没有优化措施。然而，可信度和连续性现在可以被用作可视化质量的部分独立度量。为了计算可信度和连续性，我们使用每个点包含20个最近邻居的邻域。作为第五种措施，当数据类可用时，我们使用给定显示的分类误差，使用标准的k-最近邻分类器，其中我们设置k=5。

4.3.2参数的选择

每当我们需要为任何方法选择一个参数时，我们都使用相同的标准，即从新的基于秩的度量中计算的f度量。也就是说，我们选择了产生2(P.R)/(P+R)的最大值的参数，其中P和R是基于平均排名的平滑精度和召回率。

许多方法都有一个参数k，表示用于构建邻域图的最近邻居的数量;对于每种方法和每个数据集，我们测试k值从4到20，并选择产生最佳f测量值的值。(对于MVU和LMVU，我们使用较小的参数范围来节省计算时间。MVU k范围为4 ~ 6;LMVU k为3 ~ 9)。例外是本地MDS (LMDS)，这是我们自己早期的方法之一，以及NeRV，我们只是将k设置为20而没有优化它。可能具有局部最优值的方法用不同的随机初始化运行了五次，并选择了最佳运行(同样，在f度量方面)。

4.4无监督可视化结果

接下来我们将展示几个集合的可视化，并定量测量几个集合的结果。我们首先在图3中展示一个用于普通s曲线数据集的NeRV可视化示例。在本节的后面，我们将展示一个合成人脸数据集的NeRV可视化(图8)，以及在人脸数据集的4.6节中展示真实人脸图像(图11)。定量结果分布在四个图中(图4-7)，每个图包含一对测量值和所有六个数据集的结果。

我们首先展示了平均平滑精度-平均平滑回忆的曲线，即与信息检索可视化形式化相关的损失函数。结果如图4所示。NeRV与local MDS (LMDS)形成以A为参数的曲线，NeRV的取值范围为0 ~ 1.0,LMDS的取值范围为0 ~ 0.9。NeRV显然是在所有六个数据集上表现最好的方法，这当然是意料之中的，因为NeRV直接优化了这些测量的线性组合。LMDS具有相对较好的平均平滑精度，但在平均平滑召回方面表现不佳。简单度量MDS也是一种相当不错的方法。

因为我们将可视化定义为一项信息检索任务，所以很自然地也尝试了现有的信息检索性能度量，即平均精度和平均召回率，即使它们没有考虑2.1节中讨论的相关性等级。标准平均精度-平均召回率曲线如图5所示;对于NeRV和LMDS，我们显示了由f测度选取的单个a值的曲线，如第4.3节所述。即使使用这些粗略的度量，NeRV也表现出出色的性能:就曲线下的面积而言，NeRV在四个数据集上是最好的，CDA和CCA在一个数据集上都是最好的。

接下来，我们绘制了我们更容易解释但不那么有区别的可视化性能的替代测量。基于平均排名的平滑精度-基于平均排名的平滑召回率曲线如图6所示。这些测量值介于0和1之间，因此可能更容易在数据集之间进行比较。通过这些措施，NeRV再次在所有数据集上表现最好;LMDS在处理海水温度数据方面也表现良好。

最后，我们绘制了可信度-连续性曲线，如图7所示。结果与新的基于排名的测量方法相当相似:NeRV在所有数据集上表现最好，LMDS也表现得很好，特别是在海水温度数据上。

4.4.1用已知的底层流形进行实验

为了进一步测试这些方法恢复数据中固有的邻域结构的能力，我们研究了一个合成人脸数据集，其中已知的底层流形定义了每个点的相关项(邻域)。SculptFaces数据包含698张人脸合成图像(每张图像大小为64 x 64像素)。光照的姿态和方向被系统地改变，以在图像空间中创建一个流形(http://web.mit.edu/cocosci/isomap/datasets.html;特南鲍姆等人，2000)。我们使用原始像素数据作为输入特征。

用于生成图像的姿态和照明参数是可用的。这些参数定义了嵌入在非常高维图像空间中的人脸的流形。对于任何一张脸图像，相关的其他面是那些在姿态和光照参数方面是相邻的;我们使用姿态和光照空间中的欧几里得距离定义了地面真相邻域，并将地面真相邻域的尺度固定为20个相关邻域(参见章节2.2)。

与本节的所有实验一样，我们对这些数据运行了所有方法，然后计算了四条性能曲线(平均平滑精度-平均平滑召回率，平均精度-平均召回率，基于平均秩的平滑精度-基于平均秩的平滑召回率，以及可信度连续性)，使用姿态和照明空间中的邻域作为基本真理。结果如图8所示。尽管输入空间的维数很高，而且在三维到二维的多维度上，NeRV能够很好地恢复结构。根据我们提出的可视化性能测量，平均平滑精度和召回率，NeRV是最好的;MDS和本地MDS性能也很好。在简单平均精度和平均召回率方面，NeRV是第二好的，CDA略好。在基于排名的度量方面，NeRV在精度方面是最好的;LE和MDS得到了基于平均秩的最佳结果平滑的回忆;本地MDS和CDA性能也很好。当用可信度和连续性来衡量表现时，NeRV在可信度方面是最好的，而MVU和Isomap获得了最高的连续性。

总的来说，NeRV在这些无监督可视化任务中是最好的，尽管它不是所有任务中最好的，而且在一些任务中它有激烈的竞争。

4.5无监督分类比较

对于样本分类可用的数据集，我们还遵循传统的方法来评估可视化:我们测量基于可视化的样本分类的好坏。

这里所有的方法都是无监督的，即在计算可视化时不使用样本的类标签。方法的参数再次选择，如4.3.2节所述。方法采用k-最近邻分类精度(k= 5)进行评估，即可视化中的每个样本在可视化中通过其k个最近邻的多数票进行分类，并将分类结果与ground truth标签进行比较。

我们使用四个基准数据集，它们都包含类标签，来比较这些方法的性能。表1总结了这些数据集。对于所有数据集，我们在实验中使用了随机选择的1500个样本子集，以节省计算时间。

字母识别数据集(表示为letter)来自UCI机器学习库(Blake和Merz, 1998);它是一个16维数据集，有26个类，它们是字母表中26个大写字母的4 x 4张图像。这些字母是基于20种不同的字体，这些字体被扭曲以产生最终的图像。

音素数据集(表示音素)取自LVQ-PAK (Kohonen et al.， 1996)，由由20维特征向量表示的音素样本和表示实际表示的音素的类标签组成。总共有13个班级。

landsat卫星数据集(标记为landsat)来自UCI机器学习库(Blake和Merz, 1998)。每个数据点为36维向量，对应于4个光谱波段测量的3 × 3卫星图像;点的类别标签表示图像中的地形类型(6种可能性，例如红土)。

TIMIT数据集来自DARPA TIMIT语音数据库(TIMIT)。它与LVQ-PAK的音素数据相似，但特征向量是12维的，总共有41个类。结果分类错误率如表2所示。NeRV在四个数据集中的两个上是最好的，在第三个数据集上是第二好的(我们的旧方法LocalMDS是最好的)。CDA是最好的。

4.6 NeRV、联合概率和t-分布

最近，基于随机邻居嵌入(SNE)， van der Maaten和Hinton(2008)提出了一种改进的t-SNE方法，该方法在无监督实验中表现良好。t-SNE与SNE相比有两个变化;在本节中，我们将描述这些更改，并说明可以对NeRV进行相同的更改，从而产生我们称为t-NeRV的变体。我们为t-NeRV和t-SNE提供了一种新的信息检索解释。我们首先分析t-SNE和原始随机邻居嵌入之间的差异。原始SNE最小化Kullback-Leibler散度的和其中pji和qji由式(3)和(2)定义。我们在2.2节中表明，该代价函数具有信息检索解释:它对应于检索查询点的邻居的平均平滑召回。t-SNE方法做了两个改动，我们将在下面讨论。

4.6.1基于联合概率的代价函数

t-SNE的第一个变化是代价函数:t-SNE最小化代价函数的“对称版本”，定义为

其中pij和qi j现在是i和j的联合概率，因此Σij pij = 1, qi也是如此，j。术语“对称”来自于这样一个事实，即关节概率是以一种特定的方式定义的，其中pij = Pj,i和qi,j = qj,i，注意，这并不需要对所有关节概率的定义都是如此。

4.6.2联合概率的定义

t-SNE的第二个变化是，联合概率的定义方式与公式(3)和(2)中产生的条件概率不完全相同。联合概率定义为

． 其中pi和p由式(3)计算，n是数据集中数据点的总数，前一个方程的目的是确保在输入空间中，即使是离群点也会有一些其他点作为邻居。

后一个方程表示，在可视化中，联合概率根据具有一个自由度的(归一化)t分布下降，这旨在帮助解决拥挤问题:因为小维邻域的体积增长速度比高维邻域的体积增长速度慢，邻域最终在可视化中被拉伸，以至于中等距离的点对放置的距离过远。这往往会导致数据在可视化中心聚集。由于t分布的尾部比高斯分布更重，因此使用这种分布的qi,j使可视化更少地受到中等距离的点对的放置的影响，因此能够更好地关注数据的其他特征。

4.6.3新方法:t-nerv

我们可以很容易地将上述变化应用到NeRV中的成本函数;我们称这种变异为t-NeRV。我们定义代价函数为

其中p和q是由pij和qij定义的i和j上的联合分布，单个联合概率由式(6)和式(7)给出。

可以表明，这种变化的代价函数再次具有自然的信息检索解释:它对应于两步信息检索任务的平滑精度和平滑召回之间的权衡，其中分析师查看可视化并(步骤1)选择一个查询点，然后(步骤2)为查询点选择一个邻居。选择一个查询点i的概率取决于有多少其他点接近它(也就是说，它取决于Σjqi.j)，而选择一个邻居的概率通常取决于邻居与查询的相对距离。这两个选择都是基于可视化完成的，并通过平滑精度和平滑回调与基于输入空间定义的相关查询对和邻居进行比较。参数A再次控制了精度和召回率之间的权衡。

D(p,q)与两步检索任务召回之间的联系可以用附录a中类似的证明来表示，主要区别是条件分布pili和q被联合分布pij和qij所取代，然后对i和j求和。D(q,p)与精度之间的联系可以类似地表示。

作为特例，在上述代价函数中设置a =1，即只优化两步检索任务的平滑召回，得到t-SNE的代价函数。因此，我们为t-SNE提供了一种新颖的信息检索解释，作为一种最大限度地提高查询点及其邻居查全率的方法。

NeRV和t-NeRV之间的主要概念区别是，在t-NeRV中，在可视化和输入空间中选择查询点的概率分别取决于可视化和输入空间中的密度;在NeRV中，所有潜在的查询点都被平等对待。应该使用哪种查询点处理方法取决于分析人员的任务。此外，NeRV和t-NeRV在概率的技术形式上存在差异，即使用t分布还是高斯分布等。

t-NeRV方法可以根据可视化坐标y进行优化;点，通过共轭梯度优化，如在NeRV;计算复杂度也是一样的。

4.6.4比较

我们简单地比较t-NeRV和NeRV在Faces数据集中的情况。设置与之前的对比实验相同(图4-7)。对于t-NeRV，我们使用有效邻居数k = 40计算关节概率pij;这与tSNE作者使用的perplexity值相对应.

图9显示了四个无监督评估标准的结果。从平均平滑精度和平均平滑召回度量来看，t-NeRV在召回方面表现较差。基于排名的测量也得出了类似的结果;然而，t-NeRV在基于平均秩的平滑精度方面做得相当好。可信度-连续性曲线类似于基于排名的测量方法。平均精度与平均召回率的曲线表明，t-NeRV对于小的召回值(即小的检索邻域)确实实现了更好的精度，而NeRV对于大的检索邻域略好。如上所述，这些措施对应于NeRV的信息检索解释，这与t-NeRV略有不同。图9 E显示了t-NeRV意义上的平均平滑精度/召回率，其中t-NeRV自然表现相对更好。

最后，我们计算了关于图像中人物身份的k-最近邻分类错误率(使用k= 5)。NeRV (λ= 0.3)的误差率为0.394,t-NeRV (λ= 0.8)的误差率为0.226。这里t-NeRV更好;这可能是因为它避免了在可视化中心附近拥挤样本的问题。

图10-12显示了人脸数据集的可视化示例。首先，我们展示了一个表现良好的比较方法(CDA;图10);它很好地安排了人脸，将同一个人的图像保存在一个区域;然而，每个人的区域都是分散的，并且与其他人很接近，因此在展示中人与人之间没有很强的分离。NeRV，在这里优化以最大化精度，使每个人的聚类明显更紧密(图11)，这产生更好的邻居面部图像检索。然而，NeRV在这里将几个人放置在可视化的中心，彼此靠近。t-NeRV可视化，再次优化以最大化精度(图12)减少了这种行为，更均匀地放置了面簇。

总的来说，t-NeRV是NeRV的一个有用的替代公式，并且可能对数据集有用，特别是在可视化中心附近的拥挤是一个问题。

5. 使用NeRV进行监督可视化

在本节中，我们将展示如何使用NeRV进行监督可视化。监督可视化所需要的就是以监督的方式计算输入空间距离。然后将距离插入到NeRV算法中，可视化照常进行。注意，分两个步骤模块化地进行可视化是一个优势，因为如果需要，以后可以轻松地更改其中任何一个步骤中使用的算法。

方便的是，存在从标记的数据样本中学习监督度量的严格方法。监督度量的学习最近被广泛地用于分类目的和一些半监督任务，既有简单的线性方法，也有复杂的非线性方法;例如，参见Xing等人(2003)、Chang和Yeung(2004)、Globerson和Roweis(2006)和Weinberger等人(2006)的作品。任何这样的度量原则上都可以用来计算NeRV的距离。这里我们使用早期的一个，它是灵活的，可以直接插入到NeRV中，即学习度量(Kaski et al.， 2001;Kaski和Sinkkonen, 2004;Peltonen et al.， 2004)最初是为数据探索任务而提出的。

我们将用监督距离计算的NeRV称为“监督NeRV”(SNeRV)。NeRV的信息检索解释延续到SNeRV。SNeRV可以被视为一种新的、有监督的随机邻居嵌入，但更一般地说，它就像无监督的NeRV一样，在信息检索的精度和召回率之间进行了灵活的权衡。

SNeRV有一个很有用的特性，它可以直接计算未标记训练点的嵌入，也可以直接计算标记训练点的嵌入。相比之下，一些有监督的非线性降维方法(Geng等，2005;Liu et al.， 2005;Song et al.， 2008)只给出标记点的嵌入;对于未标记的点，映射是近似的，例如通过插值或训练在一定的参数设置下一个神经网络。对于SNeRV，这种近似是不需要的。(另一方面，经过训练的神经网络不仅可以嵌入未标记的训练点，还可以嵌入之前未见过的新点;如果需要这种泛化，则可以为SNeRV学习相同类型的近似映射。)

在接下来的小节中，我们将介绍距离计算的细节，然后描述实验比较，表明SNeRV优于几种现有的监督方法。

5.1 NeRV的监督距离

SNeRV的输入空间距离是使用学习度量计算的(Kaski等人，2001;Kaski和Sinkkonen, 2004;Peltonen et al.， 2004)。这是一种形式主义，特别适合于所谓的“有监督的无监督学习”，其中最终目标仍然是在无监督学习中进行发现，但度量通过强调有用的特征来帮助集中分析，而且，对不同的样本进行局部不同的分析。学习度量以前已经应用于聚类和可视化。

简而言之，学习度量是一个黎曼拓扑保持度量，它根据类分布的变化来度量距离。类分布估计通过从标记样本的条件密度估计。拓扑保存有助于泛化到新的点，因为类信息不能覆盖输入空间拓扑。在这个度量中，我们可以计算任意两个数据点之间的输入空间距离，从而可视化具有NeRV的点，无论它们是否有已知的标签。

5.1.1定义

学习度量是所谓的黎曼度量。这样的度量是以局部的方式定义的，在两个(无穷小的)近点之间，它有一个简单的形式，并且这个简单的形式通过路径积分扩展到全局距离。在学习度量中，两个临近点x1和x2之间的距离平方由二次形式给出,

其中J(x)为Fisher信息矩阵，描述条件类分布对输入特征的局部依赖关系，即:

这里c是类，p(clx)是点x处的条件类概率。其思想是，局部距离沿条件类分布p(cx)变化最大的方向增长最大。可以证明，对于临近点，二次形式(9)等价于Kullback-Leibler散度D(p(cx1)，p(cx2))。

两个遥远点X1和X2之间的一般距离dz(X1,X2)是用黎曼度量的标准方式定义的:距离是局部距离上的最小路径积分，其中最小值是连接X1和X2的所有可能路径上的最小值。请注意，在黎曼度规中，直线路径可能不会产生最小距离。

以上述方式定义的学习度量满足任何度量所需的三个标准:距离dz是非负的、对称的，并且满足三角形不等式。由于学习度量距离被定义为最小路径积分，它们保留了输入空间的拓扑结构;粗略地说，如果两点之间的距离很小，那么它们之间一定有一条路径，沿着整个路径的距离都很小。

5.1.2实际计算

为了使用Fisher信息矩阵J(x)计算局部距离，我们需要对条件概率分布p(c)x)进行估计。我们通过优化数据的标记高斯密度的判别混合来学习分布(Peltonen et al.， 2004)。条件密度估计的形式为

其中高斯数K、质心me、类概率ßck和高斯宽度σ(标准差)是估计的参数;我们要求ßck为非负的，并且对于所有k， Σβck = 1。m和Bck通过共轭梯度算法优化，以最大化条件类似然，k和σ通过内部交叉验证选择(见章节5.3)。

给定费雪矩阵，我们接下来需要计算所有点对之间的全局距离。在大多数情况下，全局距离定义中的最小路径积分不能解析计算，我们使用基于图的近似。我们首先在所有已知数据点之间形成一个全连通图，其中每对点之间的路径近似为一条直线。对于这些直线路径，路径积分可以通过分段逼近来计算(详见Peltonen et al.， 2004);所有实验T = 10片)。然后我们可以使用图搜索(Floyd算法)来找到图中的最短路径，并使用最短路径距离作为学习度量距离。这个图近似需要O(n3)时间，其中n是数据点的数量;注意，这并不过分，因为在isomap等方法中需要类似的图计算。然而，在我们的实验中，直线路径产生了同样好的结果，所以我们简单地使用它们，这只需要O(n2)时间。因此，SNeRV作为一个整体只需要O(n2)的时间，就像NeRV。

5.2监督可视化的比较方法

对于每个要可视化的数据集，有监督可视化和无监督可视化的选择取决于分析人员;一般来说，监督嵌入将更好地保留类之间的差异，但以牺牲类内部细节为代价。在本节的实验中，我们集中比较监督方法的性能;我们将把SNeRV与最近的三种监督非线性嵌入方法进行比较。

多重关系嵌入(MRE;Memisevic和Hinton, 2005)被提出作为随机邻居嵌入的扩展(Hinton和Roweis, 2002)。MRE最小化嵌入邻域和几个不同输入邻域之间的不匹配的总和，由Kullback-Leibler散度测量:通常一个输入邻域来自输入空间坐标，其他来自标签等辅助信息。MRE能够使用未标记的数据;对于未标记的点，涉及基于标签的邻域的分歧被简单地排除在代价函数之外。

有色最大方差展开(Song et al.， 2008)是无监督最大方差展开的扩展。它根据基于交叉方差算子的Hilbert-Schmidt独立准则最大化嵌入坐标和标签之间的依赖关系。这将导致输出内核的受限优化。由于这些细节，该方法也被称为Hilbert-Schmidt独立准则的最大展开方法(MUHSIC);我们使用这个缩写。

监督异构体(s -异构体;Geng et al.， 2005)是无监督等线图的延伸。与无监督等线图的唯一区别是输入空间距离的新定义:粗略地说，不同类的点之间的距离将比同类点之间的距离增长得更快。实际嵌入的方式与无监督等线图相同(在4.1节中描述)。Li和Guo还介绍了等线图的其他有监督扩展（2006）和顾和徐（2007）。

参数嵌入(PE;Iwata et al.， 2007)用高斯混合模型表示嵌入数据，其中所有高斯在嵌入空间中具有相同的协方差，并试图通过最小化Kullback-Leibler散度之和来保留原始数据的拓扑结构。

邻域成分分析(NCA;Goldberger等人，2005年;(参见Kaski and Peltonen, 2003, Peltonen and Kaski, 2005)是一种线性和非参数降维方法，它学习马氏距离测量，使得在转换空间中，k-最近邻分类达到最大精度。

5.3监督实验的方法学

我们使用了四个具有类信息的基准数据集(第4.5节中描述的Letter、Phoneme、Landsat和TIMIT)来比较有监督的NeRV和第5.2节中描述的五种有监督的可视化方法，即多重关系嵌入(MRE)、彩色最大方差展开(MUHSIC)、有监督的等距映射(S-Isomap)、参数嵌入(PE)和邻域成分分析(NCA)。我们使用了标准的10次交叉验证设置:在每一次折叠中，我们保留一个子集用于测试，其余数据用于训练。对于每个数据集，我们使用SNeRV和比较方法来查找二维可视化。

原则上，对于无监督实验，我们可以像4.3节那样评估结果，即通过平均平滑精度和召回率;唯一的区别是使用监督学习指标进行评估。然而，与SNeRV不同的是，其他方法没有使用相同的监督度量来制定。为了对这些方法进行无偏倚的比较，我们采用了一种简单的间接评估:我们通过结果可视化的类预测精度来评估四种方法的性能。虽然这是一种间接的测量方法，但准确度是一种合理的选择，可以进行无偏性比较，并已在几篇有监督的降维论文中使用。更详细地说，我们在训练期间提供测试点位置，但不提供它们的标签;在这些方法计算出可视化结果后，我们通过在嵌入数据上运行k-最近邻分类器(k=5)对测试点进行分类，并评估这些方法的分类错误率。

我们使用标准的内部10倍验证策略来选择所有未被各自算法优化的参数:每个训练集被细分为10倍，其中9/10的数据用于学习，1/10用于验证;我们学习了不同参数的可视化值;然后选择为嵌入式验证点产生最佳分类精度的值，并用于计算整个训练数据的最终可视化。

我们使用λ=0.1和λ= 0.3运行了两个版本的SNeRV。缩放参数o是通过固定输入邻域的熵来设定的，如2.2节所述。在这里，我们指定相关邻居数量的粗略上限为0.5-n/K，其中n是数据点的数量，K是用于估计度量的混合成分的数量;这种选择大致意味着，对于分离良好的混合成分，每个数据点将平均考虑来自相同混合成分的一半数据点作为相关邻居。一个简化的验证足以满足高斯数K和宽度σ的要求:我们不需要运行嵌入步骤，而是选择在输入空间中为验证点提供最佳条件类似然的值。对于S-Isomap，我们使用验证集选择其参数a及其最近邻居的数量，并训练一个广义径向基函数网络来投射新的点，如Geng等人(2005)所建议的那样。对于MUHSIC，参数是正则化参数v，最近邻居的数量，以及Laplacian图中的特征向量的数量，我们使用线性插值来投影MUHSIC作者建议的新点。对于MRE，唯一的自由参数是它的邻域平滑度参数σMRE。对于PE，需要提供一个条件密度估计:我们使用了与SNeRV使用的相同的密度估计(见公式10)，以获得尽可能无偏的比较。邻域分量分析是一种非参数方法，因此不需要选择任何参数。

5.4监督可视化结果

图13显示了10次折叠的平均错误率以及标准偏差。最好的两种方法是SNeRV和PE，在所有数据集上都有很好的结果。在两个数据集(Letter和TIMIT)上，SNeRV显然是最好的;在其他两个数据集(Phoneme和Landsat)上，SNeRV与其余最好的方法(分别是S-Isomap和参数嵌入)一样好。MRE显然比其他方法更差，而MUHSIC和NCA结果取决于数据集:在Letter中，它们是仅次于MRE的第二和第三差的方法，而在其他数据集中，它们离最好的方法不远。

权衡参数λ的取值对SNeRV的性能影响不大;λ = 0.1和λ = 0.3都产生了良好的投影。

为了评估每个数据集上的最佳方法是否在统计上显著优于下一个最佳方法，我们对10个交叉验证折叠的性能进行了配对t检验(表3)。比较的两种最佳方法始终是SNeRV和参数嵌入，除了音素数据集上的两种最佳方法是SNeRV和S-Isomap。对于Letter和TIMIT数据集，SNeRV明显优于次优方法，而对于其他两个数据集，差异不显著。总之，所有显著差异都有利于SNeRV。

图14展示了字母识别数据集的可视化示例;投影结果显示了10个交叉验证折叠之一，包括训练点和测试点。虽然有一些重叠，但通常SNeRV显示不同的类集群—例如，字母“M”在图的顶部是一个分离良好的集群。

参数嵌入也设法分离了一些字母，如“A”和“I”，但在图的中心存在严重的类重叠。在S-Isomap中，我们看到有一些非常分离的类集群，如字母“W”和“N”，但有一个很大的区域靠近图右边缘中心的重叠类。这种重叠比SNeRV更糟糕，但仍然大致相当;相比之下，MUHSIC、MRE和NCA在这个数据集上表现不佳，导致大多数类严重重叠。

6. 结论与讨论

通过将信息可视化的非线性投影任务制定为信息检索任务，我们推导出了对可视化性能的严格激励措施，平均平滑精度和平均平滑召回率。我们证明了这些新措施是两种传统信息检索措施的扩展:平均平滑精度可以解释为平均精度的更复杂的扩展，假阳性的比例在邻居从可视化中检索到的。类似地，平均平滑召回是平均召回的延伸，即被检索的邻域引起的失误的比例。

我们引入了一种称为邻居检索可视化器(NeRV)的算法，它优化了总成本，可解释为平均平滑精度和平均平滑召回之间的权衡。根据需求调整λ，为1时得到SNE

我们证明了NeRV既可以用于无监督可视化，也可以用于监督可视化。对于无监督可视化，我们简单地使用固定的输入距离;对于监督可视化，我们学习了输入空间的监督距离度量，并将结果输入距离插入到NeRV算法中。在后一种情况下，关键思想是以一种不覆盖输入特征空间的方式使用监督(标记数据);对于输入空间，我们使用一种拓扑保留类判别度量，称为学习度量。

在无监督可视化中，对于我们尝试的六个数据集中的大多数，对于四对不同的测量，NeRV的表现优于其他方法，并且总体上是最好的方法。NeRV在无监督分类的比较中也表现良好。许多最好的流形提取方法表现得出奇地差，很可能是因为它们没有被设计成将维数降低到数据流形的固有维数以下。然而，在可视化中，我们通常别无选择，只能将数据的维数降低到2或3，即使它的内在维数更高。NeRV的设计目的是寻找一种映射，这种映射在定义明确的意义上是最适合于某种类型的可视化，而不管数据的内在维数。

在监督可视化中，监督版本的NeRV表现与最佳替代方法参数嵌入相同或更好;这表明插件学习指标可以很好地结合监督。

6.1讨论

NeRV使用概率分布对相关性进行建模，如果任何查询的相关性“总量”被标准化为固定的总和，那么这是有意义的。这样的标准化对于任何相关性度量都是可取的，因为对于任何查询(兴趣点)，检索到的邻居点的相关性应该取决于它相对于其他点的邻近性的接近性，而不是取决于它与查询点的绝对距离。(我们之前的方法，局部MDS，可以被认为是一种没有归一化的近似NeRV的尝试。)

NeRV中的Kullback-Leibler散度是测量两个概率分布之间差异的自然选择，但原则上也可以使用其他散度度量。近邻检索和概率相关模型的概念是NeRV的关键部分，而不是具体的发散度量。

我们的插件监督度量的概念原则上也可以与其他方法一起使用;其他基于距离矩阵的无监督嵌入算法也可以转换为监督版本，只需将学习度量距离插入到距离矩阵中即可。我们进行了Sammon映射的初始实验(Peltonen et al.， 2004);后来出现了类似的等距地图概念(Weng et al.， 2005)。然而，我们认为NeRV是嵌入步骤的一个特别有吸引力的选择，因为它具有信息检索解释，并且它在经验上表现良好。

NeRV和局部MDS算法的实现以及平均平滑精度-平均平滑召回措施可在http://www.cis.hut.fi/projects/上获得MI/软件/长发绺锁