通过快速搜索和发现密度峰值聚类

聚类分析的目的是根据它们的相似性将元素分类。它的应用范围从天文学到生物信息学、文献计量学和模式识别。我们提出的方法是基于这样一种观点，即集群中心的特征是密度高于其邻域，并且与密度较高的点相对距离较大。这个假设构成了聚类过程的基础，在这个过程中，集群的数量是直观的，离群值被自动发现并被排除在分析之外，不管它们的形状和它们嵌入的空间的维度，集群都能被识别出来。我们在几个测试用例中演示了算法的能力。

聚类算法试图根据它们的相似性将元素划分为类别或集群。目前有几种不同的聚类策略。但即使在聚类的定义上这些算法也没有达成共识。在K-means(2)和K-medoids(3)方法中，集群是一组数据，其特征是到集群中心的一小段距离。目标函数，通常是对一组假定的群集中心的距离的总和，在找到最佳的聚类中心之前进行了优化。然而，由于数据点总是被分配到最近的中心，所以这些方法不能检测非球形簇。在基于分布的算法中，试图将观测到的数据点看作一个预先定义的概率分布函数;这种方法的准确性取决于试验概率表示数据的能力。

基于数据点的局部密度的方法很容易检测出具有任意形状的聚类。在具有噪声的应用程序的基于密度空间的聚类中，人们选择密度阈值，丢弃噪声密度低于该阈值的区域中的点，并将其分配给不同簇的高密度的区域。然而，选择适当的阈值比较容易，不存在于均值漂移聚类方法中的缺点。定义一个簇作为收敛到密度分布函数的同一局部最大值的一组点。这种方法可以寻找非球形簇，但只适用于由一组坐标定义的数据，并且计算成本高。

在这里，我们提出另一种方法。与K-medoids方法类似，它仅基于数据点之间的距离。就像DBSCAN和mean-shift方法一样，它能够检测非球形的集群，并自动地找到正确的集群数量。将聚类中心定义为数据点密度中的局部最大值。 然而，与mean-shift方法不同，我们的程序不需要将数据嵌入向量空间中，并且明确地将每个数据点的密度最大化。

该算法的基础是假设聚类中心被邻域密度较低的点所包围，并且它们距局部密度较高的点相对较远。 对于每个数据点i，我们计算两个量：其局部密度ρ和距密度较高点的距离δ。这两个量仅取决于数据点之间的距离，假定它们满足三角不等式。 数据点i的局部密度被定义为

其中如果x <0，则χ（x）= 1，否则χ（x）= 0，并且dc是截止距离。 基本上，等于比dc更接近点i的点的个数。该算法仅对在不同点的相对大小敏感，这意味着，对于大数据集，分析结果对于dc的选择是稳健的。通过计算点i与其他密度较高的点之间的最小距离来测量

对于密度最高的点，我们通常采用。 请注意，比密度中局部或全局最大值点的最近邻距离要大得多。 因此，聚类中心被认为是的值异常大的点。

这是该算法的核心，由图1中的简单例子来说明。图1A显示了嵌入在二维空间中的28个点。 我们发现密度最大值在点1和10，我们将其确定为聚类中心。 图1B显示了每个点的作为函数的图; 我们将这个表示称为决策图。对于点9和点10，具有相似的值的的值是非常不同的：点9属于点1的集群，并且具有更高的其他点非常接近它，而最接近的更高点 点10的密度属于另一个集群。因此，正如预期的那样，高和高的唯一点是聚类中心。因为它们是孤立的，因此点26,27和28具有相对较高的和较低的; 他们被认为是由单一点组成的集群，即异常值。

在找到聚类中心之后，每个剩余点将被分配到与其最近邻居密度相同的同一个聚类。与其他聚类算法相比，聚类分配在单个步骤中执行。而其他的聚类算法是迭代计算的。

在聚类分析中，定量地度量赋值的可靠性通常是有用的。在基于函数优化的方法中，其收敛值也是一种自然的质量度量。在像DBSCAN这样的方法中，我们考虑密度值高于阈值的可靠点，这可能导致低密度簇，如图Fig.2E中的那些被归类为噪音。 在我们的算法中，我们不引入噪声信号截止点。相反，我们首先为每个群集找到一个边界区域，该边界区域被定义为分配给该群集的点集，但距离属于其他群集的数据点的距离为dc。然后，我们为每个群集找出其边界区域内密度最高的点。我们用来表示它的密度。 密度高于的集群点被认为是集群核心的一部分（鲁棒分配）。其他点被视为群晕的一部分（被视为噪音）。

为了对我们的程序进行基准测试，让我们首先考虑图2中的测试用例。数据点来自于一个非球面和强重叠峰的概率分布(图2A);与最大值对应的概率值几乎与mag- nitude的顺序不同。图2的B和C分别从图2A的分布中提取了4000和1000点。在相应的决策图(图2的D和E)中，我们只观察到5个点的大值D和较大的密度。这些点在图中表示为大的实心圆，对应于簇中心。在中心被选中之后，每一个点是分配给一个集群或光环。该算法捕获了概率峰的位置和形状，甚至是对应于非常不同密度(图2C的蓝色和浅绿色点)和非球面峰。此外，通过目视检查图2A的分布，分配给光环的点对应的区域，不会被分配到任何峰值。

为了更加定量地证明该过程的鲁棒性，我们通过从图2A中的分布中绘制10,000个点来进行分析，考虑作为在该样本上获得的聚类分配的参考。 然后，我们通过仅保留一小部分点来获得缩减的样本，并独立地为每个缩减的样本执行聚类分配。图2F显示作为缩小样本的大小的函数，分配给群集的点的比例不同于它们在参考案例中被分配的点的比例。即使对于包含1000个点的小样本，错误分类点的分数仍然远低于1%。

不同dc的数据在图2B中产生了一致的结果。根据经验选择dc，这样邻居的平均数量就在数据集中的总点数的1%到2%之间。由少量样本组成的数据集，可能会受到较大的统计错误的影响。在这些情况下，通过更准确的测量来估计密度可能是有用的。

接下来，我们在图3中给出的测试用例中对算法进行了基准测试。为了计算少数点情况下的密度，我们采用了（11）中描述的指数内核。在图3A中，我们考虑一个来自（12）的数据集，获得的结果与原始文章的结果相当，其中其他常用方法失败。在图3B中，我们考虑了从（13）中取数据分布中具有高重叠的15个簇的例子;我们的算法成功地确定了数据集的集群结构。在图3C中，我们考虑FLAME（通过隶属度局部近似的模糊聚类）方法（14）的测试用例，其结果与原始方法相当。在最初为了说明图4D所示的基于路径的谱聚类（15）的性能而引入的数据集中，我们的算法正确地找到三个聚类，而不需要生成连接图。作为对比，在S3和S4中，我们给出了这四个测试用例和图2中的例子的由K均值（2）得到的聚类结果。即使K均值使用正确的K值进行优化，在大多数情况下，聚类效果不符合视觉。

该方法对度量的变化是鲁棒的，不会显着影响dc以下的距离，即保持Eq. 1中的密度估计量不变。显然，Eq. 2的距离将受到这种度量变化的影响，但容易认识到，如图fig. S5所示，决策图的结构（特别是具有大的值的数据点的数量）是密度值排序的结果 而不是距离点之间的实际距离。

我们的方法只需要测量所有成对数据点之间的距离，而不需要参数化概率分布或多维密度函数。因此，其性能不受嵌入数据点的空间的固有维度的影响。我们验证了，在一个包含16个256维16簇的测试用例中，该算法找到了簇的数目并正确分配了这些点（图S6）。对于来自（17）的具有三种类型小麦种子的七个x射线特征的210个测量值的数据集，该算法正确地预测了3个集群的存在，并正确地分类了分配给集群核心的97%的点(图S7和S8)。

我们还将该方法应用于Olivetti人脸数据库，这是一种广泛用于机器学习算法的基准，目的是在未经过任何培训的情况下识别数据库中的主题数量。这个数据集对我们的方法构成了严重的挑战，因为“理想”数量的聚类（即不同的主题）与数据集中的元素数量（即不同图像，每个主题10个）相当。这使得对密度的可靠估计变得困难。通过以下（19）计算两图像之间的相似度。密度通过高斯核（11）估计，方差dc = 0:07。对于这样一个小集合，密度估计量不可避免地受到大的统计误差的影响;因此，我们将图像分配给一个集群，遵循比上述示例稍微严格的标准。只有在距离小于直流的情况下，图像才会被分配到密度较高的最近图像的相同聚类。因此，图像比任何其他更高图像的dc图像都要高，密度保持未分配。在图4中，我们显示了对数据集中前100个图像执行的分析结果。决策图（图4A）显示了几个不同密度最大值的存在。与其他例子不同的是，它们的确切数字并不清楚，这是其数据点的稀疏性结果。选择中心数量由按降序排列的gi-ridi图提供（图4B）。该图表明，这个数量，即按照定义大的聚类中心，开始以低于9的点异常增长。因此，我们使用九个中心进行分析。在图4D中，我们用不同的颜色显示与这些中心相对应的簇。七个集群对应不同的主题，表明该算法能够“识别”10个中的7个主题。第八个主题出现在两个不同的集群中。当对数据库的所有400幅图像执行分析时，决策图再次不能清楚地识别簇的数量（图S9）。但是，在图4C中表明通过增加越来越多的假定中心，可以明确地识别大约30个主题（图S9）。当包含更多的中心时，一些主题的图像被分割在两个集群中，但所有的集群仍然都是纯粹的，即只包括相同主题的图像。在（20）之后，我们还计算了与同一个聚类正确关联的同一主题的图像对的比例以及错误分配给同一个聚类的不同主题的图像对的比例。如果在分配中没有在dc分量上应用截止值（即如果在我们的一般公式中应用了我们的算法），那么得到〜68％和〜1.2％，约42到50个中心，聚类效果与最先进的无监督图像分类方法相当。

最后，我们将聚类算法用于分析三聚赖氨酸在300 K水中的分子动力学轨迹。在这种情况下，团簇大致对应动能盆地，即系统的独立构象，其在相当长的时间内稳定并被自由能垒隔开，在微观时间尺度上很少被穿过。 我们首先通过标准方法（22）基于动力学矩阵的谱分析来分析轨迹，其动力学矩阵的特征值与系统的弛豫时间相关联。在第七个特征值之后存在间隙（图S10），表明系统有八个盆地; 与此相一致，我们的聚类分析（图S10）产生了8个聚类，其中包括与定义动力学盆地的那些构造一一对应的构造（22）。

和其他基于密度的聚类算法（9,10）一样，用密度极大值来识别聚类是一种简单而直观的选择。但有一个重要的缺点，如果随机生成数据点，那么为有限样本大小估计的密度一致性较差，而是以几个最大值为特征。然而，决策图使我们能够区分由噪声产生的密度波纹的真正簇。定性地说，只有在前一种情况下，对应于聚类中心的点与其他点相距较远。对于一个随机分布，将观察到一个连续的分布和的值。事实上，我们对从超立方体中的均匀分布随机生成的点集进行分析。数据点之间的距离用超立方体上的周期边界条件计算。这个分析表明，对于随机分布的数据点， = 是根据幂律分布的，指数取决于嵌入点的空间的维度。具有真正群集的数据集的的分布，如Figs. 2 到4，与幂律明显不同，特别是在高部分（图S11）。这一观察结果可以为自动选择聚类中心的标准提供依据，也可以用统计学方法验证用我们的方法进行分析的可靠性。