通过线性投影可视化为信息检索

雅科·佩尔托宁

赫尔辛基科技大学信息与计算机科学系，
邮政信箱 5400，FI-02015 TKK，芬兰
jaakko.peltonen@tkk.fi

抽象的。我们应用最近的形式化作为信息检索的可视化到线性投影。我们介绍了一种优化信息检索任务线性投影的方法：仅根据输入样本的低维可视化坐标检索输入样本的邻居。简单的线性投影使该方法易于解释，而可视化任务则通过新颖的信息检索标准明确定义。该方法还有一个优点：它投影输入特征，但它保留的输入邻域可以与输入特征分开给出，例如通过样本相似性的外部数据。因此，可视化可以揭示数据特征和复杂数据相似性之间的关系。我们进一步将该方法扩展到基于内核的投影。

关键词：可视化，信息检索，线性投影

1 简介

线性投影广泛用于可视化高维数据。它们具有易于解释的优点：可视化中的每个轴都是原始数据特征的简单组合，而这些特征通常具有明确的含义。线性投影也可以快速应用于新数据。相比之下，如果映射的函数形式完全可用，则非线性投影可能很难解释。一些非线性方法还需要大量计算或映射近似来嵌入新点。基于内核的投影是线性和非线性投影之间的中间地带；它们的计算在内核空间中是线性的，它们的可解释性取决于所选的内核。

线性可视化中的关键问题是使用什么标准来找到投影。传统的答案包括在主成分分析 (PCA) 中保留最大方差；在独立成分分析中保持独立结构；保留距离和成对约束，如 [1] 中所述；或最大化类预测能力，如线性判别分析、信息判别分析 [2]、邻域成分分析 [3]、通过折叠类进行度量学习 [4] 等。

当线性投影用于可视化时，先前的标准是不够的，因为它们仅与可视化间接相关。一绝

首先形式化可视化的任务是什么，以及该任务的良好性能指标是什么。这个问题最近在 [5] 中得到了回答，其中可视化任务被形式化为*信息检索任务*，并且推导出优度度量，这是对*精确*和*记起*. 基于优度度量，可以形成优化准则，并直接*优化信息检索任务中可视化的优点*; 然而，到目前为止，这种方法仅用于非线性嵌入，其中直接优化输出坐标而无需任何参数映射 [5、6]。

我们介绍了一种用于线性和基于核的可视化的新方法，称为线性邻域检索可视化器 (LINNEA)：我们将可视化的形式化应用为信息检索任务，并优化此类检索的精度和召回率。一个有用的属性是被投影的输入要素和用于计算输入邻域的距离*可以单独给*：例如，特征可以是文本文档的单词出现向量，距离可以是引用图中文档的距离。在特殊情况下，LINNEA 与方法相关*随机邻居嵌入*[7] 和*通过折叠类进行度量学习*[4]，但更一般；它可用于无监督和有监督的可视化，并允许用户设置信息检索的精度和召回率之间的权衡。我们通过初步实验表明，LINNEA 对多个数据集产生了良好的可视化效果。

2 可视化作为信息检索

我们简要总结了 [5] 中介绍的新颖的可视化形式化。

任务是*邻居或邻近关系的可视化*在一个高维数据集中。对于一组输入点$X_{我} \in R_{d_0}$, 我 =1，。. .,, 可视化方法产生输出坐标是$_{我} \in R_d$，这应该揭示邻里关系。这被形式化为*信息检索任务*：对于任何数据点，可视化应该允许用户在原始高维数据中检索其相邻数据点。从低维可视化中完美检索通常是不可能的，检索会犯两种错误：不检索邻居减少*记起*的检索，并错误地检索非邻居减少*精确*.

应用信息检索概念*精确*和*记起*为了可视化，在 [5] 中，它们被概括为连续和概率测量，如下所示。对于每个点我，A*邻域概率分布*$p_{i,j}$在所有其他点j被定义为; 在 [5] 中，基于输入距离的指数衰减概率$d(X_{我}, X_j)$ 用来。在本文中，我们允许$d(X_{我}, X_j)$产生于点之间距离的任何定义我和j.这*从可视化中检索点*也是概率的：对于每个点我分布$q_{i,j}$被定义为告诉特定附近点的概率j从可视化中检索。这$q_{i,j}$定义类似于$p_{i,j}$,但使用欧氏距离 ||是$_{我}$−是$_j$||可视化坐标之间是$_{我}$. 这产生

$$p_{i,j} = \Sigma \frac{电子^{-d2个}(X_{我}, X_j)/2\sigma 2个_{我}}{\sum_{k6=我} 电子^{-d2个}(X_{我}, X_k)/2\sigma 2个_{我}}, \quad q_{i,j} = \Sigma \frac{电子^{-||是_{我}-y_j||2个/2\sigma 2个_{我}}}{\sum_{k6=我} 电子^{-||是_{我}-y_k||2个/2\sigma 2个_{我}}} \tag{1}$$

其中σ我是尺度参数，可以通过固定熵来设置pi,j
正如 [5] 中所建议的那样。自从pi,j和qi,j是概率分布，很自然地使用 Kullback-Leibler 散度来衡量检索到的分布有多好
butions 对应于输入邻域。分歧丁吉隆坡(p我, q我) =
Σ j6=我pi,j日志（pi,j/qi,j) 结果是一个*回忆的泛化*和丁吉隆坡(q我, 磷我) 结果是一个*精度的泛化*. 分歧的价值是对点的平均我这产生了最终的善意措施。

3 方法：线性邻域检索可视化工具

上面对precision和recall的概括可以直接作为优化目标，但是由于precision和recall通常不能一起最大化，用户必须在它们之间设置一个tradeoff。考虑到权衡，可以定义单个成本函数，并且可以根据成本函数直接优化可视化。在早期的作品 [5, 6] 中，这种方法用于计算非线性嵌入，即输出坐标是我直接优化了数据点。在本文中，我们改为考虑参数化线性投影是我=体重我在哪里W∈Rd×d0是投影矩阵。我们希望优化W使得投影有利于可视化的信息检索任务。我们称该方法为线性邻域检索可视化工具 (LINNEA)。我们使用与 [5] 中相同的成本函数，即

$$E = \lambda \sum_{\text{我}} \text{丁}_{\text{吉隆坡}}(p_{\text{我}}, q_{\text{我}}) + (1-\lambda) \sum_{\text{我}} \text{丁}_{\text{吉隆坡}}(q_{\text{我}}, \text{磷}_{\text{我}})$$

$$= \sum_{\text{我}} \sum_{j6=\text{我}} \left[-\lambda p_{i,j} \text{日志} q_{i,j} + (1-\lambda) q_{i,j} \text{日志} \frac{q_{i,j}}{p_{i,j}} \right] + \text{常量}。 \quad (2)$$

其中权衡参数 λ 由用户设置，以反映准确率或召回率哪个更重要。我们简单地使用共轭梯度算法来最小化乙关于矩阵W。梯度∂E

$$\frac{\partial}{\partial \text{W}} \text{是} \sum_{i,j6=\text{我}} \left[\lambda(p_{i,j}-q_{i,j}) + (1-\lambda) q_{i,j} \text{丁}_{\text{吉隆坡}}(q_{\text{我}}, \text{磷}_{\text{我}}) - \text{日志} \frac{q_{i,j}}{p_{i,j}} \right] (\text{是}_{\text{我}} - \text{是}_{\text{j}})(\text{X}_{\text{我}} - \text{X}_{\text{j}}) \text{吨} \quad (3)$$

哪个产量在σ2我个) 每个梯度步骤的计算复杂度。

*优化细节。*在本文中，我们简单地初始化元素W在 0 和 1 之间统一随机数；更复杂的初始化，比如通过初始化W作为主成分分析投影，自然是可以的。为了避免局部最优，我们使用两种简单的方法。首先，在每次运行中，我们首先将邻域尺度设置为较大的值，并在每个优化步骤后减小它们直到达到最终尺度，之后我们使用最终尺度运行 40 个共轭梯度步骤。其次，我们从 10 次随机初始化开始运行算法，并取具有最佳成本函数值的结果。

4个　　雅科·佩尔托宁

3.1 内核版本

我们现在展示 LINNEA 的内核版本。我们优化内核空间的投影而不是简单的线性投影：我们设置是我=WΦ(x我) 其中 Φ(・)是对具有由核函数给出的内积的潜在无限维空间的某种非线性变换k(X我,Xj) = Φ(X我)吨Φ(Xj).像往常一样，内核就是我们所需要的，不需要知道 Φ。

任务与之前相同：优化投影（可视化），以便根据成本函数（2）有利于信息检索。

可以合理地假设行w吨　　Σ升的W可以表示为的线性组合 Φ(X我）， 以便w升= 系数。然后投影具有简米Φ(X米） 在哪里A升 米是单的形式

$$是_{我}=\Big[\sum_{米} A_{米}\Phi(X_{米}),\ldots,\sum_{米} A_{米}\Phi(X_{米})\Big]_{吨}\Phi(X_{我}) = AK(x_{我}) \tag{4}$$

其中矩阵A∈R$_{d\times 否}$包含系数A(l, m) = 一个升　　米和K(x我) = [k(X1个,X我),. . . , k(X否,X我)]吨.和以前一样，坐标是我可用于计算邻域q$_{i,j}$,成本函数等。

为了优化这个基于内核的投影，优化关于系数矩阵的成本函数就足够了A。我们可以再次使用标准的共轭梯度法：关于A与等式（3）相同，除了X我和Xj被替换为K(x我） 和K(xj). 自从A有 否列，计算复杂度变为在3个) 每个梯度步骤。

3.2 LINNEA 的特性

LINNEA 的一个关键属性是输入特征X我被投影和距离d(X我,Xj) 用于计算输入邻域的可以单独给出。最简单的d(X我,Xj) 可以是欧式距离 ||X我−Xj||,但它也可以基于其他数据：例如，X我可以是文本文档的单词出现向量，并且d(X我,Xj)可以是引用图中文档的距离（我们在第 4 节中测试了这个例子）。当直接从输入特征计算距离时，投影是无监督的。当距离单独给出时，投影由距离监督；然后优化投影以允许检索基于单独给定距离的邻居，因此它揭示了特征与距离之间的关系。

请注意，仅基于距离的可视化，例如根据文档之间的引用图距离计算的多维缩放，不会提供可视化与特征（文档内容）之间的任何关系；相比之下，LINNEA 可视化直接是特征的投影，它针对基于距离的邻居检索进行了优化。

如果我们在 (2) 中设置 λ = 1，即我们最大化召回率，这将产生成本函数*随机邻居嵌入*(神经元；[7]); 因此 LINNEA 包括 SNE 的线性版本作为特例。更一般地说，LINNEA 的成本函数实现了精确度和召回率之间灵活的用户定义权衡。

如果输入邻域来自数据点的类标签，则会出现另一个有趣的特殊情况。考虑一个简单的社区$p_{i,j}$: 对于任何一点我在班上$C_{我}$，邻居是来自同一类的其他点，具有相等的概率。很容易证明，如果我们在成本函数中设置 λ = 1（即我们最大化召回率），这相当于最大化

$$\sum_{我}\sum_{j6=我}\delta_{C_{我},C_j}\text{日志}\frac{\text{电子}^{-||\text{是}_{我}-y_j||^2/2\sigma^2_{我}}}{\sum_{j\ 6=我}\text{电子}^{-||\text{是}_{我}-y_j\ ||^2/2\sigma^2_{我}}} \tag{5}$$

其中$\delta_{C_{我},C_j}$=1 如果类 ($C_{我}$， C_j) 是相同的，否则为零，并且为简单起见，假设类是等概率的。这是成本函数*通过折叠类进行度量学习*(MCML；[4]）作为 SNE 的监督线性版本引入。LINNEA 将 MCML 作为特例包括在内。因此我们给出了 MCML 的新解释：它最大化*召回同级点*.

请注意，LINNEA 为上述类型的直接输入邻域产生了有意义的解决方案，因为映射是输入特征的线性投影。相比之下，自由优化输出坐标的方法可以产生将每个类的所有输入点映射到单个输出点的简单解决方案。为了避免琐碎的解决方案，这些方法可以例如应用 [6] 中的保留拓扑的监督指标；LINNEA 不需要如此复杂的指标。

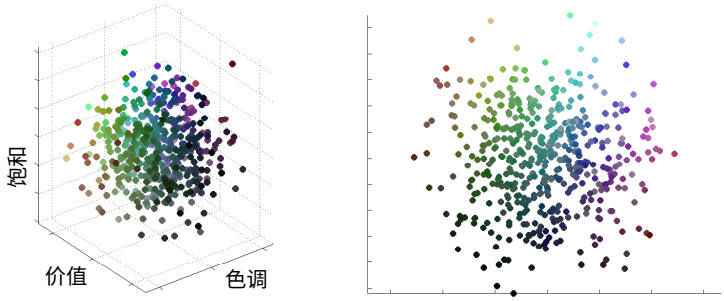
总之，LINNEA 可用于有监督和无监督的可视化；它与众所周知的方法有关，但更通用，允许用户在信息检索的不同成本之间进行权衡。

4 实验

在第一篇论文中，我们还没有对 LINNEA 和早期方法进行彻底比较。我们通过四个实验展示了 LINNEA 的潜力；我们使用主成分分析 (PCA) 作为基线。我们使用 LINNEA 的非内核版本（第 3 节），并在实验 1-3 和 λ =0.1在实验4中。其他参数是[5]代码中的默认值。

*实验一：提取相关维度。*我们首先在玩具数据上测试 LINNEA，其中可视化可以完美地恢复给定的输入邻域。考虑在色相饱和度值 (HSV) 颜色空间中包含 500 个点的球形高斯云，如图 1（左）所示。人们无法在一个二维可视化中表示所有三个维度，并且在没有额外知识的情况下，人们无法分辨哪些特征需要保留，因为数据的形状在所有方向上都是相同的。但是，如果我们还给出了点之间的成对距离，它们将决定要保留哪些特征。假设这些距离仅根据 Hue 和 Value 秘密计算；那么正确的可视化是采用这两个维度，忽略饱和度。

我们用 LINNEA 优化了一个二维投影；我们将每个数据点的 HSV 分量作为输入特征，并根据已知的成对距离计算输入邻域。如图1（右）所示，LINNEA



图。1。色相-饱和度-值色彩空间中点的投影。左边：原始三维数据集为高斯点云；坐标对应于每个点的 Hue、Saturation（灰度-色彩度）和 Value（明度-暗度）。成对输入距离仅根据 Hue 和 Value 计算得出。正确的：LINNEA 已正确找到投影中的 Hue 和 Value 维度并忽略了 Saturation。

根据需要找到色相值维度并忽略饱和度；Saturation 在投影方向上的权重接近于零。

*实验 2：S 曲线。*我们可视化具有简单底层流形的数据集：沿二维流形采样 1000 个点，嵌入三维空间中作为 S 形曲线，如图 2（左）所示。没有给出外部成对距离，输入邻域是根据三维输入特征计算的。任务是找到一个可视化，可以很好地检索流形上的原始邻居。展开流形就足够了；然而，线性投影无法完美展开非线性 S 曲线。图 2（中）中的 PCA 解决方案遗漏了原始 Z 轴，这对于检索原始邻居而言不是最优的，因为它仅留下底层二维流形的一个坐标可见。图 2（右）中的 LINNEA 结果强调了 Z 和 Y 方向；

*实验三：人脸图像的投影。*我们将人脸数据集可视化（[8]；可在http://web.mit.edu/cocosci/isomap/datasets.html）。该数据集有 698 张不同方向和光照方向的合成人脸图像；每个图像有 64×64 像素。我们首先使用像素图像作为输入特征找到面部图像的线性投影，而不提供任何额外的知识。如图 3（左上）所示，PCA 揭示了部分数据结构，但检索相邻面孔的结果并不令人满意，因为 PCA 已将背光面聚集在一起。相比之下，如图 3（右上）所示，LINNEA 展开了前光和背光面。投影方向可以解释为图像的线性滤波器。对于 PCA，水平轴上的滤波器大致响应左向头部；过滤器

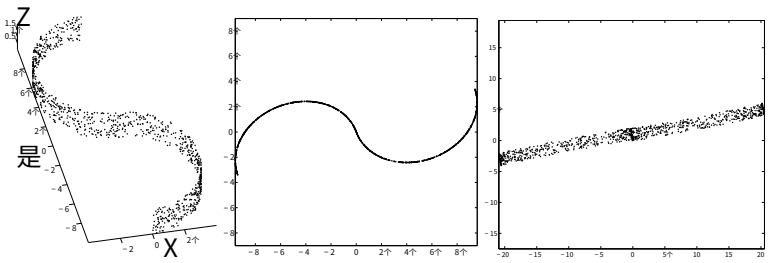


图 2。S 曲线的投影。左边：原始三维数据。中间： PCA 忽略原来的 Z 方向。正确的：LINNEA 找到一个投影，可以从可视化中很好地检索底层流形上的邻居。

垂直轴上的 粗略检测左右照明方向。LINNEA 过滤器很复杂；在未来的工作中需要对过滤器进行更多分析。

*检索已知姿势/照明邻居的投影。*对于面部数据，可以使用面部的姿势和光照参数。然后，我们可以根据这些参数计算成对输入距离，并使用 LINNEA 找到像素图像的监督可视化，最能检索每张脸的姿势/光照邻居。LINNEA 投影如图 3（底部）所示。面部图像在姿势和光照方面都安排得很好；左上-右下轴粗略地分隔左面和右面的前照面，右上-左下轴粗略地分隔左面和右面的背光面。相应的过滤器有些复杂；垂直轴和水平轴上的过滤器似乎分别粗略地检测边缘和照明方向。底层的姿势/光照空间是三维的，不能用二维映射精确表示，因此过滤器是表示姿势/光照的几个方面之间的折衷。请注意，在已知姿势/光照参数上运行例如 PCA 不会产生像素图像的过滤器，因此它不会告诉像素数据如何与姿势/光照相关；相比之下，LINNEA 优化了用于检索姿势/光照邻居的过滤器。

*实验四：科学文献的可视化。*我们将包含科学文章及其引用的 CiteSeer 数据集可视化。数据集可在http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html。 每篇文章都用一个二进制的 3703 维向量来描述文章中出现了哪些词；我们使用这些向量作为输入特征。为了减少计算负荷，我们选取 了 1000 篇文章的子集，这些文章的入站引用和出站引用次数最多。我们提供单独的成对输入距离，简单地取引文图中的图距离：即，一个引用另一个的两个文档的距离为 1，引用同一其他文档的文档的距离为 2，依此类推。作为简化，我们假设引文是对称关系，作为正则化，我们将图形距离上限设为 10。我们使用 LINNEA（$\lambda = 0.1$）优化二维可视化，其中根据该图形距离的邻居可以

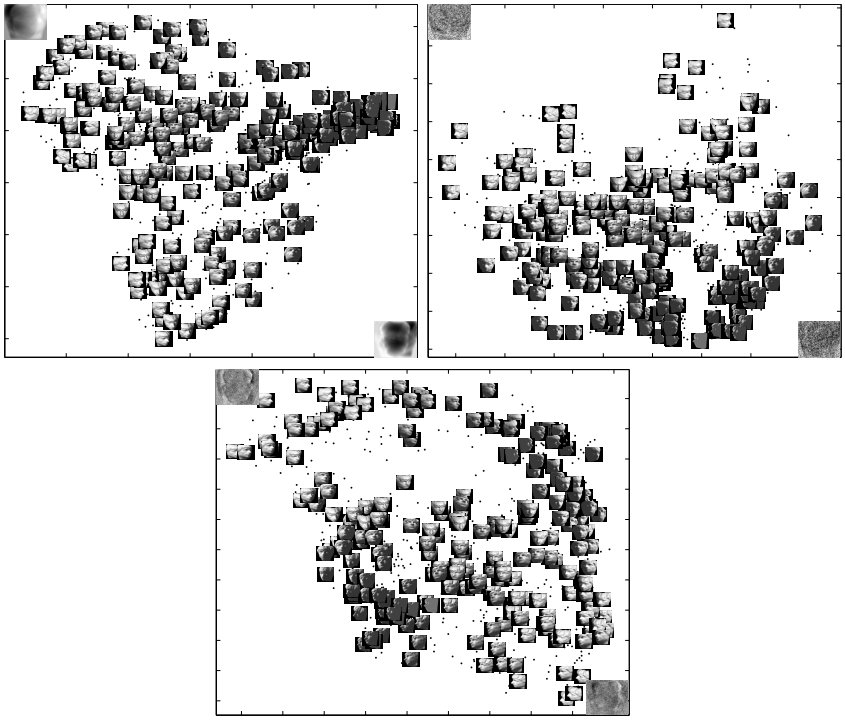


图 3。面部图像的投影。顶部：PCA（左）和 LINNEA（右）对像素图像的无监督投影。线性投影方向可以解释为图像的线性滤波器，它针对每个轴显示。底部：LINNEA 的监督投影。成对距离来自已知的面部姿势/光照参数。LINNEA 优化了像素图像的投影，用于检索具有相似姿势/光照参数的邻居。更多分析见文。

最好找回。结果如图 4 所示。在基线 PCA 投影（左子图）中，引用分布在几乎没有可见结构的数据中，而 LINNEA 投影（右子图）显示出清晰的结构：文档相互引用的集群，以及集群之间的引用连接。对于此数据，每个特征都是一个词；不幸的是，这些词的身份无法获得。一般来说，可以通过为每个方向列出具有最大权重的单词来解释 LINNEA 给出的投影方向。

5。结论

我们介绍了一种通过线性或基于内核的投影进行可视化的新方法。该投影针对从可视化中检索原始相邻点的信息进行了优化，并在精度和召回率之间进行了用户定义的权衡。该方法既可以找到输入特征的投影，也可以找到

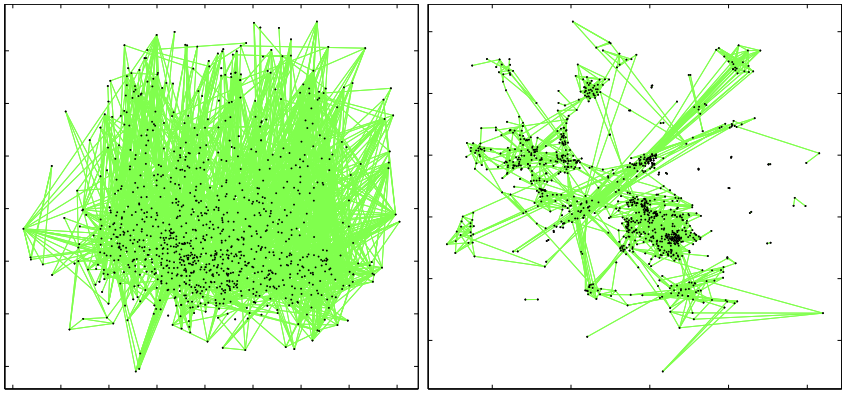


图 4。科学文献的预测。文档显示为点，两个文档之间的引用显示为线。左边：文档内容向量的 PCA 投影不能很好地揭示引用邻域。正确的：LINNEA 的投影显示了引证文献的聚类和聚类之间的联系。

揭示输入特征与单独给定的输入距离之间关系的投影。该方法产生了几个数据集的良好可视化。

致谢。作者属于自适应信息学研究中心和赫尔辛基信息技术研究所 HIIT。他得到了芬兰科学院的支持，第 123983 号决定，以及 PASCAL2 卓越网络的部分支持。他感谢 Samuel Kaski 进行了非常有益的讨论。

参考

1. Cevikalp, H., Verbeek, J., Jurie, F., Kläser, A.：使用成对等价约束的半监督降维。在：过程。VISAPP 2008，第 489–496 页。（2008 年）

2. Peltonen, J., Kaski, S.：数据的判别成分。IEEE 跨。神经网络16(1), 68–83 (2005)

3. Goldberger, J.、Roweis, S.、Hinton, G.、Salakhutdinov, R.：邻域成分分析。在：过程。NIPS 2004，第 513-520 页。麻省理工学院出版社，马萨诸塞州剑桥 (2005)

4. Globerson, A., Roweis, S.：通过折叠类进行度量学习。在：过程。NIPS 2005，第 451–458 页。麻省理工学院出版社，马萨诸塞州剑桥市 (2006)
5. Venna, J., Kaski, S.：作为信息检索的非线性降维。在：过程。人工智能统计*07。(2007)

6. Peltonen, J., Aidos, H., Kaski, S.：通过近邻检索监督非线性降维。在：过程。ICASSP 2009。出版中。
7. Hinton, G., Roweis, ST：随机邻域嵌入。在：过程。NIPS 2002，第 833–840 页。麻省理工学院出版社，马萨诸塞州剑桥 (2002)
8. Tenenbaum, JB, de Silva, V., Langford, JC：非线性降维的全局几何框架。科学290 (2000 年 12 月)