山东大学 计算机科学与技术 学院

可视化技术 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：201900130176 | 姓名： 李伟国 | | 班级： 智能 |
| 实验题目： | | | |
| 实验学时： | | 实验日期： 2021/10/16 | |
| 实验目的：  对mnist数据分析不同的参数对t-sne结果的影响（60k 的数据量，只是用其中的train set就可以） | | | |
| 硬件环境：  处理器：AMD Ryzen 5 3600 6-Core Processor 3.60 GHz  Ram 16.0 GB | | | |
| 软件环境： | | | |
| 实验步骤与内容：  数据的基本信息  测试的数据时MNIST 是手写识别的图片的集合，每个图片是28\*28的大小。并且每个像素是单通道的灰度图。  将这28\*28的每张图片strench 成一个个的列向量，因为是单通道的图片，所以每张图片的维度都是28\*28 = 784，即维度是784维度的，现在采用t-sne方法进行降维。  由于数据较多，为了达到测试的目的，这里仅仅使用了6000个样本，随机抽选6000个样本，做为本次实验的数据集。  下面是matlab自带的T-sne 函数对该数据集降维后的效果    默认参数的情况（**peplexity** = 30）  通过对上面的降维后的结果可视化后，可以非常容易的发现，几乎每个数字的类别都被分隔的非常好。仿佛达到了一种 聚类的效果，也就是说，在高维空降的聚类在低维度的空间中任然被保留了。但是peplexity 只有30 ，并不能对拥有6000个数据点的集合有较好的global structure。所以下来来调整peplexity 的值。  将perplexity 调整为50    Perplexity = 50      Perplexity = 100    Perplexity = 200    Perplexity = 500  通过调整不同的perplexity ，我们可以发现，在perplexity 很大的时候，在embedding 中其cluster 的 效果远远不如在默认情况下即perplexity = 30的情况下的效果，这是因为，perplexity 越大，所要在低维空间中保持的点与点之间的距离的数目就越多，也就是说要保持原来高维的数据越多，导致降维后的效果会越复杂，就会导致类与类之间的空间会很小，会更加的紧凑。但是，如果perplexity 设置的很小，类与类之间就不能很好地分离开来  Exaggeration parameter：这个参数增加在连个点之间的吸引力并且允许点移动的更加的自由。默认的情况下，Exaggeration = 4（在matlab中是这样），该参数用来指定在数据集上的natrual cluster。 Exaggeration 设置的越大会让tsne学习到跟多的联合概率分布，那么tsne 也就更能将不同类之间跟清晰的分开。Tsne在前99次迭代中使用这个参数.  由于Mnist 训练集上有10 种类型不同的cluster，所以不妨将exaggeration 设置为10，perplexity 任然设置为30.    Exaggration = 2    Exaggeration = 10    Exaggration = 12    Exaggeration = 15    Exaggeration = 20  通过对上面不同的exaggeration 的设置，可以发现，其在embedding space 中的位置在改变，该参数设置的越大，那么在embedding space中各个类之间的空隙也就越大，在将exaggeration 设置为2的时候，可以发现,数字 7这个 clusters被分成了2 块.而随着exaggeration 的增大，这种情况有被缓解。  深层原理  1）  t-sne 使用高斯分布来衡量两个点的相似性，以要测量的点在为中心（该点称为point of interest），计算每一个点与该点的距离，并且在高斯分布的概率密度函数上计算出每个点相对于该点的值，直觉告诉我们，越近的点，它们彼此之间就有又更大的相似性。  similarity：pj∣i=，（以i点为interest，衡量j点的相似性），这个式子表示的是经过归一化后的结果，分子相当于是标准的高斯分布的函数值，输入是两个点之间的欧式距离。由于对J点作为interest ，i点对其的similarity并一样（因为，对i点和j点来说，它们个周围点密度并不是同等程度的，这体现在标准差上），所以用同样的方式计算P i|j , 通过将二者平均一下，作为j点和i点彼此的similarity，即：pij= 。  在embedding space 中，使用t-distrubted，该分布函数具有中心点对应的值比Gaussian要低，但是尾巴比Gaussian 要高一点。可以在embedding space中计算出两个点的相似度Pij  不妨分别在original space 和 embedding space中根据计算出啦的相似性构造matrix Q 和 P，t-sne的目的就是让Q matrix 和 P matrix 很相似。  通过使用KL 做为cost function ，然后用gradinet descent 的方法来优化损失函数。    还有一点需要做的就是以每个data point 为中心的Gaussian function 的σ应该怎么设置。这里t-sne 引入了 perplexity 这个hyperparameter，perplexity 可以被认为是对最近的k个邻居的连续模拟。σ被设置成可以包含这k个邻居的值。这k个邻居的值可以认为的设定，是试出来的，有一个最适的取值。即perplexity 就是在local 和 global的一种tradeoff（权衡），会在embedding space空间中保留这k个neighboor的距离。如果要考虑的邻居越多，那么Gaussian function的分布越广才行，即标准差就越大，and vice the visa。 Perplexity 可以认为是一个点附近的有效的近邻点的个数，tsne对perplexity 的调整比较有鲁棒性，通常选择5-50之间。给定之后用二分搜索寻找合适的σ。  在原始的空间中使用Gaussian 分布，在 embedding 中使用t – distributed 。  2） tsne 的不足之处  Crowing 问题（tsne 在embedding space 中使用t 分布来避免crowding问题）  拥挤问题就是说各个簇聚集在一起，无法区分。比如有一种情况，高维度数据在降维到10维下，可以有很好的表达，但是降维到两维后无法得到可信映射，比如降维如10维中有11个点之间两两等距离的，在二维下就无法得到可信的映射结果(最多3个点)。  假设数据点在高维度空间中时均匀分布的，intuitively， 那么以第i个点为中心的附近的点，在离其越远的地方，点的数量会越多，与到i点的距离分布及其不均衡，如果直接将这种距离关系保持到低维，就会出现拥挤的问题  加速:四叉树的加速 Barnes-Hut  3）注意事项    Tsne 具有随机性：每次的实验结果都可能不一样高  Intrisic：如果数据经过t-sne后再2D平面上的效果不好，可能并不是算法不好，而是数据本身的内在结构不足以在2D平面上表示 | | | |
| 结论分析与体会： | | | |

附录：程序源代码