**聚类集成：模型一致和弱化分区**

**摘要：**聚类继承是作为一种强大的方法去同时提高解决无监督分类问题的健壮性和稳定性。但是，在多样的分区中找到一个统一的模型是非常困难的，这可以从基于图的、组合的以及统计的角度来解决。本次研究在先前的聚类集成的几个方面上做扩展。首先，我们引入多个聚类的统一表达，并制定了相应的分类聚类问题。第二，我们在聚类空间中使用有限的多项分布混合提出了共识的概率模型。使用EM算法找到组合分区作为对应的解决最大似然问题的解决方案。第三，我们使用广义互信息定义一个新的共识函数，该函数与经典的类内方差标准相关。最后，我们展示了使用数据投影和随机数据分割的弱聚类算法生成的组合分区的成效。为这种弱聚类组件的组合行为提供一个简单的解释型模型。组合精度呗分析为控制组合分区的功率和分辨率以及分区数量的若干个参数的函数。我们同时分析具有不完整信息的聚类集合以及缺乏聚合标签对整体共识质量的影响。实验结果证明提出的方法在一些实际数据集上具有有效性。

# 1.介绍

对比有监督分类，聚类本质上是一个不适定问题，它的解决方案至少违反了关于尺度不变性、丰富性、聚类一致性这些常见假设中的一个。如果没有关于底层数据分布的先验知识，不同的聚类解决方案可能看起来同样合理。每个聚类算法都含蓄或明确地设定一个明确的数据模型，当这些设想不适用于样本数据时它可能会产生错误或者没有意义的结果。因此，有关的数据域的先验信息的可用性对于成功的聚类至关重要，尽管这样的信息难以得到，甚至是专家。识别相关的子空间或可视化可以帮助建立样本数据与基础分布的一致性，或者至少是适当数量的簇的一致性。

集群任务的探索性性质要求有效的方法，这些方法将受益于结合许多单独聚类算法的优势。这是集群集合研究的重点，寻找多个分区的组合，以提供改进的给定数据的整体集群。聚类集合可以超越单个聚类算法在多个方面通常实现的范围：

1. 稳健性 跨域和数据集的平均性能更好
2. 新颖性 找到任何单一聚类算法无法达到的联合解决方案
3. 平行化和可扩展性 数据子集的并行聚类以及随后的结果组合。能够集成来自多个分布式数据源或属性的解决方案

聚类集成同时能够用于多目标聚类，用于在包含冲突的目标函数的各个聚类之间进行折衷。在分布式数据挖掘中，使用多多源的数据或者特征变得越来越重要，例如，我们在[41]的回顾中看到。一些最近的个人研究[10]、[12]、[14]、[15]、[43]、[47]已经倡导将聚类集成作为传统聚类算法分类学中的一个新的分支。其他相关的工作包括[7]、[11]、[16]、[19]、[28]、[31]、[35]。

聚类组合的问题大致上可以定义为如下：给定数据集的多个聚类，找到一个质量更好的组合聚类。虽然聚类组合问题具有经典聚类问题的特征，但是它也有三个主要的问题，这些问题是组合设计所特有的。

1. 共识函数 怎么样去组合不同的聚类？如何解决标签对应问题？如何确保所有组件分区的对称性和公正共识。
2. 聚类的多样性 如何生成不同的分区？组件中多样性的来源是什么？
3. 成分/组件强度 每个输入分区的“弱”程度如何？确保一个成功的组合的最小复杂度的组合聚类是什么？

在多个分类器系统的框架中已经解决了类似的问题。从多个有监督的分类器中得到的组合结果是一个热门的研究领域（[44]、[6]），它提供了聚类组合的主要动机。但是，不可能将组合算法从分类（监督）域机械地应用于聚类（无监督）域。实际上，在聚类中没有标记的训练数据可用；因此，不能使用提高整体精度所必需的地面实况反馈。另外，不同的聚类可能会产生不兼容的数据标签，导致相应的棘手问题。尽管如此，有监督的分类组合原则上证明了多个解决方案如何减少预期错误率的方差分量并提高方案的健壮性。

从监督案例中，我们还了解到弱分类器[32]、[25]、[18]、[6]的正确组合可以在训练数据上实现任意低的错误率，并减少预测误差。可以预期，使用许多简单但计算成本低廉的组件将优先组合通过复杂但计算上涉及算法得到的聚类。

本文进一步推进了集合方法的几个方面，即新的有效共识函数的设计，新分区生成机制的开发，以及聚类精度的研究。

## 我们的贡献

我们提供多个聚类的表示，作为表征数据项的一组新属性。这样的观点直接导致组合问题的公式作为这些属性空间中的分类聚类问题，或者换句话说，是中值分区问题。可以将中值分区视为给定输入分区的最佳摘要。作为优化问题，中值分区是NP-complete问题,具有连续的启发式算法用于近似解。这项工作的重点是聚类集合的主要问题，即共识函数，它创建了组合聚类。我们展示了当广义互信息被用作评估函数时，中值分区如何与经典的类内方差标准相关。提出了基于二次互信息（QMI）的共识函数，并将其简化为特殊变换聚类标签空间中的k-means聚类。

我们还提供了一种新的无监督决策的融合方法，该方法基于贡献聚类空间中共识分区的概率模型。找到共识分区是给定聚类集成的最大似然问题的解决方案。集合的似然函数相对于有限混合分布的参数被优化。该分布中的每个组件对应于目标共识分区中的簇，并且假设为多元多项分布。使用EM算法解决最大似然问题[8]。

QMI和EM共识函数有几个优点。 这些包括：

1. 完全避免解决标签对应问题
2. 低计算复杂度
3. 处理缺失数据的能力，即缺少集合中某些模式的集群标签（例如，当引导方法是用于生成整体）。

我们工作的另一个目标是采用弱聚类算法并结合其输出。 模糊定义，弱聚类算法产生一个分区，它只比数据的随机分区略好。 我们提出两种不同的弱聚类算法作为组件生成机制：

1. 聚类多维数据的随机1维投影。 这可以推广到原始数据空间的任何随机子空间中的聚类
2. 通过使用多个随机超平面分割数据来进行聚类。 例如，如果仅使用一个超平面，则将数据分成两组。

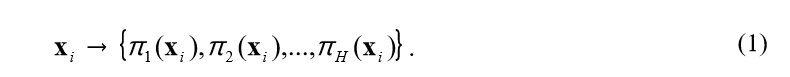
最后，本文比较了不同共识函数的性能。 我们已经研究了基于分类聚类的一系列共识函数的性能，包括基于协同关联的分层方法[15,16,17]，超图算法[47,29,30]和我们新的共识函数。 根据聚类组件的数量和分辨率分析组合精度。此外，我们研究了丢失某些群集标签时的群集性能，这在分布式数据或重新采样方案中经常遇到。

# 2. 多个分区的表示

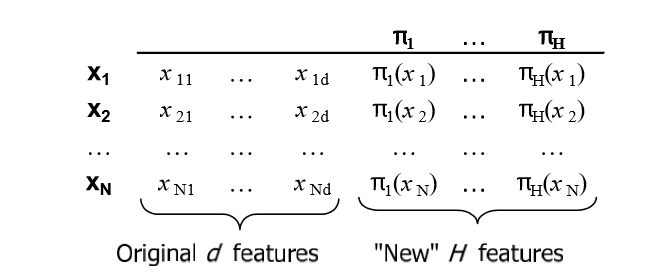
多个分区的组合可以视为一个分区任务本身。 通常，组合中的每个分区都表示为由聚类算法分配的一组标签。 组合分区是另一种聚类算法的结果，该算法的输入是参与分区的聚类标签。 我们将假定标签为标称值。 通常，聚类可以是“软的”，即，由表示分区中每个聚类中的模式成员资格程度的实数值来描述。 我们在下面仅考虑“硬”分区，但是要注意，“软”分区的组合可以通过多种聚类算法来解决，而且看起来并不复杂。

,

假设给定一个N个数据点的集合X = {x1,…, xN}，和一组H个分区的集合Π={π1,…, πH}在X中的对象。X的不同分区为每个点xi返回一组标签



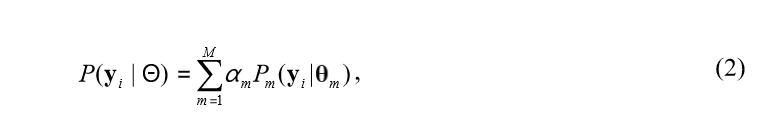
这里，表示H个不同的聚类，并且表示由第j个算法分配给xi的标签。我们不对由不同聚类算法产生的标签之间的对应性做出假设，并且现在也不需要假设。关于数据输入：可以用非度量空间或N×N不相似矩阵表示。为简单起见，我们使用符号。聚类组合是找到一个新的数据X的分区，该分区汇总来自收集的分区的信息。我们的主要目标是在没有X原始模式帮助的情况下，仅从贡献聚类算法传递的标签Y上构造共识分区。因此，为了解决无监督共识问题，忽略了分区和数据的基础结构之类的潜在重要问题。我们强调，新特征空间是由集合引起的。 可以将每个组件分区视为具有分类值（即簇标签）的新特征。假定第i个新特征的值只是分区中的群集标签。因此，将对象x在不同分区中的隶属关系视为新的特征向量y = π (x)，一个H元组。在这种情况下，可以将分区视为特征提取函数。如果仅使用现有的聚类{π1，…，πH}，而不使用数据X的原始特征，则聚类的合并就等同于H元组的聚类问题。因此，合并分区的问题可以转换为分类聚类问题。这样的观点可以洞悉预期组合的属性，可以通过各种统计和信息理论技术来推断。特别是，可以估计组合对组件（特征）相关性的敏感性，并分析各种样本量问题。这种表示的主要优点也许在于，它有助于使用已知的算法进行分类聚类[37，48]，并允许人们以透明的方式设计新的共识启发式算法。数据X的扩展表示形式可以通过具有N行和（d + H）列的表来说明：



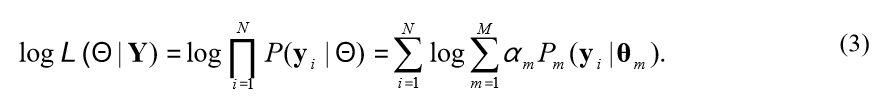
发现共识聚类是向量Y = {yi}的集合的分区，该向量直接转换为基础数据点{xi}的分区。

# 共识的混合模型

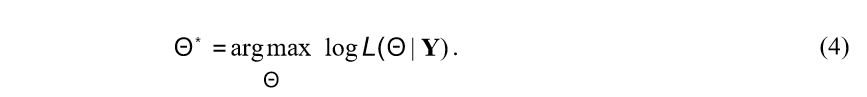
我们针对共识问题的方法是基于有限混合模型，用模式/对象x的簇标签y =π（x）的概率。 主要假设是将标签yi建模为随机变量，该随机变量是从概率分布中抽取的，该概率分布描述为多元变量密度的混合：



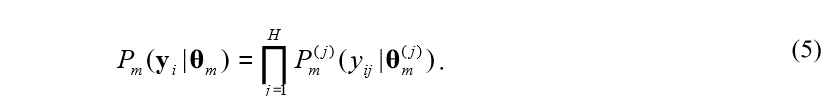
其中每个分量都由θm参数化。 混合中的M组分由共有分区πC的簇标识。混合系数αm对应于簇的先验概率。在该模型中，假定数据点{yi}分两步生成：首先，根据概率质量函数αm绘制一个分量，然后从分布中采样一个点。假定所有数据是独立的并且分布相同。在给定数据集Y的情况下，这允许用参数表示对数似然函数：



现在将共识聚类的目标表述为最大似然估计问题。为了找到给定数据Y的最佳拟合混合密度，我们必须针对未知参数Θ最大化似然函数：

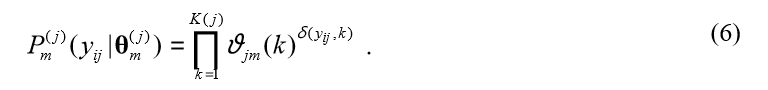


接下来的重要步骤是指定组件条件密度的模型。 注意，在多个数据聚类算法的帮助下，在数据X空间中聚类的原始问题已转换为新的多元特征y =π（x）的空间。为了使问题更易于处理，对向量yi的分量进行了条件独立性假设，即yi的条件概率可以表示为以下乘积：

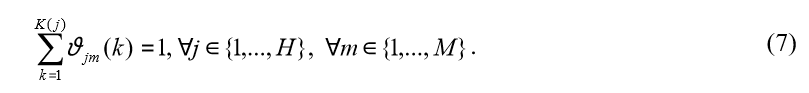


为了激发这一点，可以注意到，即使不同的聚类算法（j索引）不是真正独立的，也可以用等式（5）中的乘积近似，可以用朴素贝叶斯分类器在离散域中的出色表现来证明[34]。 我们的最终目标是通过Y的密度估计的间接途径对X中的数据进行离散标签分配。对于πC中的聚类，模式分配对条件独立性的近似要比概率的估计值敏感得多，如[9]中朴素贝叶斯分类器的分析所支持。

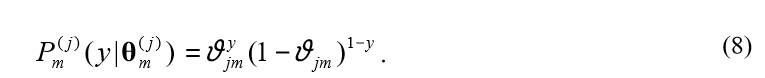
混合模型的最后一个成分是选择向量yi分量的概率密度，因为变量取自变量集中一组聚类标签的标称值划分，自然将它们视为多项式试验的结果



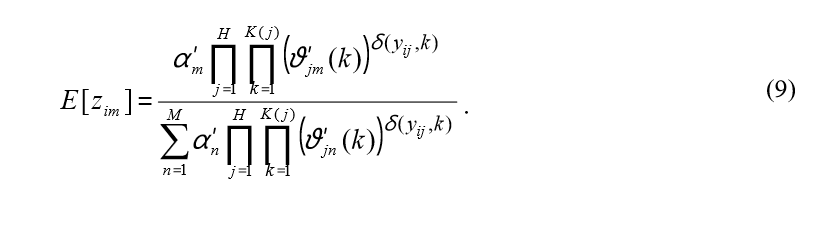
这里，在不失一般性的情况下，将中的簇的标签选择为{1，…，K（j）}中的整数。为了澄清这种表示法，请注意将结果的概率定义为，并且乘积超过分区的标签的所有可能值。而且，这些概率的总和为1:

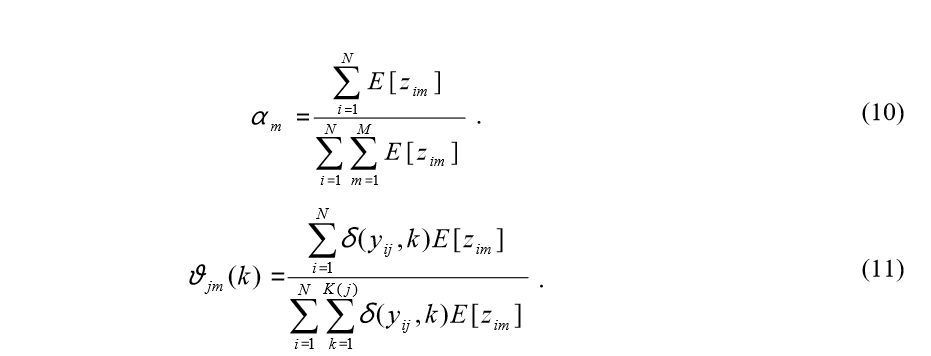


例如，如果第j个分区只有两个群集，并且可能的标签为0和1，则等式（5）可以简化为：



等式（3）中的最大似然问题常不能以闭合形式求解，当所有参数未知时。然而，方程（2）中的似然函数可以使用EM算法进行优化。为了采用EM算法，我们假设存在隐藏数据Z和存在完整数据的可能性（Y，Z）。如果知道的值，则可以立即判断出使用了M个混合成分中的哪一个去生成数据点。附录中给出了具有多元，多项式混合模型的EM解的详细推导。在这里，我们仅给出E和M步骤的方程式，这些方程式在算法的每次迭代中都会重复：

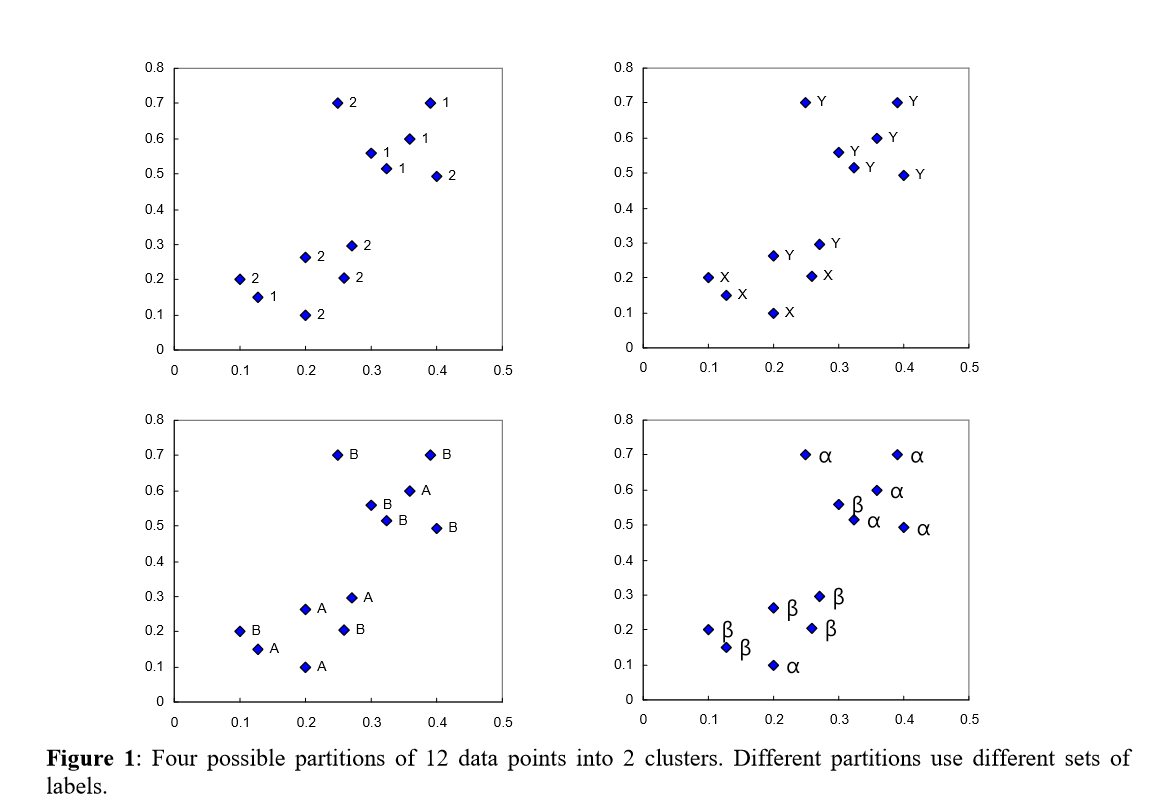


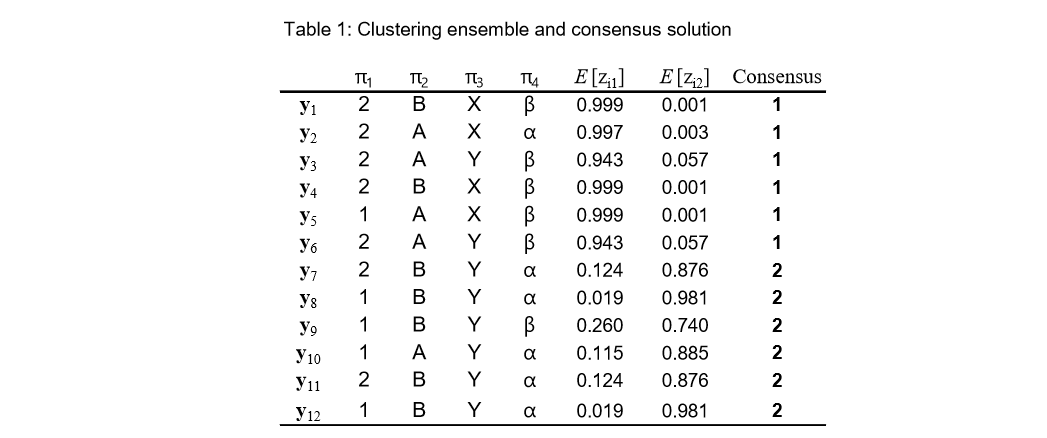


通过简单检查变量E []的期望值即可获得共识聚类问题的解决方案，这是因为E []表示模式由第m个混合分量生成的可能性 。一旦收敛，就将模式分配给具有最大隐藏标签值的组件.

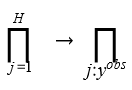
考虑一个简单的ensemble示例很有启发性。图1显示了12个二维数据点的四个2群集分区。 分区使用的不同标签系统强调了对应问题。表1显示了EM算法经过6次迭代以及潜在的共识聚类后潜在变量的期望值。实际上，早在第三次迭代时就出现了稳定的组合，它对应于数据的真实基础结构。

我们的共识混合模型承认具有不完整分区的聚类集成的一般化。这样的分区可能是由于子样本聚类或数据集重采样而出现的。例如，引导程序样本的分区仅提供所选点的标签。因此，这种分区的集合由具有潜在缺失成分的一组簇标签向量表示。此外，聚类标签的不同向量可能会丢失不同的成分。当某些聚类算法未将异常值分配给任何聚类时，也会出现不完整的信息。不同集合中的不同聚类可以将相同点xi视为异常点，否则将导致向量yi的成分丢失。



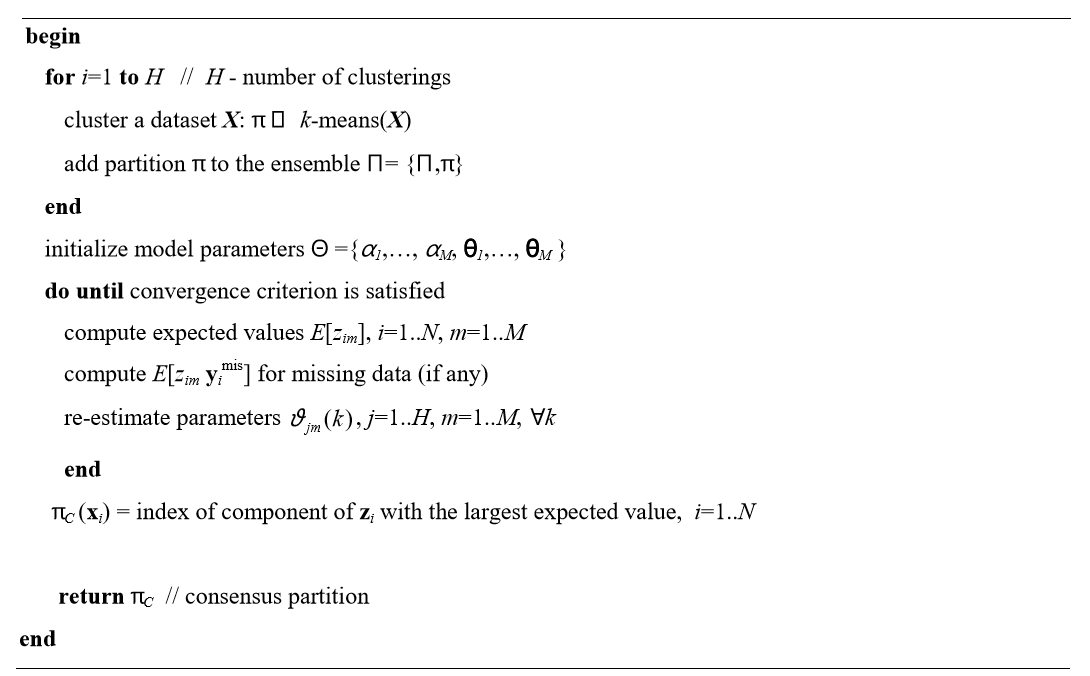


导致信息丢失的另一种情况可能发生在分布式数据的聚类组合或数据集不同副本的聚类集合中。

缺少数据的情况下可以应用EM算法[20]，即某些数据点的聚类标签丢失。在这些情况下，可以将Y中的每个向量yi分为观察分量和缺失分量。合并丢失的数据会导致E和M步骤的计算略有修改。首先，现在从观察到的向量yi分量（即等式1中的乘积）推断出期望值。公式（9）接管已知标签：另外，必须计算期望值并替换它们，以及，在M步重新估计参数。有关处理丢失数据的更多详细信息，请参见[20]。

尽管缺少聚类标签的数据可以通过不同的方式获得，但我们仅分析yi的分量随机完全缺失的情况[46]。这意味着组件丢失的可能性不取决于其他观察到的或未观察到的变量。注意，数据子样本的聚类结果（例如，bootstrap）与聚类整个数据集然后删除标签的随机子集不同。但是，我们的目标是为常规设置提供共识功能。我们希望具有标签缺失的ensemble的实验结果至少在定性上是适用的，即使对于bootstrap簇的组合也是如此。

下面总结了基于混合模型共识算法的集成聚类。请注意，可以使用任何聚类算法来生成集合，而不是此伪代码中所示的k-means算法:

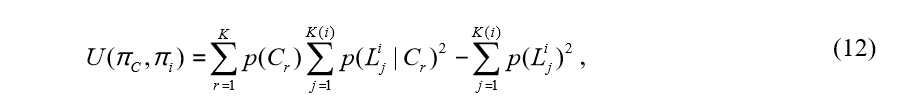


M的值（混合中的组的数）值得单独讨论，这超出了本文的范围。在此，我们假设目标簇数是预定的。然而，应该指出的是，无监督分类中的混合模型极大地促进了聚类的真实数量的估计[13]。问题的最大似然公式特别允许我们在推理过程中使用其他目标函数（例如模型的最小描述长度）来估计M。此外，提出的共识算法可以看作是潜在类别分析的一种版本（例如，参见[4]），该算法具有严格的统计手段来量化候选混合模型的合理性。

尽管有限混合模型可能不适用于原始空间中的图案（初始表示），但该模型更自然地解释了“提取”特征（由分区生成的标签）空间中图案组的分离。这有点让人想起基于核方法的分类方法，该方法依赖于变换空间中的线性判别函数。例如，支持向量聚类[5]在内核转换后寻找与原始模式空间中更复杂的簇形状相对应的球形簇

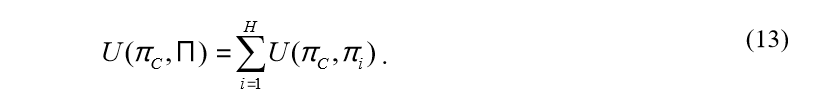
# 聚类的信息-理论共识

另一个候选共识函数基于中位数划分的概念。中值分区σ是Π中现有分区的最佳总结。与协同关联方法相反，中位数分区是根据属性（此处的“属性”（功能）是指整体的分区，而对象是原始数据点。）之间相似性的估算值（即Π中的分区）得出的，而不是根据对象之间的相似性得出的。 在概念聚类的背景下，在COBWEB算法中实现了该方法的一个著名示例[48]。 COBWEB聚类准则估计分区效用，它是Gluck和Corter [21]引入的类别效用函数的总和。 用我们的术语来说，类别效用函数用标签评估候选中间分区与其他分区，通过第j个簇的：



其中，

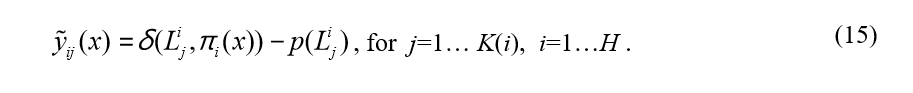
函数将两个分区之间的一致性评估为分区的预期标签数量之间的差异，该期望值可以通过聚类的知识而无需聚类就能正确预测。类别效用函数也可以写为两个分区之间的列联表的Goodman-Kruskal索引[22，39]。相对于C中的所有分区，分区的总体效用可以衡量为成对协议的总和:



因此，最佳的中位数分区应该使整体效用的价值最大化：

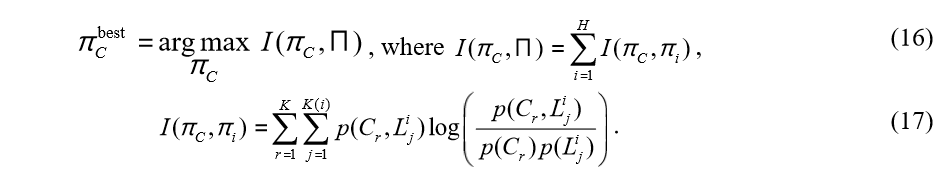


重要的是，Mirkin [39]证明了等式中分区效用的最大化。如果目标分区中的簇数K是固定的，则等式（13）等效于最小化平方误差聚类标准。这有点令人惊讶，因为等式中的分区实用程序功能。（14）仅使用等式（12）的属性间相似性度量，而平方误差准则则利用对象与原型之间的距离。中分类标签的简单标准化有效地将它们转换为定量特征[39]。这使我们能够计算实值距离和聚类中心。该变换用K（i）个二进制特征替换假设K（i）个值的第i个分区，并将每个二进制特征标准化为零均值。换句话说，对于每个对象x，我们可以计算新特征的值，如下所示：



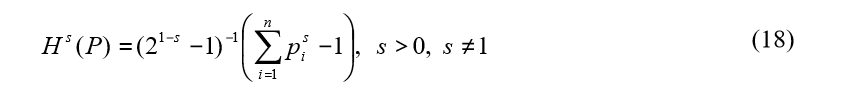
因此，方程(4)中的中位数分配问题的解决方案。如果目标聚类的数量是预先确定的，则可以通过在特征空间中运行的k均值聚类算法来逼近（4）。 我们将这种启发式方法用作共识功能实证研究的一部分。

让我们考虑中位数分割问题的信息理论方法。在此框架中，共识分区的质量由它与Π中给定分区共享的信息决定。 Strehl和Ghosh [47]提出了一个基于经典Shannon互信息定义的目标函数：

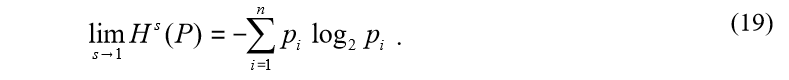


同样，可以通过解决此优化问题找到最佳中位数分区。 但是，尚不清楚如何直接使用这些方程式寻求共识。

我们表明，熵的另一种信息理论定义将把互信息准则简化为前面讨论的类别效用函数。我们从离散度概率分布 [23]的程度s的广义熵出发：



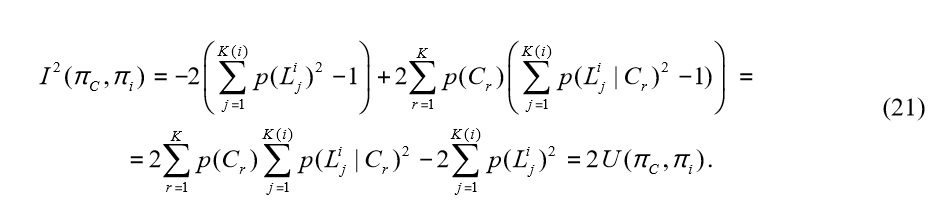
Shannon的熵是式（18）的极限形式：



σ和π之间的广义互信息可定义为：



二次熵（s = 2）特别令人感兴趣，因为当用于类间距离的概率度量时，它与分类误差密切相关。 当s = 2时，广义互信息变为：



因此，广义互信息给出了与等式（13）中的效用函数相同的共识聚类准则。而且，传统的用于属性选择的基尼指数测度也可以从等式（12）和（21）中得出。根据Mirkin的结果，所有这些标准都等同于简单标签转换后的集群内差异最小化。在我们的比较经验研究中使用了二次互信息，混合模型和其他有趣的共识函数。

# 弱聚类的组合

前面的部分讨论了聚类组合的问题，即无论组合中各个分区的性质如何，如何表述共识函数。 现在我们来看为组合生成不同聚类的问题。有几个主要问题。我们是否使用文献中提供的众多聚类算法产生的分区？我们可以放宽对聚类组件的要求吗？现有几种提供不同分区的方法：

1.使用不同的群集算法，例如 k均值，高斯混合，频谱，单链接等[47]。

2.利用某些算法的内置随机性或不同参数，例如 k均值算法[35、15、16]中的初始值和k的各种值。

3.使用数据集的许多子样本，例如引导样本[10，38]。

这些方法依赖于聚类算法，聚类算法本身功能强大，因此涉及计算。我们认为可以使用较弱但较便宜的群集算法生成分区，并且仍然可以获得可比或更高的性能。当然，关键动机是许多此类组件的协同作用将弥补它们的弱点。我们考虑两种简单的聚类算法：

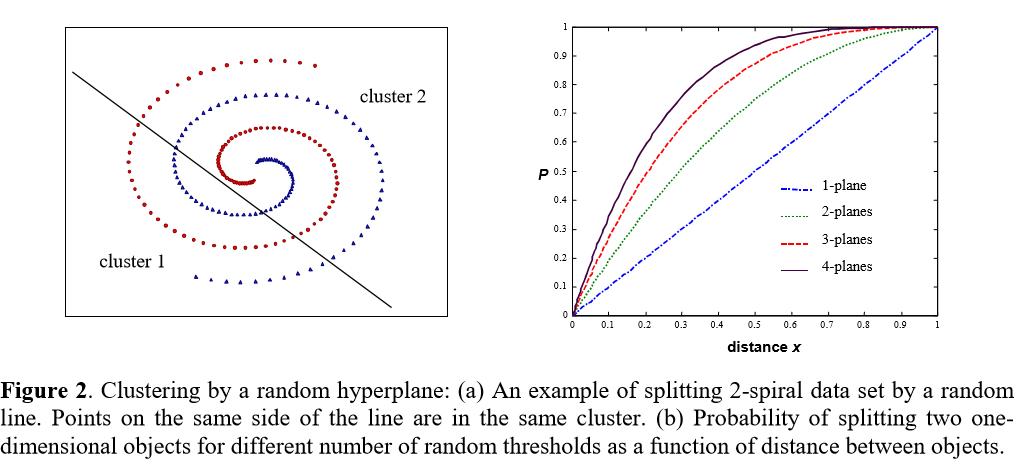
1.将数据聚类到随机子空间。在最简单的情况下，数据会投影到一维子空间（一条随机线）上。k均值算法对投影数据进行聚类并给出组合的分区。

2.通过超平面随机分割数据。例如，单个随机超平面通过将超体积切割成两个区域，将创建一个相当琐碎的d维数据聚类。

我们将证明这两种方法都能够结合适当的共识功能来产生高质量的共识聚类。

* 1. **通过随机超平面分裂**

通过使用随机超平面的直接聚类说明了如何从低信息量组件中获得可靠的共识。随机分裂方法将弱聚类的概念几乎推向了极端。通过剖析包含这些点的d维空间的原始体积的随机超平面切割数据集。由超平面分隔的点被声明为处于不同的群集中。因此，输出簇是凸的。在这种情况下，因为需要的唯一信息是模式是否在相同的集群中，所以使用协关联共识函数是合适的。因此，对于任何一对对象，超平面分区对协关联值的贡献可以为0或1。通过计算分隔对象的超平面的数量，可以实现更精细的距离分辨率，但是为简单起见，在此不使用。考虑一条随机线，它剖析了图2（a）中所示的经典2螺旋数据。尽管任何一个这样的分区都无法揭示真正的基础群集，但对超平面生成机制的分析表明，多个此类分区如何能够发现真正的群集。



首先考虑一维数据的情况。一维空间中的对象拆分由中的随机阈值完成。通常，如果随机选择r个阈值，则会形成（r + 1）个簇。很容易得出，在一维空间中，将两个点间距离为x的对象分离的概率为：



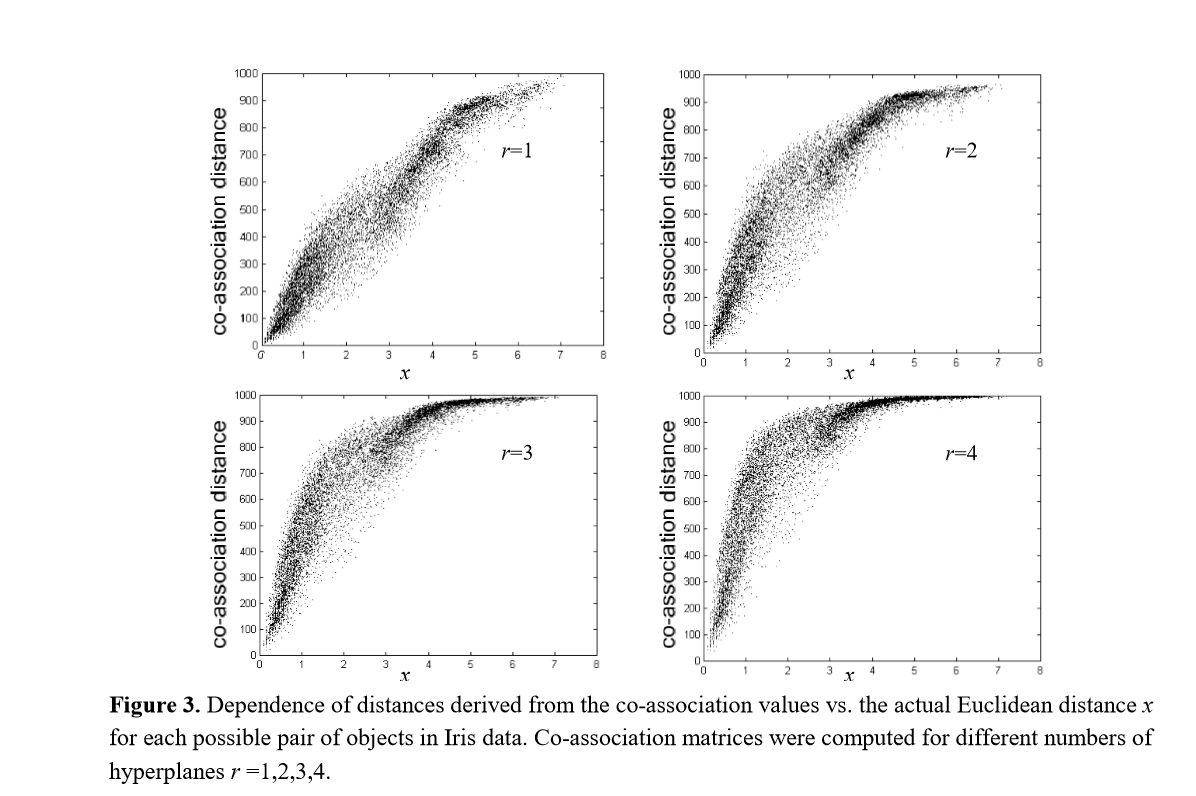
其中L是包含对象的间隔的长度，并且从该间隔的均匀分布中随机抽取了r个阈值点。图2（b）说明了L = 1和r = 1,2,3,4的依赖性。如果使用协关联矩阵来组合H个不同的分区，则两个对象之间的协关联的期望值为，其源自在H次尝试中分裂次数的二项式分布。因此，在组合许多随机分割分区后发现的协关联值通常被认为是各自距离的非线性函数和单调函数，多维数据的情况相似，但是随机超平面的生成要多一些为了在d维上生成随机超平面，我们首先应该在多维区域中绘制一个随机点作为原点，然后我们随机选择一个定义超平面的单位法向矢量u，这两个物体如果，向量p和q将位于同一簇中，否则将分开（此处ab表示a和b的标量积）。如果生成r个超平面，则总概率两个对象保留在同一群集中只是每个超平面不拆分对象的概率的乘积。因此，我们可以预期，控制协关联值的定律接近于公式（22）中在一维空间中获得的定律。

让我们将共关联值的实际依赖性与方程（22）中的函数进行比较。图3显示了通过Iris数据集的随机分割对1000个不同分区进行实验的结果。Iris数据为4维数据，包含150个点。数据项之间成对的距离为11,175。对于所有可能的点对，图3中的每个图都显示了一对被分裂的次数。实际上，可以通过等式（22）中的函数来描述观察到的从互关联值得出的点间“距离”对真实欧几里得距离的依赖性。

显然，点间距离决定了各个关联值的行为。在任意两个给定对象之间进行剪切的概率不取决于数据集中的其他对象。因此，我们可以得出结论，任何与原始点间距离配合得很好的聚类算法，也有望与通过随机分割从多个分区的组合中获得的协同关联值一起协同工作。 但是，当可获得真实距离时，此结果更多是理论值，因为可以直接使用它们来代替关联值。它说明了该方法的主要思想，即多重弱点的协同作用聚类可能非常有效。我们在实验部分中对该算法的聚类质量进行了实证研究。

* 1. **随机子空间中聚类的组合**

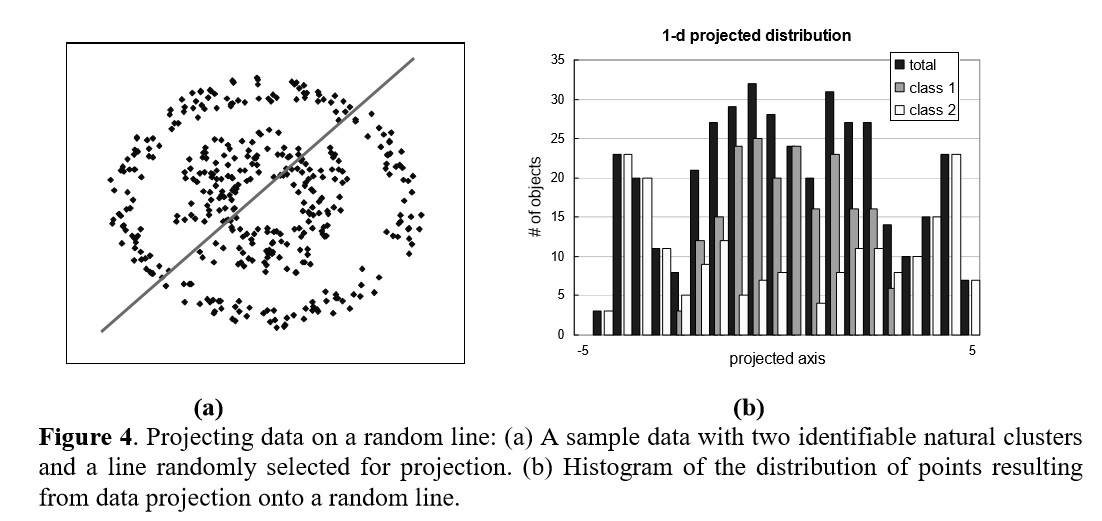
随机子空间是聚类多样性的绝佳来源，它提供了数据的不同视图。投影集群是数据挖掘中的活跃主题。例如，诸如CLIQUE [2]和DOC [42]之类的算法可以发现有用的投影以及数据簇。 但是，在这里，我们仅关注出于聚类组合目的使用随机投影。



（对于Iris数据中每个可能的对象对，从关联值得出的距离与实际欧几里得距离x的关系。 计算了不同数量的超平面r = 1,2,3,4的协关联矩阵）

每个随机子空间的维数可能非常小，它本身就没有任何信息意义。另一方面，在一维空间中的聚类在计算上很便宜，并且可以通过k-means算法有效地执行。K-means算法的主要子程序（距离计算）在一维空间中的速度提高了d倍。 投影的成本相对于样本大小和维数是线性的，并且小于一k-means迭代的成本。

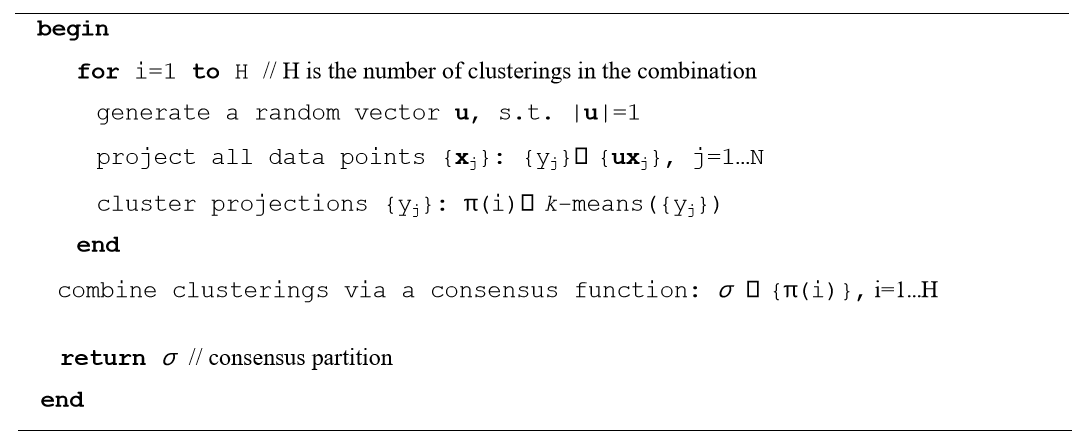
我们方法的主要思想是通过将数据投影到随机线上来生成多个分区。 快速简单的算法（例如k均值）将投影的数据聚类，结果分区成为组合的组成部分。 之后，将选定的共识函数应用于组件。 我们在实验部分讨论并比较了几种共识函数。



(在随机线上投影数据：（a）具有两个可识别的自然簇和一条随机选择进行投影的线的样本数据。 （b）数据投影到随机线上的点分布的直方图。)

如图4（a）所示，考虑一个简单的二维数据及其投影是很有启发性的。数据中有两个自然簇。在任何一维投影中，此数据看起来都相同，但是在投影子空间中的不同聚类中，点的实际分布是​​不同的。例如，图4（b）示出了该数据的一维投影中的点的一种可能的直方图分布。共有三种可识别的模式，每种模式都有来自两类之一的明显多数点。可以预期，通过k均值算法进行聚类将可靠地将至少一部分点与外圈聚类分开。不难想象，将图4（a）中的数据投影到另一条随机线上会导致点的分布不同和标签分配不同，但是对于此特定数据集，它将始终显示为三个钟形的混合体形的组件。这些模式很可能会通过k-means算法识别为聚类。因此，每个新的一维视图正确地有助于对一些数据点进行分组。最终积累多个视图应该会得到正确的组合聚类。

过以下过程描述了使用随机一维投影组合聚类的主要步骤：



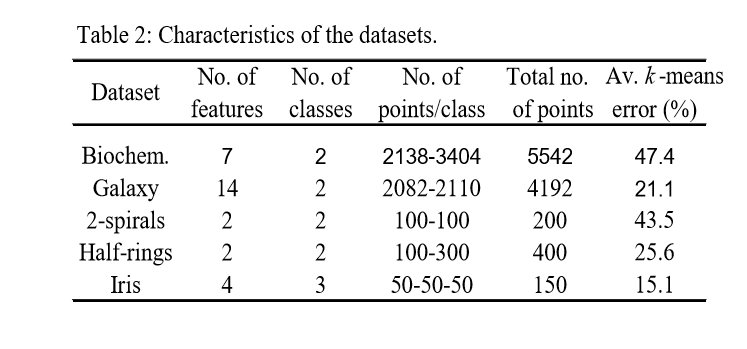
重要的参数是在每次迭代中由k-means算法返回的组件分区πi中的簇数，即k的值。 如果k的值太大，则分区将过度拟合数据集，这又可能导致协关联值不可靠。中的簇太少可能不足以捕获数据集的真实结构。另外，如果组合中的聚类数太少，则用于估计与协关联值的距离的有效样本大小也将不足，从而导致估计值的较大差异。这就是为什么基于协关联值的共识函数比基于超图算法的共识函数对组合中分区数（H值）更敏感的原因

# 实证研究

实验是在人工和真实世界的数据集（已知真实的自然簇）下进行的，以通过混合模型验证共识的准确性和稳健性。 我们使用五个不同的共识函数探索了数据集。

**6.1数据集**

表2汇总了数据集的详细信息。实验中使用了五个不同性质的数据集。[1]和[40]中分别描述了“Biochemical”和“ Galaxy”数据集。



我们通过匹配检测到的数据集和已知数据集来评估证据累积聚类算法的性能。群集的最佳匹配提供了一种性能度量，表示为错误分配率。为了确定聚类误差，需要解决已知聚类和衍生聚类的标签之间的对应问题。可以使用匈牙利方法(Hungarian method)针对k个簇的复杂度最小的权重二分匹配问题获得最佳对应关系。

**6.2参数和算法的选择**

已将QMI和EM共识算法的准确性与其他六个共识函数进行了比较:

1. CSPA用于根据协关联值对超图进行分区。它的复杂度为，这会导致严重的计算限制。我们没有将此算法应用于“ Galaxy” [40]和“Biochemical” [1]数据。出于相同的原因，我们没有使用其他协同关联方法，例如单链接群集。这些方法的性能已经在[14,15]中进行了分析。

2.用于超图分区的HGPA。

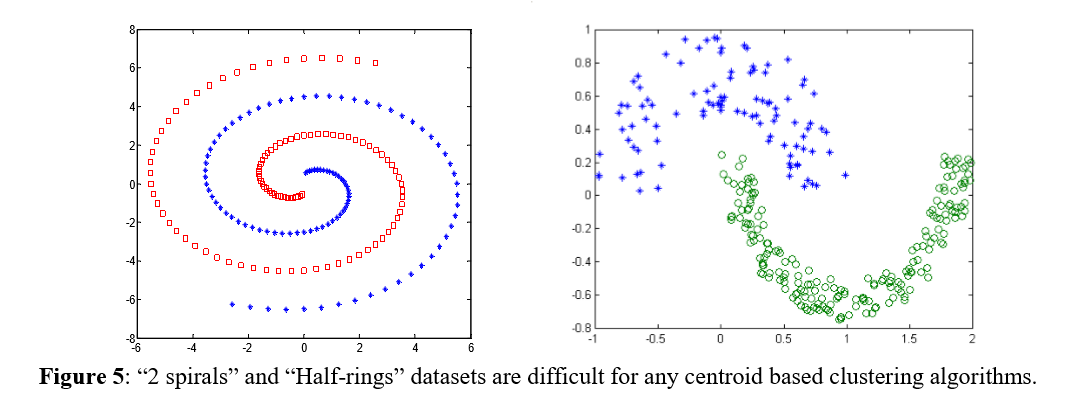
3. MCLA，通过扩展的超边缘操作集和其他启发式方法修改HGPA。

4.共识函数在协同关联矩阵上运行，但具有三种不同的层次聚类算法来获取最终分区，即单链接，平均链接和完全链接。

[47]中介绍了前三种方法（CSPA，HGPA和MCLA），其代码可从http://www.strehl.com获得。

k-means算法被用作为组合生成分区的方法。通过算法的随机初始化之后获得的解来确保分区的多样性。 聚类集成的以下参数特别重要：

i：H –组合聚类的数量。我们在[5..50]范围内更改了此值



ii：k –由k-means算法产生的组件聚类中的聚类数在[2..10]范围内

iii：r — 随机分裂算法用于获得聚类{π1，…，πH}的超平面数

EM和QMI算法都容易受到目标函数局部最小值的影响。为了降低收敛到较低质量解决方案的风险，我们使用了由这些算法的低计算复杂度提供的简单启发式算法。根据目标函数的值，从三个运行（随机初始化）的结果中选择了最终分区。似然函数的最大值用作EM算法的标准，而群集内方差是QMI算法的标准。

**6.3使用完整分区的实验**

由于空间限制，表3-7仅列出了每个数据集的主要结果。这些表格报告了相对较大的生化和天文数据集的10个独立运行和其他较小数据集的20个运行的聚类组合的平均错误率(%)

首先观察到，共识功能都不是绝对赢家。通过在参数k和H的值上使用不同的组合算法，可以实现良好的性能。对于较小尺寸的集合，EM算法略胜于其他算法，而当聚类数H> 20时，MCLA则更胜一筹。在实践中不太重要。 在簇数H <50时，所有协关联方法通常都是不可靠的，这就是我们放置提出的EM算法的位置。EM和QMI共识函数都需要至少估计kHM参数。 因此，当样本大小固定时，随着分区数量的增加，不可避免地会发生精度下降。 但是，在当前实验中，EM算法的准确性没有明显下降。由于提高了模型参数估计的可靠性，因此EM算法也应受益于大型数据集。

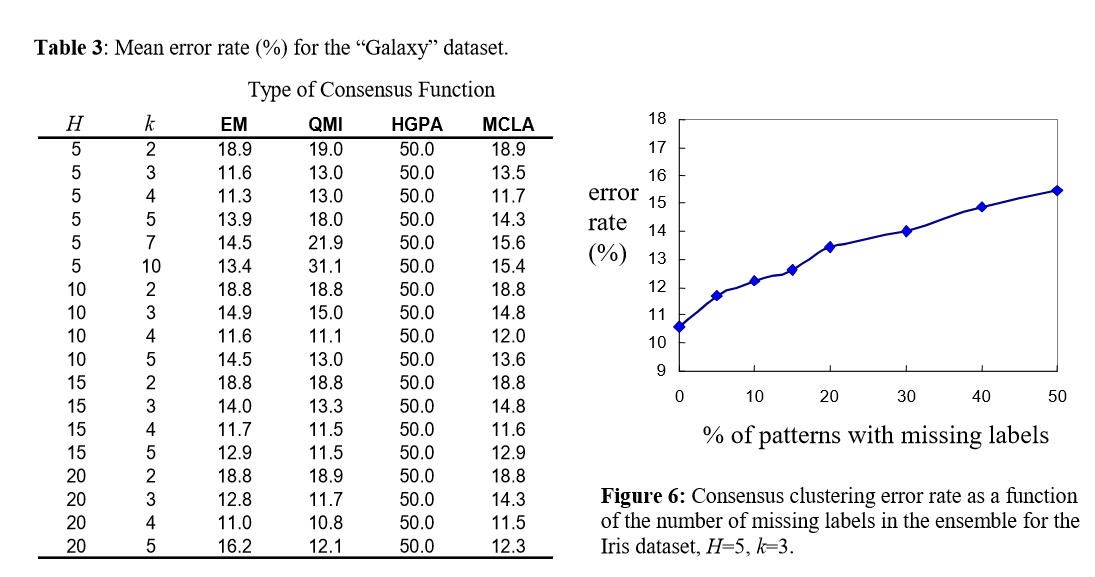
EM共识算法的一个有价值的特性是它的快速收敛速度。对于所有数据集，混合物模型参数估计几乎总是在少于10次迭代中收敛。此外，模式分配通常以4-6次迭代来解决。

聚类组合的准确性还取决于集合分区中的聚类数M，或更确切地说，取决于其与聚类的目标数之比，即k / M。例如，EM算法在Iris数据集的k = 3，对于“ Galaxy”数据集的k = 3.4，以及对于“ Halfrings”数据的k = 2的情况下效果最佳。这些k值等于或略大于组合分区中的簇数。相比之下，MCLA的准确度随着集成中簇数的增加而略有提高。图7显示了对于星系数据的不同共识函数而言，误差与k的关系。

有趣的是，正如预期的那样，当选择k等于真实簇数时，共识聚类的平均误差低于集合中k均值聚类的平均误差（表2）。此外，对于“Biochemistry”数据，通过EM和MCLA算法在k = 4时获得的聚类误差[1]与应用于该数据集的监督分类器发现的聚类误差相同[45]。

**6.4不完整分区的实验**

这组实验着重于聚类精度对缺少聚类标签的图案数量的依赖性。如前所述，使用k-means算法生成了一组分区。然后，我们随机删除每个分区中固定数量模式的群集标签。 EM共识算法用于此类合奏。每个分区中丢失标签的数量在样式总数的10％到50％之间变化。 表8列出了不同H和k值的“ Galaxy”和“Biochemistry”数据集的10次独立运行平均的主要结果。此外，图6中针对Iris数据显示了误差对具有缺失数据的图案数量的典型依赖性（H = 5，k = 3）

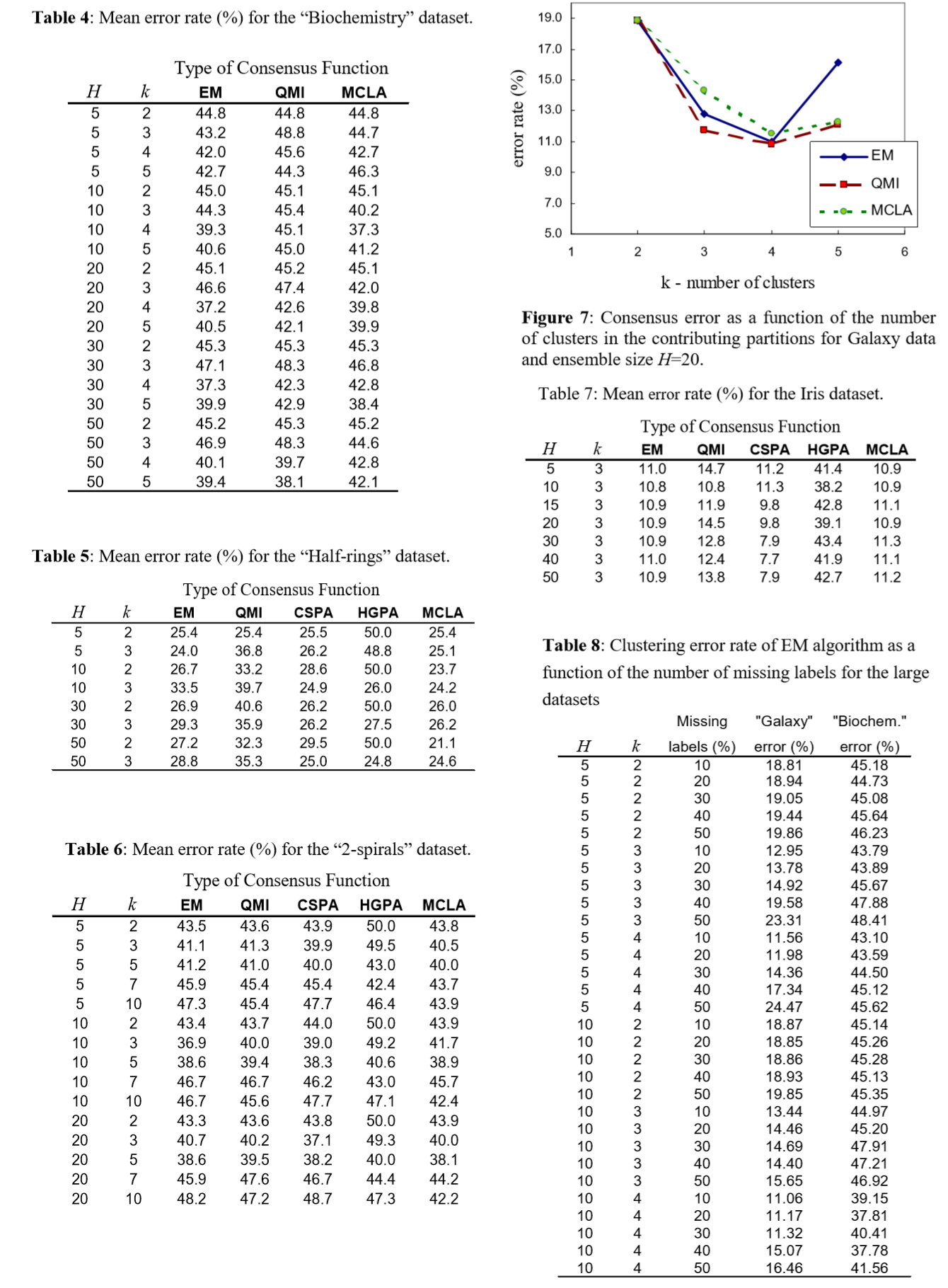


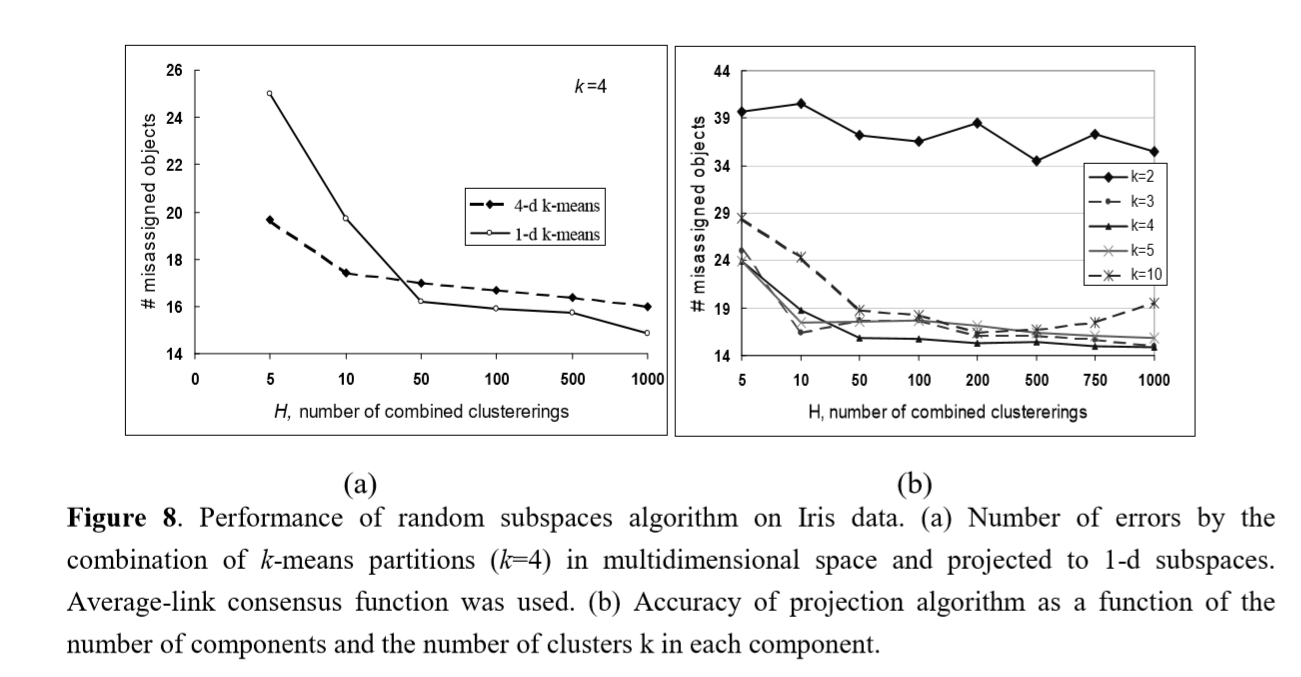
可以注意到，当多达50％的标记缺失时，对于“生物化学”数据，组合准确度只会显着降低。可以通过此数据的固有准确性低来解释这一点，从而几乎没有留下进一步退化的空间。对于“ Galaxy”数据，当k = 3,4时，精度下降近10％。但是，当仅丢失10-20％的簇标签时，则准确性仅有很小的变化。同样，使用不同的k值，我们看到结果对缺失标签的敏感性不同。 例如，在k = 2的情况下，精度仅下降1％以上。大小为H = 10的集合比缺少大小为H = 5的集合遭受的数据丢失更少。

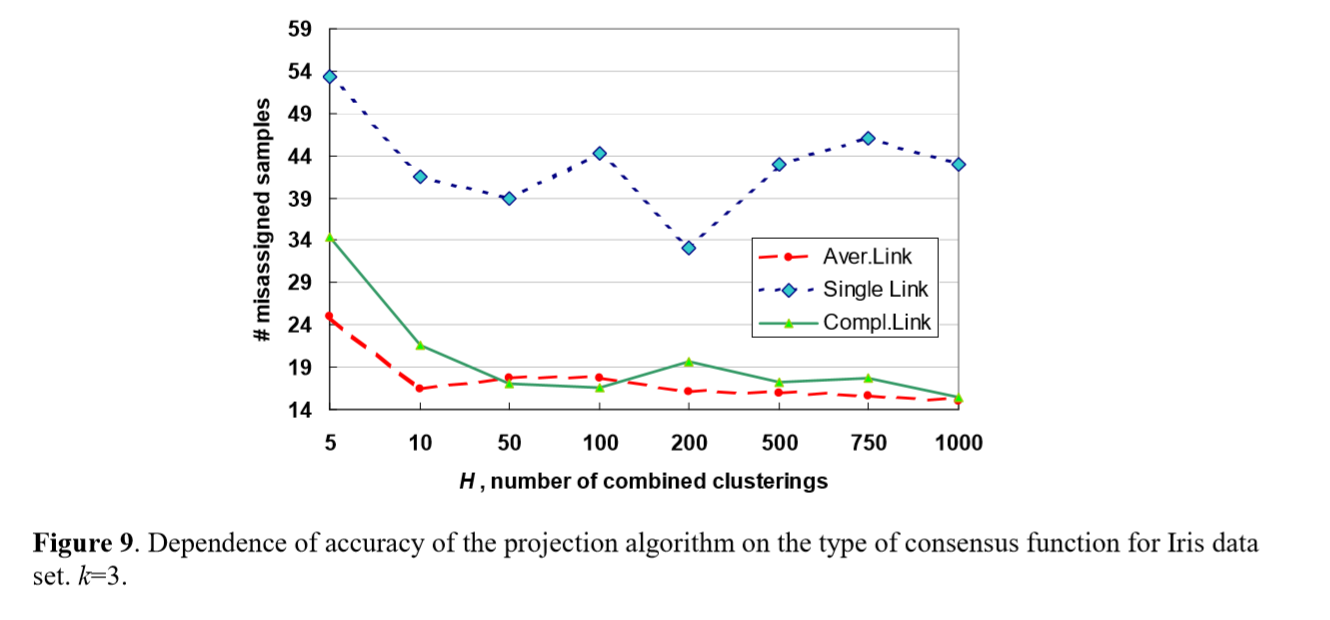
**6.5随机子空间算法的结果**

让我们首先说明投影一维子空间中聚类的组合如何胜过原始多维空间中聚类的组合。图8（a）显示了使用基于协关联值的平均链接共识函数对Iris数据和k = 4的学习动态。注意，每个分量中的簇数设置为k = 4，并且与实际簇数（= 3）不同。显然，在完整的多维空间中，每个单独的群集都比任何一个1维分区都强得多，因此，只有少量的分区（H <50）时，较弱的分区组合仍然无效。但是，对于大量的合并分区（H> 50），一维投影一起可以更好地揭示数据的真实结构。这是完全出乎意料的，因为k = 3的k-means算法平均在原始4维空间中产生19个错误，在1维随机子空间中产生25个错误。此外，在投影子空间中的聚类比多维空间中的聚类快d倍。虽然，在两种情况下计算共识分区σ的成本是相同的。

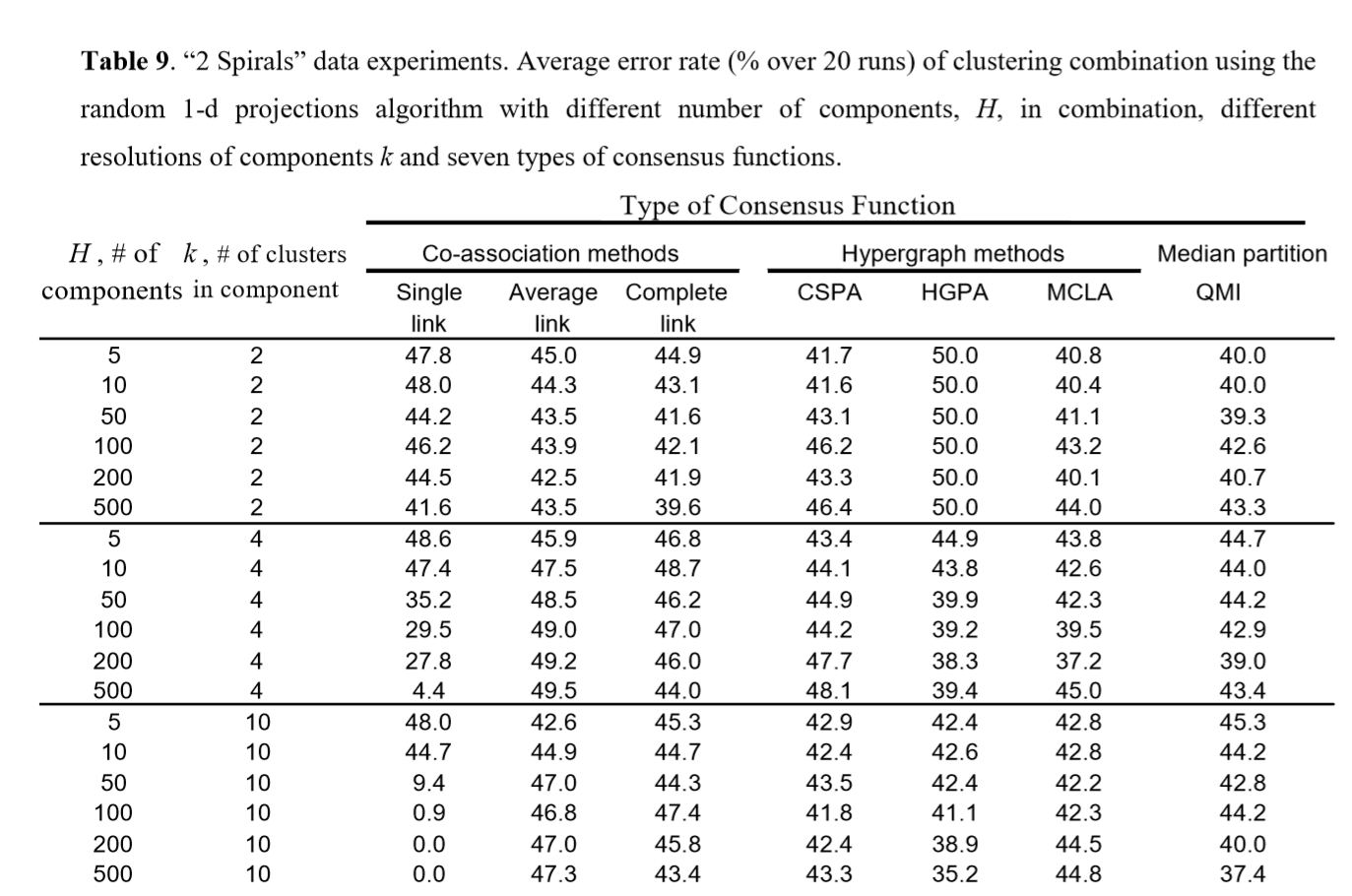
关于k值影响的结果报告在图8（b）中，该结果表明对于Iris数据集存在k的临界值。当关联距离的平均链接用作共识函数时，会发生这种情况。在这种情况下，值k = 2不足以分离真实的群集。共识函数的作用如图9所示。在Iris数据集上比较了三个共识函数。它们都使用了来自关联矩阵的相似性，但是使用三个不同的准则函数对对象进行聚类，即单链接，平均链接和完整链接。显然，使用单链接的组合的性能明显比其他两个共识功能差。这是可以预期的，因为虹膜数据中的三个类别具有超椭球形状。在图5中的“半环”和“ 2螺旋”数据集上获得了更多结果，这对于任何基于分区质心的算法传统上都是困难的。表9报告了使用7个不同的共识函数，分量分区的不同数量H = [5..500]和每个分量中不同的群集数量k = 2,4,10的“ 2个螺旋”数据的错误率。由于空间限制，我们在相同的实验条件和k的某些中间值的情况下，省略了“半环”数据集的相似结果。如我们所见，单链路共识功能表现最好，并且能够识别“半环”簇和螺旋形。与虹膜数据的结果相反，平均链接和完全链接共识不适用于这些数据集。





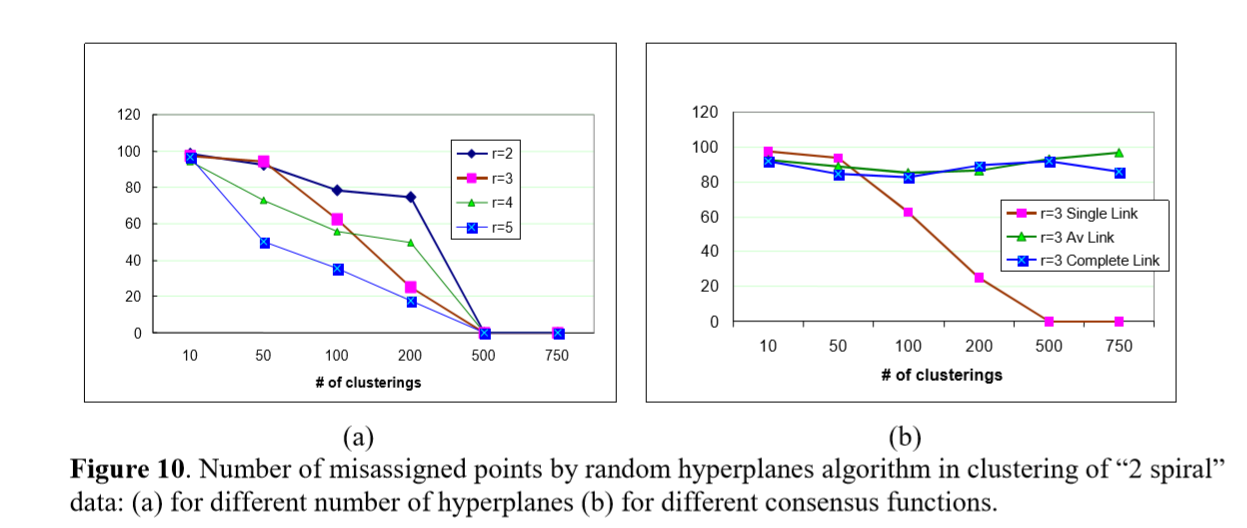


该观察结果支持这样的想法，即基于协关联值的共识函数的准确性对数据集的选择敏感。 通常，如果标准平均链接（单链接）聚集聚类适用于数据，则可以期望平均链接（单链接）共识是合适的，反之亦然。此外，这三个超图共识功能都找不到正确的组合分区。 鉴于超图算法在Iris数据上表现良好，这有点令人惊讶。但是，Iris数据的问题要少得多，因为其中一个类是线性可分离的，而其他类则很好地描述为两个多元正态分布的混合。



自然簇的完美分离是通过在簇组合中使用大量分区（H> 200）以及对于“半环”和“ 2个螺旋”的k> 3的值实现的。再次表明，对于每个问题，组件分区的分辨率都有一个关键值，可以保证良好的群集组合。 这进一步支持了Fred和Jain [15,16]的工作，他们的研究表明，每个分区中的随机簇数可确保组件的更大多样性。 我们看到，Iris数据的分辨率的最小要求值为k = 3，“ 半环”为k = 2，“ 2螺旋(2 spiral)”为k = 4。通常，k的值应大于群集的真实数目。

分区的数量会影响共识函数的相对性能。 H值较大（> 100）时，关联关联共识会更强，而H值较小时，最好使用超图算法或k均值中位数划分算法。



将组合的聚类精度与某些经典聚类算法的精度进行比较很有趣。例如，对于Iris数据，EM算法具有6.17％的最佳平均错误率。在我们的实验中，Iris数据的最佳性能是超图方法，其准确度低至3％，H> 200且k>5。对于“半环”数据，最佳标准结果是5.25％的误差通过平均链接算法，使用单链接协关联算法的组合聚类在H> 200时实现了0％的误差。此外，对于“ 2螺旋”数据，聚类组合也实现了0％的误差，与通过常规的单链接群集。因此，通过适当选择共识函数，聚类组合的性能优于许多标准聚类算法。但是，好的共识函数的选择与好的常规聚类算法的选择问题相似。猜测正确的共识函数的好选择也许只是运行所有可用的共识函数，然后根据等式（4）或等式（6）中的分区实用程序标准选择最终的共识分区。我们希望在该方法的未来应用中解决此问题。

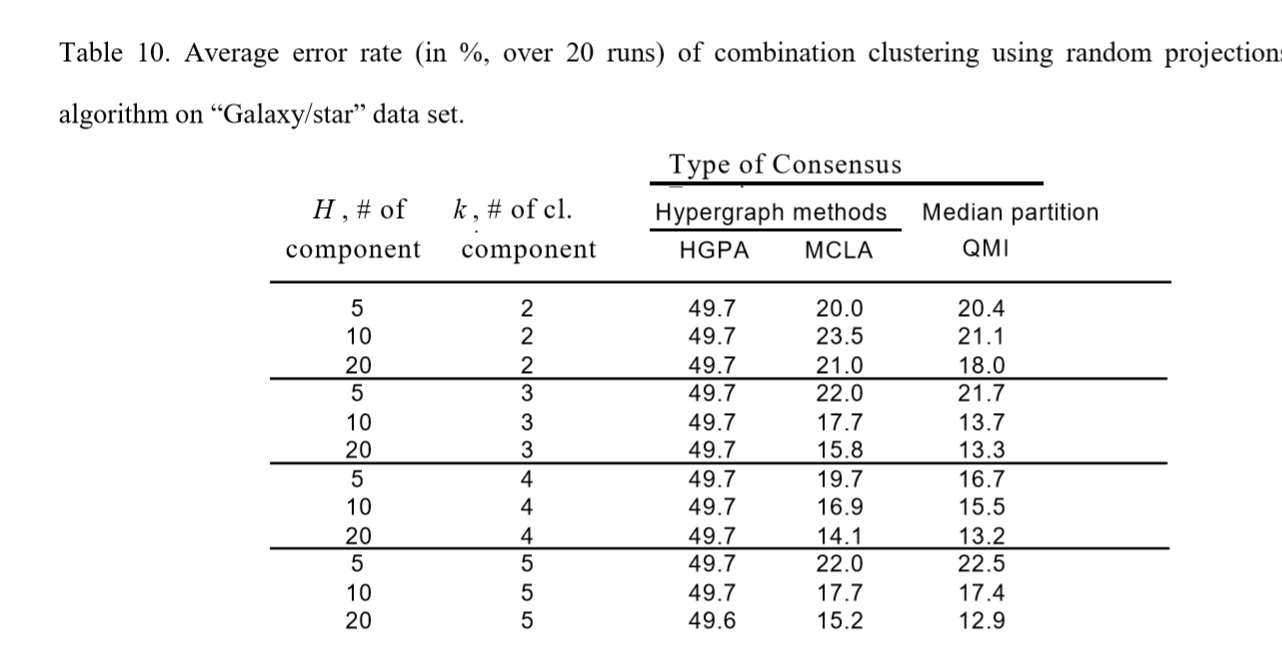
在“Galaxy”数据集上进行了另一组实验，该数据集具有大量样本N = 4192，特征数d =14。任务是将星系模式与恒星分离。我们从原始数据中使用了最困难的“模糊”对象集[40]。对象的真实标签是由专家手动提供的。即使由于投影而使组件分区的计算速度加快了d倍，但总体计算量仍可以由计算共识分区的复杂性来控制。由于建立N个对象的关联矩阵的复杂度，二次计算复杂度有效地禁止了基于协关联的共识函数在大数据集上使用。因此，对于大型数据集，我们不会使用Strehl和Ghosh的三种分层凝聚方法以及CSPA超图算法。就速度而言，通过QMI进行中位数划分的k均值算法最具吸引力，其复杂度为。另外，我们还使用了其他两个基于超图的共识函数，因为它们在实践中工作很快。表10报告了使用这些共识函数实现的错误率。由于数据量大，我们还将组合中的组件数限制为H = 20。结果表明，k-means算法用于中值划分具有最佳性能。 HGPA由于偏向于平衡的群集大小而无法很好地工作，“半环”数据集也是如此。我们看到，当分区和群集的数量增加时，准确性会提高。

重要的是要注意，“ Galaxy”数据的标准k均值算法的平均错误率约为20％，而最著名的解决方案的错误率为18.6％。 值得注意的是，k均值中位数分割算法和MCLA获得了更好的分割，并且k> 3时的错误率仅为13％左右。

**6.6随机分裂算法的结果**

与5.1节中一样，通过随机超平面的分裂通过聚类组合进行了同一组实验。在这里，我们只想强调最有趣的观察结果，因为许多细节上的结果都接近于使用随机子空间获得的结果。 就绝对性能而言，存在一点差异：随机超平面算法在使用单链接共识函数的“半环”数据上略胜一筹，在“两个螺旋”上近似相同，而在虹膜数据集上则较差。

区分组件分区中的群集数k和超平面r的数量很重要，因为超平面随机相交并形成变化的群集数。 例如，一条平面上的3条线可以在4到7个不同的区域之间创建任何地方。



“ 2螺旋”数据集的结果还表明，当H达到500且r = 2，…，5时，共识聚类收敛到理想解。 参见图10（a）。大量的超平面（r = 5）改善了收敛性。图10（b）说明了共识函数的选择对于成功的聚类至关重要。对于“ 2-螺旋”数据，只有单链路共识能够找到正确的聚类。

# 结论与未来工作

这项研究从多个方面扩展了先前对聚类集成的研究。首先，我们为多个聚类引入了统一表示，并提出了相应的分类聚类问题。其次，我们提出了解决聚类组合问题的方法。共识聚类是从分区集合的有限混合模型的最大似然问题的解中得出的。将Ensemble建模为聚类标签空间中多元多项式分布的混合。使用EM算法可有效解决最大似然问题。基于EM的共识功能还能够处理不完整的贡献分区。第三，证明了使用通用的互信息定义，另一个共识函数可以与经典的类内方差准则相关。在特殊变换的标签空间中，可以通过k-means算法高效地找到基于二次互信息的共识解决方案。实验结果表明，该方法对于多个数据集具有良好的性能，并且可以与其他共识函数进行良好的比较。这些算法的优势包括计算复杂度低和基础牢固的统计模型。

我们还考虑过结合使用数据投影和随机数据分割的弱聚类算法。 提供了一个简单的解释模型来说明这种弱组件的组合行为。 我们已经分析了组合精度作为参数的函数，该参数控制组件划分的功能和分辨率以及学习动态与所涉及的聚类数量。 实证研究比较了应用于弱分区的几种共识函数的有效性。

继续在多个方向上研究聚类组合很有趣：

1）设计有效的启发式搜索以实现共识函数；

2）更精确和可量化的弱聚类概念；

3）更好地理解组件分辨率对整体的影响 性能。可以扩展此研究，以考虑集合中分区的非独立性。这里介绍的共识函数等效于某种潜在类分析，它提供了已建立的统计方法来度量和使用变量之间的依赖性（至少成对）。考虑不同质量的分区的组合也很有趣。在这种情况下，需要开发一种共识函数，根据不同分区的强度对不同分区的贡献进行加权。我们希望在未来的工作中解决这些问题。