A neural algorithm for fundamental computing programming

求解基本计算问题的神经算法

【作者】

Sanjoy Dasgupta

Charles F.Stevens

Saket Navalakha

【摘要】

相似性搜索，例如，在数据库中识别相似的图像，或者在网页上识别相似的文章，是大规模信息检索系统所面临的一个基本的计算问题。我们发现果蝇的嗅觉回路通过一种利用了计算机的科学算法（局部敏感哈希）的新型变体解决了这个问题。果蝇的嗅觉回路为相似的气味提供相似的神经元的活动模式，所以果蝇因一种味道产生的活动模式可以应用到其它相似的味道上。然而果蝇使用了三种偏离传统方法的计算策略。这些算法能够被用来提高相似性搜索的性能。这一观点能够帮助阐明支持重要的感觉功能的逻辑，并且它为解决基本的计算问题提供了一种概念上的新算法。

许多神经回路的基本任务就是对输入的刺激生成神经活动模式，这样不同的输入刺激就能被明确地识别出来。我们研究了果蝇嗅觉系统用来处理气味的回路，并且由此发现了解决基本机器学习问题的计算策略:近似相似(或近似邻域)搜索。

果蝇的嗅觉回路会对每一种气味赋予一个由一组神经元组成的标签，每当这种气味出现时，对应标签（即一组神经元）就会被激活。这个标签对于研究果蝇对不同气味的行为反应至关重要。举个例子：如果一种奖励（例如糖水）或者惩罚（例如电击）与一种气味相联系，那么这个气味就会有吸引力（即果蝇将会靠近这个气味）或者有排斥性（即果蝇将会远离这个气味）。被分配给气味的标签是稀疏的——只有一小部分接收气味信息的神经元对其作出反应(3-5)，而且并不重叠：如果两种随机选择的气味对应的标签有相重叠的部分，也不会是活跃的神经元，所以不同的气味能够被轻易地区分。

气味标签的计算方法是三步法(图 1A)。第一步包括一个前馈连接，从果蝇鼻子里的气味神经受体（ORNs）到肾小球中的投射神经元（PNs）。有 50 种不同的气味受体神经元，而且每一种都对不同的气味有不同的敏感度和选择性。。所以每一种输入的气味在由 50 个气味受体神经元组成的 50 维的空间中都有特定的位置。对于每一种气味对应的分布在 50 种受体神经元中的神经元的激活率是呈指数分布，而且其平均值取决于气味的浓度。对于投射神经元来说，这个浓度的影响消失了，也就是说，50 种投射神经元中的激活率呈现指数分布，而且其平均值对于所有的气体以及气体浓度是相同的。因此，三步法的第一步的本质实际上是——一种标准的预处理步骤——名叫“除法归一法”的方法。这一步非常重要，经过这一步后，果蝇就不会将气味的强度和气味的类型混合在一起。

第二步，算法的主要部分开始了，包括神经元数量 40 倍的拓展：由 50 种 PNs 到 2000 种凯尼恩细胞（KCs）由一个稀疏的二进制的随机矩阵连接。每一个凯尼恩细胞都接收大约 6 个随机选择的 PNs，并且总结其激活率。第三步是使用一种能够单一抑制的 APL（前对侧神经元）神经元对反馈为“win”（赢）的回路进行强烈抑制。结果是，激活率最高的 5% 的凯尼恩细胞都不活跃了（1,3,4），剩余的 5%分给了输入气味对应的标签。

从计算机科学的角度来看，我们认为这个在果蝇中的回路就像是一种输入为气味，输出为对应的标签（被称作哈希）。尽管标签能够区分气味，但果蝇的优势是能够用将相似的气味同相似的标签联系起来（Fig.1B）所以我们从一种气味中学习到的条件反射，可以被当成一种经验应用到相似的气味上，或者是已经学习的气味的嘈杂版本。这使得我们可以进行这样的推测：果蝇的回路产生的标签是局部敏感的，而且当气味越相似（就是 50 种 ORN 的激活率），它们被分配的标签越相似。局部敏感哈希[LSH (10, 11)]是解决计算机科学中众多相似搜索问题的基础。我们从果蝇的回路中得到一些灵感，可以利用其开发一类 LSH 算法以有效地找到高维度点的邻近区域。

假设你被提供了一张大象的图片，并且尝试从网上的十亿多张的图像中找到与之最相近的 100 张。这被称为近领域搜索问题，它在信息检索，数据压缩和机器学习中具有重要的意义。每一个图像通常表示一个 D 维上的特征向量（果蝇进行操作的每一种气味都是一个 50 维的激活率的特征向量）。使用距离度量来计算两个图像（特征向量）之间的相似度，目的是为了有效的找到任何查询图像的最近邻。如果网络上只包含少量的图像，那么只需要进行蛮力的线性搜索就可以很容易地找到最近邻。如果网络上包含许多土星，但是每一个图像都有一个低维的向量表示（例如 10 个或 20 个特征），那么空间区分方法（12）也就足够了。然而，对于具有高维数据的大型数据库，那么两种方法都不适用。（11）

在许多应用程序中，只要能够很快的找到它们，就可以返回一个接近于查询的足够近的近似邻。这激发了一种寻找近似邻的方法 LSH（10）。正如所指出的那样，对于苍蝇来说，环境敏感的特性说明了两种相似的气味将会被两种相似标签的包含的 ORN 所响应（Fig.1B）。同样的，对于图像搜索而言，大象图像的标签将与另一个大象图像的标签类似，而不会像一个摩天大楼图像的标签。

不像传统的（非 LSH）哈希函数，输入点是随机分布的，并且在范围均匀分布，LSH 函数提供了一个从 D 维空间到 M 维空间（后者对应标签）的保持距离的嵌入点。因此，在输入空间中彼此更接近的点被分配相同或相似的标记的概率要比彼此相隔远的点高。【正式定义在（13）】

为了设计一个 LSH 函数，一个常见的技巧是计算输入数据（10,11）的随机投影，——即将输入的特征向量乘以一个随机矩阵。 Johnson-Lindenstrauss 引理（14,15）和它的许多变体，为我们使用各种类型的随机投影（16-18）将数据从 D 维嵌入 M 维时，并且能够在很大程度保持位置的相对结构提供了强大的理论边界。

果蝇还通过随机投影（Fig.1A 中的 step2；50 PNs →2000 KCs）为气味分配标签，这为这部分的回路提供了一个关键的线索。然而，果蝇的算法和传统的 LSH 算法有三个不同之处。第一点，果蝇使用稀疏的二元随机投影，而 LSH 函数通常使用稠密的需要更多的数学运算来计算的高斯随机投影；第二点，果蝇拓宽了投影后输入的维度（D《M），而 LSH 则降低了维度（D》M）；第三点，果蝇用 WTA 机制来稀疏化高纬度的表示，而 LSH 则保留了稠密的表示。

在补充材料（13）中，我们分析了果蝇在嗅觉器官回路中的稀疏的，二元随机投影产生标签，从而保持了输入点的领域结构。这证明了果蝇的回路代表了一个不为人知的 LSH 家族。

然后，我们根据每个算法如何识别给定查询点的最近邻，对果蝇算法与传统的 LSH

（10,11）进行经验评估。为了进行公平的比较，我们将两种算法的的计算复杂度改成相同的（图 1C）。也就是说，对于每一个输入，这两种方法固定的使用相同数量的数学运算来生成长度为 K 的哈希（即一个带有 K 个非零值的向量）

我们比较了这两种算法在三个基准数据集中寻找最近邻的算法：SIFT (d = 128), GLOVE (d = 300), and MNIST (d =784) (13)；SIFT 和 MNIST 都包含了用于图像相似搜索的图像的矢量表示，而 GLOVE 则包含用于语义相似搜索的词的向量表示。我们使用了每个数据集的子集，每个数据集有 10000 个输入，其中每个输入被表示为 D 维空间中的一个特征向量。为了测试性能，我们从 10000 个随机查询输入中选择了 1000 个，并将真实值的和预期的最近领域进行了比较。也就是说，对于每个查询，我们在输入空间中找到了最接近的 2%的最近邻（200），这是根据特征向量之间的欧氏距离确定的。然后我们发现在 M 维空间中最近邻的预测值的前 2%，基于欧式距离标记（哈希）的距离确定。我们改变了哈希（K）的长度，并且利用中间平均精度（19）计算出了真实的排序列表和最近邻之间的重叠.我们改变随机投影矩阵和查询输入进行了 50 次不同的实验,并且平均了中间的平均精度.我们分离了果蝇算法和 LSH 算法之间的三个差异,用来测试它们对近邻检索性能的影响.

我们使用稀疏的二元投影代替 LSH 的稠密高斯投影,而不损害最近邻的精确位置.(图 2A) 计算局部敏感的果蝇的随机投影的这些结果支持我们的理论.此外,稀疏的二进制随机投影实现了相对于稠密的高斯投影(图 S1)(13)20 个相关联的计算因子的节省。

当扩展维度时，使用 WTA 对标签进行稀疏化，结果会比使用随机选择更加有效。（图 2B）WTA 会选择前 K 个激活性能较好的凯尼恩细胞作为标签，不像随机标记选择，会随机选择 K 个凯尼恩细胞。对于这两种情况，我们都会使用 20K 的随机投影，这样能够将果蝇算法和 LSH 所使用的而数学运算的数量等同起来（13）。例如，对于筛选 hash 长度 K=4 的 STF 数据集上，随机选择产生了 17.7%的平均精度，WTA 使用的平均值大约是其两倍（32.4%）。因此，选择激活性能靠前的神经元最好地保持了输入之间的相对距离；增加的维度也使得区分不同的输入变得更容易。对于随机的标签选择，我们随机选择了 7 个（但所有的输入都是固定的）KCs 作为标签。因此，它的性能与仅仅只是在 LSH 中做个 K 的随机投影是完全相同的。

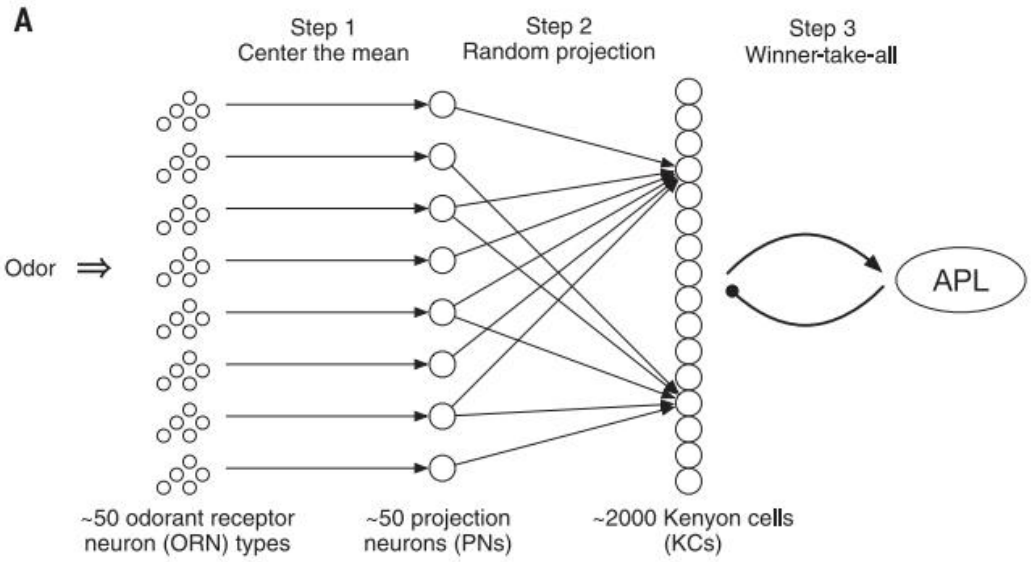
随着维度的进一步扩展（从 20K 到 10dK KCs，更加接近果蝇的实际回路），我们获得了与 LSH 相比，根据所有数据集合，哈希长度（图 3），识别最邻近值的显著增量。在非常短的哈希长度中获得最高的增益，并将平均值的平均精度提高了将近 3 倍。（例如，对于 MNIST 来说，k = 4 时，LSH 是 16%，而对于果蝇算法则是 44.8%）。

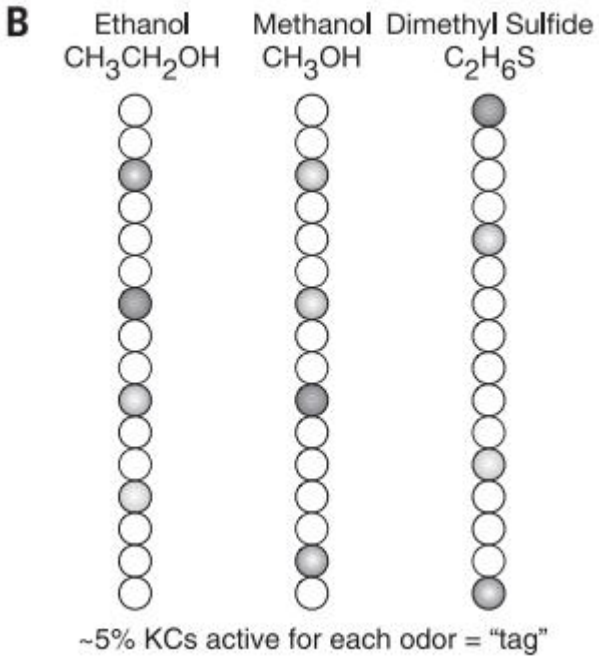
在更高维度的果蝇算法，和二进制 LSH（20）（图 S2--图 S3），我们同样发现了类似性能的提升。这些表明，果蝇算法是可扩展的，并且在其它的 LSH 家庭中也适用。

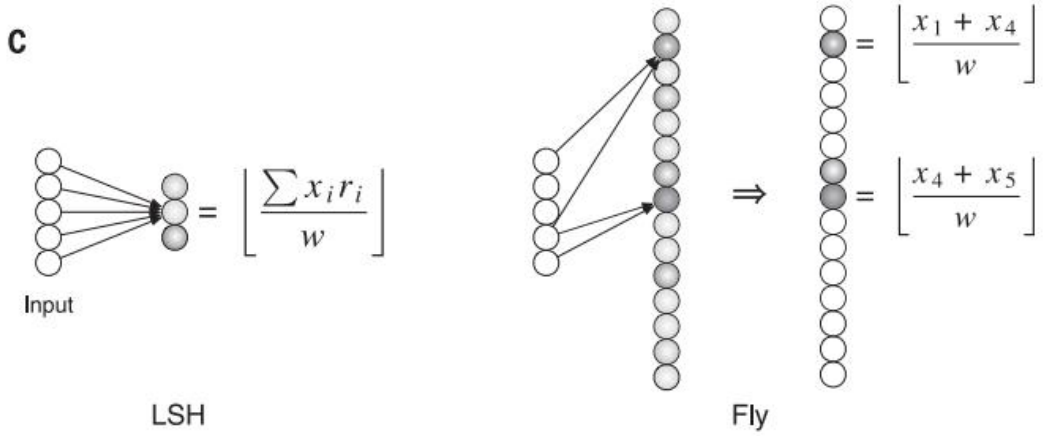
我们的研究表现，在大规模的信息检索系统中，大脑中的相似性匹配策略（21）和针对最近邻搜索的哈希算法存在协同作用。它们也可以被应用在重复检测，聚类和高能深度学习（22）等方面。LSH（23）有许多扩展，包括使用多个哈希表（11）来提高精度（过去两种算法都使用一个哈希表），多功能探针（24）的使用，可以将类似的标签组合在一起（这可能更容易实现果蝇算法，因为标签是稀疏的）用于离散哈希（25）和学习（被称为数据依赖哈希（13）的各种量化技巧）。还有一些方法可以加快随机投影的乘法运算，这两种方法都适用于快速的约翰逊-林登施特劳斯变换(26,27)和快速稀疏矩阵乘法。我们的目标是比较两个概念上的不同方法来解决最近邻的搜索问题，在实际应用中，所有的这些扩展都要移植到果蝇算法中。

一些果蝇算法曾经被使用过。举个例子，MinHash 和 Winner-all 哈希（29）都使用了 WTA 相似的插件，但都不建议扩展维度。类似的，在许多 LSH 家庭中也适用随机投影，但在我们的知识中，没有使用稀疏的二进制投影。果蝇的嗅觉回路似乎已经进化到使用这些计

算成分的独特组合，已经有证据表明，用于果蝇回路图形的三个印记也可以出现在大脑的其他区域或者其他物种中（表一），因此，局部敏感哈希可能是一个在大脑中使用的一般计算原则（30）。







图一所示：

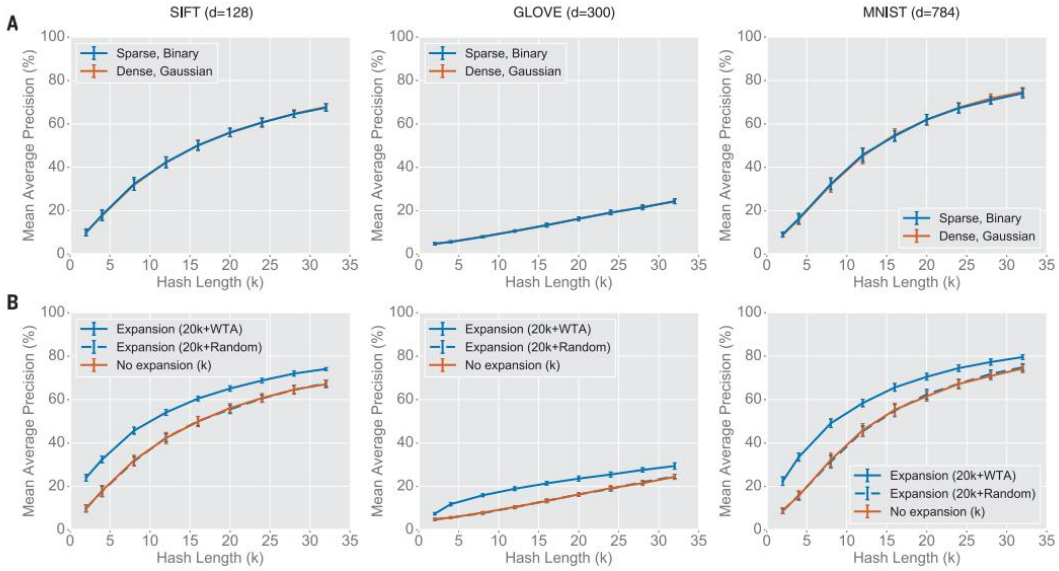
在果蝇嗅觉回路和局部敏感哈希（LSH）之间的映射。

（A）果蝇嗅觉回路原理图。第一步，果蝇鼻子中的 50 个气体受体神经元把轴突送到肾小球里的 50 个投射神经元上，由于这个投影，每一种气味都将以激活率的指数分布，所有的气味和气味浓度都拥有相同的均值。第二步，投射神经元进行了维度的拓展，投射到 2000 个 KCs，并且通过一个稀疏的二元随机投影矩阵连接。在步骤 3 中，KCs 接收来自前双侧（APL）神经元的反馈抑制，这样剩下的前 5%会保持对气味的激活刺激，这 5%对应于气味的标签（哈希）。

（B）气味反应说明。类似的气味（如甲醇和乙醇）被分配的标签与不同气味分配的标签相比更相似。较深的阴影表示较高的活性。

（C）传统的 LSH 算法与果蝇算法之间的区别。在这个例子中，LSH 和果蝇算法的计算复杂度是相同的。输入维数 D=5。LSH 计算 M = 3 个随机投影，每一个投影都需要 10 个操作（5 个乘法，5 个加法）。果蝇算法计算 M=15 个随机投影，每一个都需要两个加法操作。因此，两者都需要 30 个总操作

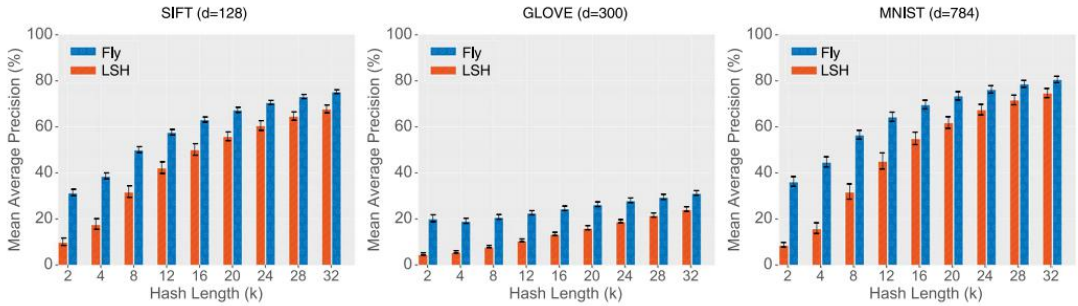
X,输入特征向量；R，高斯随机变量；W 离散化的宽度。



图二所示：不同的随机投影类型和标签的选择方法比较。在所有的图中，X 轴都表示哈希的长度，y 轴表示最近邻精确程度的平均值（越高越好）。

（A）随机稀疏二进制的预测与密集的高斯随机投影几乎达到了相同的性能，但是前者的计算量要远小于后者

（B）扩展维度（k-20k），以及通过‘WTA’进行稀疏化进一步提高了相对于不进行扩展维度方法的性能。三个基准数据集的结果都是一致的。误差条表示标准差的检测已经超过了 50 次。



图三所示：果蝇算法与 LSH 算法的比较。在所有的图中，X 轴代表哈希的长度，Y 轴精度的平均值（越高越好）。在果蝇算法中使用了大概 10D（10 维）的拓展。在三个数据集中，果蝇的算法都要优于 LSH，而且其中最突出的是在哈希长度较短时。误差条表示关于标准差的检测已经超过了 50 次。

Table 1. The generality of locality-sensitive hashing in the brain. Shown are the steps used in the fly olfactory circuit and their potential analogs in vertebrate brain regions.

	Step 1	Random projection	Step 2 (expansion)	Step 3 (WTA)
Fly olfaction	Antennae lobe; 50 glomeruli	Sparse, binary; samples six glomeruli	Mushroom body; 2000 Kenyon cells	APL neuron; top 5%
Mouse olfaction	Olfactory bulb; 1000 glomeruli	Dense, weak; samples all glomeruli	Piriform cortex; 100,000 semi-lunar cells	Layer 2A; top 10%
Rat cerebellum	Precerebellar nuclei	Sparse, binary; samples four precerebellar nuceli	Granule cell layer; 250 million granule cells	Golgi cells; top 10 to 20%
Rat hippocampus	Entorhinal cortex; 30,000 grid cells	Unknown	Dentate gyrus; 1.2 million granule cells	Hilar cells; top 2%

表 1 所示：大脑中局部敏感的哈希的普遍性。表中所显示的是在脊椎动物大脑区域存在的与果蝇嗅觉回路类似的结构。

	步骤一：	随机投影	步骤二（拓展）	步骤三（WTA）
果蝇的嗅觉回路	触角神经叶，50 个肾小球	稀疏二进制，6 个肾小球样品	蕈形体，2000 个凯尼恩细胞	取 APL 神经元前 5%
老鼠的嗅觉回路	嗅觉球体，1000 个肾小体	密集，弱，所有的肾小球都成为样品	梨状皮质，100000 半月板细胞	取前 10%Layer 2A
鼠的小脑	小脑前核	稀疏，二进制 4 个小脑前核样品	颗粒细胞层 2.5 亿个颗粒细胞	取前 10%-20%的高尔基细胞
鼠的海马体	内嗅皮层和 3000 个网格细胞	不清楚	齿状回区域，120 万颗粒细胞	取前 20%肺部细胞