# 基于优化初始点的K-means聚类算法

张昱、曹泽宇、付少伟、潘港

（安徽大学 计算机科学与技术学院 计算机科学与技术专业，安徽 合肥 230601）

摘要：聚类的实用方法是使用迭代程序（例如K-Means，EM），使其收敛到许多局部最小值之一。已知这些迭代技术对初始启动条件特别敏感。我们提出了一个算法，用于根据给定的初始条件计算精确的起始条件，该条件基于估计分布模式的有效技术。优化的初始启动条件允许迭代算法收敛到“更好”的局部最小值。该程序适用于离散数据和连续数据的一大类聚类算法。我们演示了这种方法在流行的K-Means聚类算法中的应用，并证明精确的初始起点确实导致了改进的解决方案。细化运行时间比集群完整数据库所需的时间要短得多。该方法是可扩展的，并且可以与可扩展聚类算法结合以解决数据挖掘中的大规模聚类问题。

关键词：K-means，局部最小值，迭代，大规模聚类

## 研究背景

聚类是包括数据挖掘[1]，统计数据分析[2]，压缩[3]和矢量量化等多个领域的重要应用领域。 在机器学习[4]，模式识别[5]，优化[6]和统计学文献[7]中已经以各种方式制定了聚类。 基本的聚类问题是将彼此相似的数据项组合在一起（聚类）。 最常用的聚类方法是将其视为密度估计问题[8]。 我们假设除了每个数据项的观察变量之外，还有一个隐藏的未观测变量，表示给定数据项的“群集成员资格”。 因此，假设数据是从混合模型到达的，混合标签（簇标识符）是隐藏的。

这是EM的一个特例，它假设：

1）每个簇由球状高斯分布建模;

2）每个数据项被分配到单个集群;

3）混合权重（Wi）相等。

请注意，K-Means [DH73，F90]是根据数值（连续值）数据定义的，因为它需要计算平均值的能力。 K-Means的离散版本存在并且有时被称为苛刻的EM [9]。 KMeans算法找到局部最优解，使每个数据点与其最近的聚类中心之间的L2距离的平方和（“失真”）最小化，这相当于在给出上述假设的情况下最大化可能性。

考虑到数据迭代细化方法（包括EM和K均值）是最有效的，存在各种方法来解决确定（局部）最优参数值的问题。基本算法的工作原理如下：

1）将模型参数初始化为当前模型;

2）假定当前模型是正确的，确定数据项的成员集群;

3）假定在2）中获得的数据成员是正确的，重新估计当前模型的参数，产生新的模型;

4）如果当前模型和新模型彼此足够接近，则终止，否则转到2）。

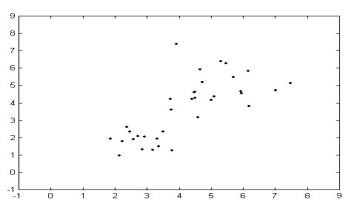
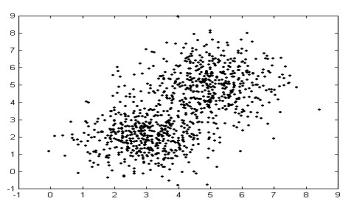


图1. 2-d中的两个高斯凸起：全样本与小子样本

我们专注于初始化步骤1.给定步骤1的初始条件，算法定义从初始点到解决方案的确定性映射。 K-Means和EM算法都有限地收敛到给定模型的数据可能性最大的一个点（一组参数值）。确定性映射意味着局部最优解对初始点的选择很敏感。

有关聚类的初始化方法的工作很少。根据[10]中提到了通过运行K聚类问题来初始化平均值的递归方法。这种方法的一个变种是取整个数据的均值，然后随机扰动K次。在离散数据上的EM情况下，这种方法似乎并不比随机初始化好。在中，通过选择沿该坐标的K个最密集的“分档”来确定沿任何一个d坐标轴的初始值的值。

初始化EM的方法包括K-Means解决方案，分层凝聚聚类（HAC）和“边际+噪声”。已经发现对于使用HAC或“边际+噪声”初始化的离散数据的EM显示对随机初始化没有改进。

对于本文的其余部分，我们将重点放在K-Means算法上，尽管该方法可以改进其他聚类算法的初始点。我们关注K-Means的理由如下：

1）它是一种标准的聚类技术，用于各种应用，甚至用于初始化更昂贵的EM聚类算法;

2）无论使用哪种聚类算法，K-Means都是通过我们的初始化细化方法内部使用的;

3）本文的目的是说明细化过程，而不是评估各种聚类算法。

## 二、具体实现

### 1．集群

为了克服估计噪音的问题，我们采用以下程序。多个子样本，如J，被绘制并独立聚类，产生真实聚类位置的J估计。为了避免与每个J解决方案相关的噪音，我们采用“平滑”程序。然而，为了“最好”执行这种平滑，需要解决以“最佳”方式将K \*个点（每个具有K个群集的J个解）分组为K个群组的问题。图3显示了对于K = 3，J = 4获得的4个解。 “真”的群集手段由“X”表示。 A表示从第一个子样本获得的3个点，B的第二个，C的第三个和D的第四个。然后问题是确定D1将与A1分组，但A2不应与{A1，B1，C1，D1}分组。

### 2.细化算法

细化算法最初选择数据的J个小随机子样本，Si，i = 1，...，J。子样本通过K-Means进行聚类，条件是终止处的空簇将重新分配其初始中心，并且子样本将被重新聚类。集合CMi，i = 1，...，J是形成集合CM的子样本上的这些聚类解决方案。然后CM通过用CMi初始化的K-Means产生解决方案FMi。然后选择精确的初始点作为在集合CM上具有最小失真的FMi。

聚类CM是对CMi的平滑处理，以避免由子样本Si中包含的异常值“破坏”解决方案。细化算法将输入作为输入：SP（初始起始点），Data，K和J（从数据中取出的小子样本的数量）：

算法细化（SP，Data，K，J）

CM =φ

1.对于i = 1，...，J

一个。让Si是一个小的随机子样本

数据

湾令CMi = KMeansMod（SP，Si，K）

C。 CM = CM∪CMi

2. FMS =φ

3.对于i = 1，...，J

一个。令FMi = KMeans（CMi，CM，K）

设FMS = FMS∪FMi

让FM = ArgMin {Distortion（FMi，CM）}

5.后退（FM）

我们定义了细化算法调用的以下函数：KMeans（），KMeansMod（）和失真（ ）。 KMeans只是对经典K-Means的调用

算法采用：初始起点，数据集和聚类数K，返回一组K个二维矢量，K个聚类的质心估计。

KmeansMod采用与KMeans（上面）相同的参数，并执行与经典K-Means相同的迭代过程，除了以下轻微修改。当经典的KMeans已经收敛时，将检查K个群集的成员资格。

成员资格（这通常发生在对小子样本进行聚类时），将空群集质心的相应初始估计值设置为离其指定聚类中心最远的数据元素，并从这些新的初始中心再次调用经典K均值。

启发式重新分配的动机如下：如果在终止K均值时存在空集群，那么将所有空集群重新分配到离其各自中心最远的点将最大程度地降低失真。图2（左）描述了具有零成员资格的群集的一个例子。

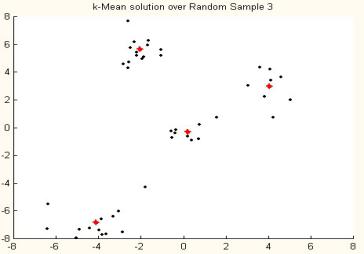
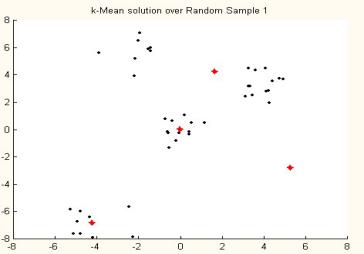


图2：对从相同分布中抽取的两个不同样本进行聚类的结果，并用相同的起点（用'+'表示的产生的解决方案）进行初始化

失真采用一套K均值和数据集的估计值，并计算每个数据点距其最近平均值的平方距离之和。这个标量度量了一组数据集对数据集的拟合程度。 KMeans算法终止于对这种失真函数局部最优的解决方案。

3．大型数据库的计算复杂性和可扩展性

细化算法主要用于处理大型数据库。当处理小数据集（例如Irvine存储库中的大多数数据集）时，从许多不同的起点应用经典的K-Means算法是可行的选择。但是，随着数据库大小的增加（特别是维度），高效和准确的初始化变得至关重要。对具有多维和数万或数百万记录的数据集的聚类会话可能需要数小时至数天。在[BFR98]中，我们提出了一种将聚类扩展到超大型数据库的方法，专门针对不适合RAM的数据库。我们证明，与应用于数据库[BFR98]的适当大小的随机子样本的经典K均值相比，可以实现精确的聚类。可扩展的集群方法显然受益于更好的初始化。

由于我们的方法适用于非常小的数据样本，因此初始化速度很快。例如，如果我们使用整个数据集大小的1％（或更少）的样本大小，则可以在时间复杂度上运行超过10个样本的试验，该时间复杂度低于聚类整个数据库所需时间的10％。对于非常大的数据库，初始样本的大小可以忽略不计。

对于数据集D，如果是聚类算法r要求Iter（D）迭代对其进行聚类，则时间复杂度为| D | \* Iter（D）。一个小的子样本S⊆D，其中| S | << | D |通常需要明显更少的迭代进行聚类。从经验上来说，期望Iter（S）<Iter（D）是安全的。因此，给定用户分配给细化过程的特定时间预算，我们只需确定在细化过程中使用的子样本的数量J.当| D |非常大，而| S |是| D |的一小部分，细化时间实际上可以忽略不计，即使对于大J也是如此。细化算法的另一个理想特性是它可以轻松扩展到非常大的数据库。唯一的内存要求是在RAM中保存一个小的子样本。在二级聚类阶段，只有从J子样本获得的解决方案需要保存在RAM中。图5：左：来自随机初始点（蓝色方块）的K均值解（大红色圆圈）。右：精确的初始点（红色圆圈），随机初始点（蓝色方块）。注意，我们假设可以从大型数据库中获得随机样本。虽然这听起来很简单，但实际上这可能是一项艰巨的任务。除非可以保证数据库中的记录不是由某个属性排序，否则随机抽样可能与扫描整个数据库一样昂贵。请注意，在数据库环境中，人们认为数据表（视图）可能不会作为物理表存在。查询的结果可能涉及连接，分组和排序。在很多情况下，数据库操作会对结果集进行特殊排序，并且通常不能假设生成的数据库视图的“随机性”。

## 三、实验数据

图3总结了在数据范围内统一确定的10个随机初始点的平均结果。注意，从“Refined（J = 10）”计算得到的K-Means解与从随机初始点或“Refined（J = 1）”计算出的K-Means解相比，始终更接近真正的高斯均值生成数据集）“的初始点。在左边，我们总结了从精确的初始点计算的经典KMean解的平均距离与真正的高斯均值相对于平均距离的比率。在这些结果中值得注意的是以下事实：

1.对于维度2-50，细化方法（Refined（J = 10））始终比随机起始点（Unrefined）和在1个子样本（Refined（J = 1））上细化的点好。

2.对于维度100，在10个独立试验中的9个中，我们的细化方法比随机起点更好。

3.精炼剂溶液在离真正高斯平均值近2.34（d = 3）和6.44（d = 50）倍之间，比随机初始点的溶液和1.09（d = 3）和4.80（d = 50）比从“精制（J = 1）”初始点计算的解决方案要多。

在一次运行中，我们做得更糟。这解释了100个维度的大差异数量。如果我们排除一个数据点，那么方差就会下降到所有其他维度的范围内。对于具有小维度和最大比率的数据集出现最小比率的情况出现在具有大维度的数据集上，这表明大规模数据集的改进算法的实用性。

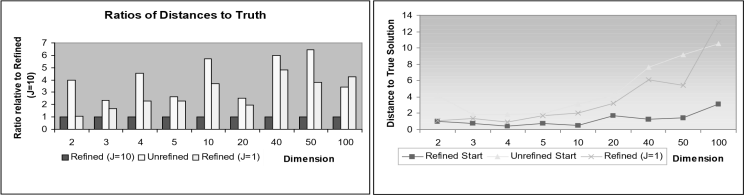


图3.以维度增加作为性能评判标准

## 总结

我们已经提出了一种用于改进一般类别聚类算法的初始起点的快速且有效的算法。细化算法对给定数据库的小子样本进行操作，因此需要存储完整数据库所需的总内存的一小部分，并使这种方法对于大规模聚类问题非常有吸引力。

然而，我们注意到，同样的方法很容易被推广到其他算法，甚至是离散数据（在这些数据上没有定义）。中介绍了广义方法及其用于初始化EM算法的用法以及实验结果。这里的关键洞察是，如果使用某种算法ClusterA来对数据进行聚类，那么ClusterA也可用于对子样本进行聚类。算法ClusterA将产生一个模型。该模型基本上由其参数描述。参数在一个连续的空间中。对聚类进行聚类的阶段保持不变;即在这一步我们使用K-Means算法。使用K-Means的原因是，这个阶段的目标是找到模型的“质心”，在这种情况下，K-Means的严格成员分配是可取的。

## 参考文献

[1] J. Banfield and A. Raftery, “Model-based gaussian and non-Gaussian Clustering”, *Biometrics,* vol. 49: 803-821, pp. 15-34, 1993.

[2] C. Bishop, 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

[3] D. Fisher. “Knowledge Acquisition via Incremental Conceptual Clustering”. *Machine Learning*, 2:139-172, 1987.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego, CA: Academic Press, 1990

[5] L. Kaufman and P. Rousseeuw, 1989. Finding Groups in Data, New York: John Wiley and Sons.

[6] 王玲，薄列峰，焦李成． 密度敏感的普聚类［J］． 电子学报，2007，35

[7] 姚跃华，史秀岭． 一种优化初始中心的 K-means 粗糙聚类算法

[8] 袁方、周志勇，宋鑫．初始聚类中心优化的 K-means 算法［J］．计算机工程，2007，33(3) : 65-66．

[9] 汪静，印鉴，郑利荣，等． 基于共同评分和相似性权重的协同过滤推荐算法［J］． 计算机科学，2010，37( 2) : 99-103．

[10] 黄创光，印鉴，汪静，等． 不确定近邻的协同过滤推荐算法［J］． 计算机学报，2010，33( 8) : 1369-1376．