# 基于样本稳定性的聚类集成

**摘要：**聚类集成的目的是根据一组聚类结果找到数据的底层结构。 已经观察到，样本可以在不同聚类结果中的聚类之间改变。这种变化表明样品可能对基础结构的检测有不同的贡献。但是，现有的聚类集成方法均等地对待所有样本。为了解决这一缺陷，我们介绍了样本的稳定性以量化其贡献，并提出了确定这种稳定性的方法。我们提出了两个与该方法一致的公式来计算样品的稳定性。然后，我们根据样本的稳定性开发一种聚类集成算法。无论采用哪种公式，该算法都会将数据集分为两类：聚类核和聚类环。借助聚类核和聚类环，然后，所提出的算法使用聚类核中的样本发现清晰的结构，并将聚类环中的样本逐渐分配给该清晰结构。在八个合成数据集上进行的实验说明了该算法的工作原理。在来自UCI的十个真实数据集和六个文档数据集上，该算法在统计上优于十二种最新的聚类集成算法。对图像分割情况的实验分析表明，通过稳定性发现的聚类核是合理的。

## 1.简介：

数据聚类[1]是机器学习中最重要的领域之一。它试图发现数据集的基础结构。通常，该结构意味着将相似的样本分配给相同的集合，而将不相似的样本分配给不同的集合。尽管已有文献提出了许多聚类算法，但是缺乏先验知识使得聚类分析仍然是一个非常具有挑战性的问题。公认的是，单个聚类算法无法有效处理所有类型的数据分发。每种聚类算法都有其自己的策略，可以从数据集中发现结构。不同的算法或算法的不同参数可能导致不同的聚类结果。没有监督信息，很难判断哪种结构最符合实际分布。因此，选择合适的算法是一项艰巨的任务。为了避免这个任务，许多研究集中在集成多个聚类结果上，这被称为聚类集成[2]。与单个聚类算法相比，聚类集成可以显着提高聚类解决方案的鲁棒性，稳定性和质量。聚类集成技术已被有效地用于处理许多聚类任务，例如分类数据[3,4]，高维数据[5]，噪声数据[6]，时间数据[7]，特征选择[8]等。

Strehl和Ghosh [9]首次提出了聚类集成问题。在[9]中，聚类集成问题被描述为在不访问原始特征的情况下组合了一组对象的多个聚类结果。聚类集成方法应该能够将多个聚类结果组合到一个与基本聚类最相似的一致分区中，而无需调用原始数据集。一些研究人员使用多个聚类结果和原始特征作为输入，以进一步提高聚类性能[10,11]。与不使用原始特征作为输入的方法相比，可以在更多领域应用不访问原始特征的聚类集成。字段主要包含原始功能不可用的场景，例如由于某些保密原因而导致的分布式数据源或属性以及未公开的数据[12,13]。另外，可以将大多数不调用原始数据集的聚类集成算法扩展为调用原始数据集的版本。

在本文中，我们关注于传统的聚类集成问题，该问题无法访问原始特征。有了这个要求，已经提出了许多聚类集成方法。这些方法利用了很多技术，包括但不限于：聚类技术[14,3,15–17,1,6]，图技术[18,9,19]和优化方法[20-24]。另外，加权聚类集成和选择性聚类集成是两种进一步提高聚类集成算法性能的方法。这两种方法增加了基本分区的效果，这对于后续集成是有益的。关于加权聚类集成（WCE）和选择性聚类集成（SCE）的研究主要集中在探索有益基础聚类结果的特征。特征主要涉及两个问题，即基本分区的多样性和准确性。关于WCE和SCE的先前研究探讨了多样性还是准确性是选择基本聚类的决定因素。最近，更多的研究人员倾向于在选择基本分区的子集时将多样性和准确性结合起来。尽管上述工作在解决聚类集成问题上取得了良好的性能，但仍有很大的空间可以提高集成质量。现有的集成算法在基础结构的构造中均等地对待每个样本。但是，给定一组聚类结果，样本对基础结构的检测可能有不同的贡献。在聚类分析中，一个聚类具有一个聚类核心（表示稳健的分配）和一个聚类环（可以视为噪声）[25]。在[26]中，作者介绍了轮廓和图案。给定一组聚类结果，配置文件会计算样本在一个聚类中出现的相对频率，而基序是在同一聚类中具有高度一致性的样本子集。但是，要发现主题，就应该处理簇对应发现和簇融合问题，但尚未很好地解决[27]。另外，在文献[28-31]中已经指出，基础聚类的多样性和准确性对于整体表现很重要。对于一组基本聚类结果，多样性主要来自对聚类环中样本的不同决定，并且通过核心样本的一致分配来保证准确性。区分聚类核中的样本和聚类环中的样本的集成算法可能会产生更好的解决方案。根据以上讨论，出现了一个有趣的问题：如何在聚类集成问题中有效地发现聚类核和聚类环？

为了回答这个问题，本文介绍了样本在聚类集成问题中的稳定性。样本的稳定性反映了其在基本聚类结果中改变其聚类的趋势。在这种稳定性下，聚类核应由具有高稳定性的样本组成，而聚类环应包含具有低稳定性的样本。对于这样的聚类核和聚类环，应该采用不同的策略来处理聚类核和聚类环中的样本。

聚类核中的样本具有来自大多数基础集群的一致分配。这些样品代表易于发现的清晰结构。聚类核样本的策略侧重于发现分配的预结构。对于聚类环中的样本，在基本聚类中存在许多分歧。因此，仅基于基础聚类很难为环中的样本生成一致的分配。发现的预结构将为环样本的分配提供指导性信息。现在，聚类环样本的策略只需要正确地将环样本分配到预结构中即可。基于预结构的光环中样品的分配可能比以前更准确。考虑到上述因素，本文提出了一种新的聚类集成算法，该算法集成了从聚类核到聚类环的基础聚类。

简而言之，论文的贡献如下：

提出了一种确定样本在聚类集成问题中的稳定性的方法。 开发了符合该方法的两种公式。

提出了一种基于样本稳定性的聚类集成方法，该方法可以区分稳定样本和不稳定样本。 该算法根据稳定样本（称为聚类核）发现结构，然后将不稳定样本（称为聚类环）逐渐分配给预先发现的结构。

在图像分割的情况下，通过实验验证了所提方法测量的样品稳定性的合理性。 使用八个综合数据集来说明所提出的聚类集成算法的工作机制。 此外，对基准数据集进行了实验，以证明该算法的有效性。

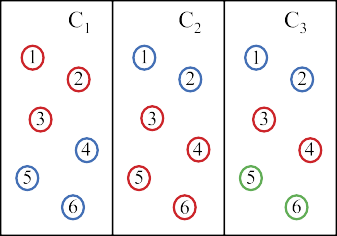
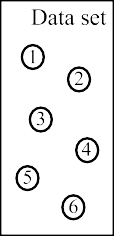


图1.基本集群的示例

本文的其余部分安排如下。在第2节中对聚类集成问题进行了概述。在第3节中，我们介绍了一种测量样本稳定性的方法，并根据该方法提出了两个公式。在第4节中，将详细介绍根据样本的稳定性提出的聚类集成方法。在第5节中，进行了实验以展示该算法如何工作以及所提出算法的有效性。最后，第6节总结了论文。

## 2.聚类集成

总的来说，关于聚类集成问题的研究主要包括三个方面：集成生成，集成选择和集成。 假设X = {x1，x2，...，xn}是具有n个样本的数据集。聚类后多种不同的聚类方法，将获得一组聚类结果= {C1，C2，...，CL}，其中L是整体大小并指示聚类数。 Cl（xi）表示聚类结果Cl诱导的xi的标记。聚类集成的目标是找到一个新的聚类结果C \*，它与中的每个元素相似。 一般而言，集成选择部分和集成部分仅基于一组聚类结果，而无需调用原始数据集。

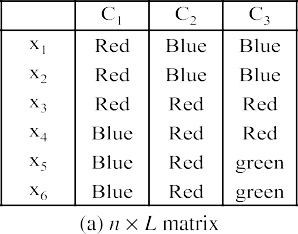
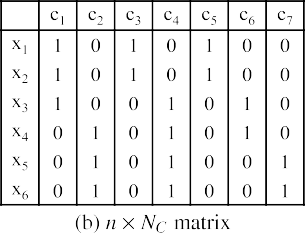
聚类集成的第一个任务是获得一组基本聚类结果。 图1给出了三个简单示例5个样本的聚类结果= {C1，C2，C3}。

众所周知，基本聚类的多样性和准确性是整体性能的两个关键因素技术。 为了生成更多样化和可接受的准确集合，建议使用以下策略：

* 不同的参数设置。 基本聚类结果可以通过利用具有随机性的聚类算法（例如k均值类型算法的初始中心）来生成。 另一个重要参数是簇数。 在文献中建议将聚类数目设置为大于预期的聚类数目。
* 不同的聚类算法。 关于如何发现数据集的基础结构，每种聚类算法都有其自己的特定观点。 不同的聚类算法通常会产生不同的结果。 因此，可以使用多种聚类算法来生成各种基础聚类结果。
* 特征的不同表示。 特征的表示形式主要由两种形式组成，一种是数据投影，另一种是特征的子集。 对于多维数据集，两种形式的要素表示都尝试从不同的视图描述数据集。 因此，当利用数据特征的多个表示时，将获得一组多样化的聚类结果。
* 弱聚类。 已经证明，整合多个弱聚类可以产生良好的整体解决方案。 通过在数据集的随机一维投影上运行聚类算法或通过将数据集划分为随机超平面，可以实现弱聚类，而弱聚类的效果优于随机聚类。 一组基本的弱聚类应该包含更高的多样性。

受特征选择技术成功提高机器学习算法性能的启发，在聚类集成中提出了集成选择技术，以提高获得的聚类结果集的质量。 集成选择方法基于预定义的原则选择基本聚类结果的子集，该原则被认为对后续的集成步骤有益。 关于集成选择原理的研究主要涉及三个问题，分别列出如下：

* 基于多样性的集成选择。 聚类结果的多样性通常通过其与集合中其他聚类结果的平均差异来衡量。 多样性原则是基于这样一个事实，即低多样性限制了整体集成的提高，首要的聚类选择集成方法主要利用多样性来选择聚类结果的子集。
* 根据准确性进行集成选择。 聚类结果的准确性通常通过其与集成中其他聚类结果的平均相似度来衡量。 准确性是集成选择阶段的重要元素。 在中已经说明，准确性与聚类集成性能呈正相关。

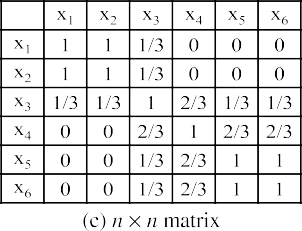
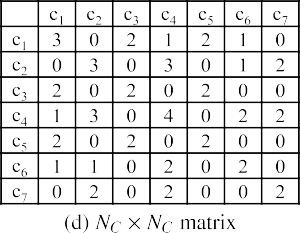
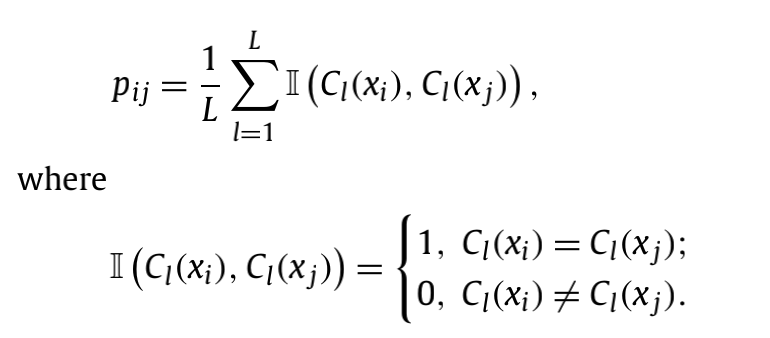
 

图2.四个信息矩阵

基于多样性和准确性的集合选择方法。 多样性和准确性是两个相反的目标，它们对于集成性能都很重要。 最近提出的大多数集成选择方法都同时考虑了两个目标。 这些方法使用结合了多样性和准确性的度量作为选择准则，或者设计一个同时考虑多样性和准确性的复杂过程。

获得了一组基本的聚类结果后，最后一步是生成一个综合的聚类结果。 为了从基本聚类结果中发现结构C ∗，提出了许多聚类集成方法。集成方法所依赖的不同类型的信息矩阵[45]，现有方法可以分类如下：

* 基于特征的方法（n×L矩阵，如图2（a）所示），基本聚类结果集实际上是一个n×L矩阵，其中每一列表示一个聚类结果。 基于n×L矩阵，提出三种类型的方法。 第一种是寻找聚类结果之间的对应关系，然后转换到集成分类器里面。 第二类是将聚类集成视为聚类分类数据。 而且，许多中值分割方法也基于该矩阵。 这些方法尝试发现与基于优化技术的基本聚类结果集相似的聚类结果。
* 基于聚类的方法（n×NC矩阵，如图2（b）所示，如果我们使用每一列来表示集群，那么将生成一个n×NC矩阵，其中NC = k1 + k2 + ... + kL，ki是聚类的簇数结果Ci。 该矩阵是二进制且稀疏的。 要基于此n×NC矩阵生成共识结果，一般方法是正在寻找聚类之间的关系，丰富n×NC矩阵，并利用现有的聚类方法或图分区方法生成最终结果。
* 基于样本协关联的方法（n×n矩阵，如图2（c）所示）。 给定一组聚类结果= {C1，C2，... CL}，两个样本xi和x j出现在同一簇中的频率由下式计算：



所有成对的频率将形成一个n×n矩阵，称为协关联矩阵。 该矩阵反映了每对样本之间的关系。 它提供了一个近似相似性矩阵，可以供分层类型聚类算法使用。 此外，此矩阵是许多基于图的聚类集成的基础方法。

//（以a矩阵为参考矩阵）L=3，Cl(xi)=Cl(xj)=1是因为两个相同的样本始终在同一个簇，所以两个相同的样本在同一个簇的概率为1，Cl(xi)≠Cl(xj)的意思是xi和xj不在同一个簇中，举例说明，为什么x4行x3列的数字时2/3呢？一共有3个簇，x4和x3同时出现在C2和C3簇中，所以就是2/3，计算过程：（1/3\*（0+1+1）=2/3）再举例，为什么x2行x6列的数字时0呢？因为x2和x6从来都没有出现在同一个簇中过，所以就是（1/3\*（0+0+0）=0）或直接计算0/3=0，再举例，为什么x5行x3列的数字时1/3呢？因为x5和x3只在C2簇中同时出现过，所以概率为1/3（1/3\*（0+1+0））。

* 基于聚类相交的方法（NC×NC矩阵，如图2（d）所示）。 根据每对聚类也可以访问之间的关系，例如常见样本。 它们形成一个NC×NC矩阵。 该矩阵反映不同聚类中聚类之间的关系，有助于找到聚类之间的对应关系。

//（以b矩阵为参考矩阵）这个方法的意思是，每个聚类之间有几个相同的样本，或者说聚类相交时是几个样本在相交，举例说明，c1行c1列为3，因为c1簇和c1簇相交时相同的样本（b矩阵中c1和c1有几个相同的1），分别为x1,x2,x3;

c1行c2列为0，因为c1簇和c2簇相交时没有相同的样本，故为0；

c1行c3列为2，因为c1簇和c3簇相交时相同的样本（b矩阵中c1和c1有几个相同的1），分别为x1,x2;（实际意义就是x1,x2同时属于c1簇和c3簇）

c1行c4列为1，因为c1簇和c4簇相交时相同的样本（b矩阵中c1和c1有几个相同的1），为x3;

c1行c5列为3，因为c1簇和c5簇相交时相同的样本（b矩阵中c1和c1有几个相同的1），分别为x1,x2;

c1行c6列为1，因为c1簇和c6簇相交时相同的样本（b矩阵中c1和c1有几个相同的1），为x3;

c1行c7列为0，因为c1簇和c7簇相交时没有相同的样本，故为0；

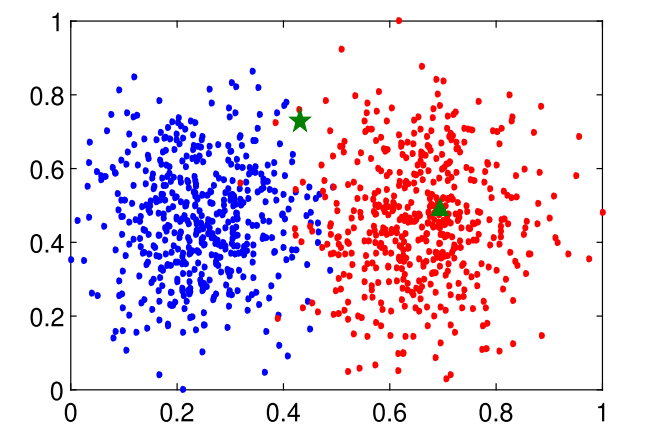


图3. 2k2d数据集和两个标记的样本。

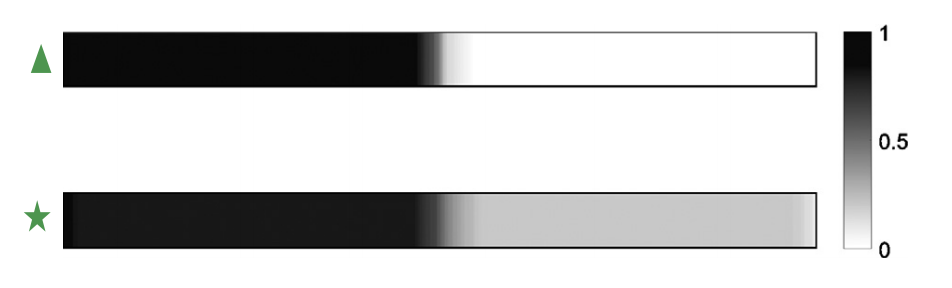


图4.标记的样本与其他样本之间的关系。

上面关于集成集成方法的讨论几乎涉及所有形式的信息矩阵，这些信息矩阵可以由构造。 大多数现有的集成方法都将所有样本视为一个整体，以发现数据集的结构。 很自然地得出结论，不同的样本对底层结构可能具有不同的表征能力。 因此，使用不同策略来处理聚类核和聚类环中的样本可能会产生更有效的结果。 为了设计这样的聚类集成算法，聚类核应该基于找到。 为了有效解决此问题，下面引入样本的稳定性来衡量量样本属于聚类核的程度。

## 3.样本在聚类集成中的稳定性

给定一组基本的聚类结果，一些样本在一个组中始终保持一致，而另一些样本则经常一组变化到另一组。 这种现象具有样本改变其组趋势的特征。 这个趋势对于聚类集成中的许多任务很有帮助，例如测量基本聚类结果集的质量并区分簇核和簇环。 为了量化样本的群体变化趋势，我们引入了测量值称为样品稳定性。

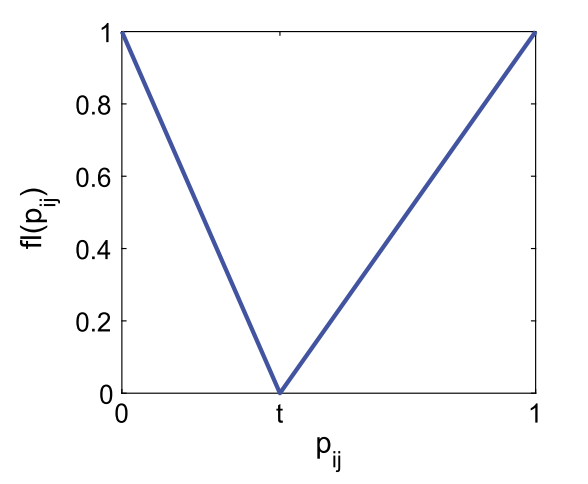
让我们首先考虑一个简单的可见的人工数据集2d2k，它具有1000个样本和2个高斯分布簇。图3示出了2d2k的分布。 在图3中，选择了两个样本，分别标记为三角形和星。 在这里，我们在2d2k上运行50次随机一维k-均值（k = 2），以获得不同的基本聚类。随机一维k均值将数据集投影到随机线上，并使用k均值生成聚类结果。 至于样品，很容易

以获得与其他群集在同一群集中出现的频率。 在图4中，我们使用灰度图像分别显示三角形样本和星形样本与其他样本的频率按降序排列，其中黑色表示关联频率为1，白色表示关联频率为0。（//关联频率为1代表这个样本始终与三角形样本在同一个簇中，关联频率为0代表这个样本从来没有与三角形样本在同一个簇中，类似于（C）矩阵的计算）从图4可以很容易地发现星形样本有许多灰色区域，表示样本在各簇之间的变化。对于三角形样本，有许多黑色区域和白色区域，表示样本分配一致。

从示例中可以看出，可以通过样本与其他样本之间的关系来测量样本的稳定性（//（C矩阵中的数字））。对于一个样本，如果与其他样本的关联频率相对较高或相对较低，则可以将其视为样品稳定（//（C矩阵中的数字越接近于1或者越接近于0）），否则稳定性低。 在本节中，我们提出一种量化这种稳定性的方法。

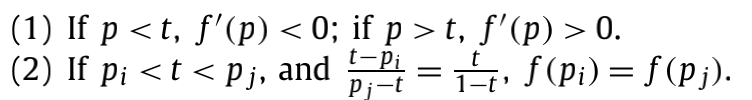
### 3.1 测量样品稳定性的方法

根据以上分析，应首先测量两个样品之间关系的稳定性。 考虑到如果两个样本xi和x j被所有聚类结果分配到同一组，则两个样本的协关联频率即pij = 1，xi和x j之间的关系是确定的（//（C矩阵xi行xj列中的数字为1））。此外，如果所有聚类，则xi和x j具有一定的关系结果一致地将它们分配给不同的组，即pij =0。如果xi和x j之间的关系不稳定，关联频率值在（0，1）中。 可以基于协关联的分布来学习阈值t表示最不稳定的频率。 根据以上讨论，基于一组聚类结果，两个样本之间的稳定关系包含两个方面：（1）大多数聚类结果将它们分配给相同的簇; （2）大多数聚类结果将它们分配到不同的聚类中。 稳定水平直接由协关联频率不能反映第二个方面。 因此，我们设计了确定性函数f来投影协关联频率进入关系稳定空间。 为了衡量两个样本之间关系的确定性，函数f在两端应具有较高的值（//在两端越稳定值越高），但在t处应具有较低的值。 通过以上讨论，确定性函数f的定义如下：

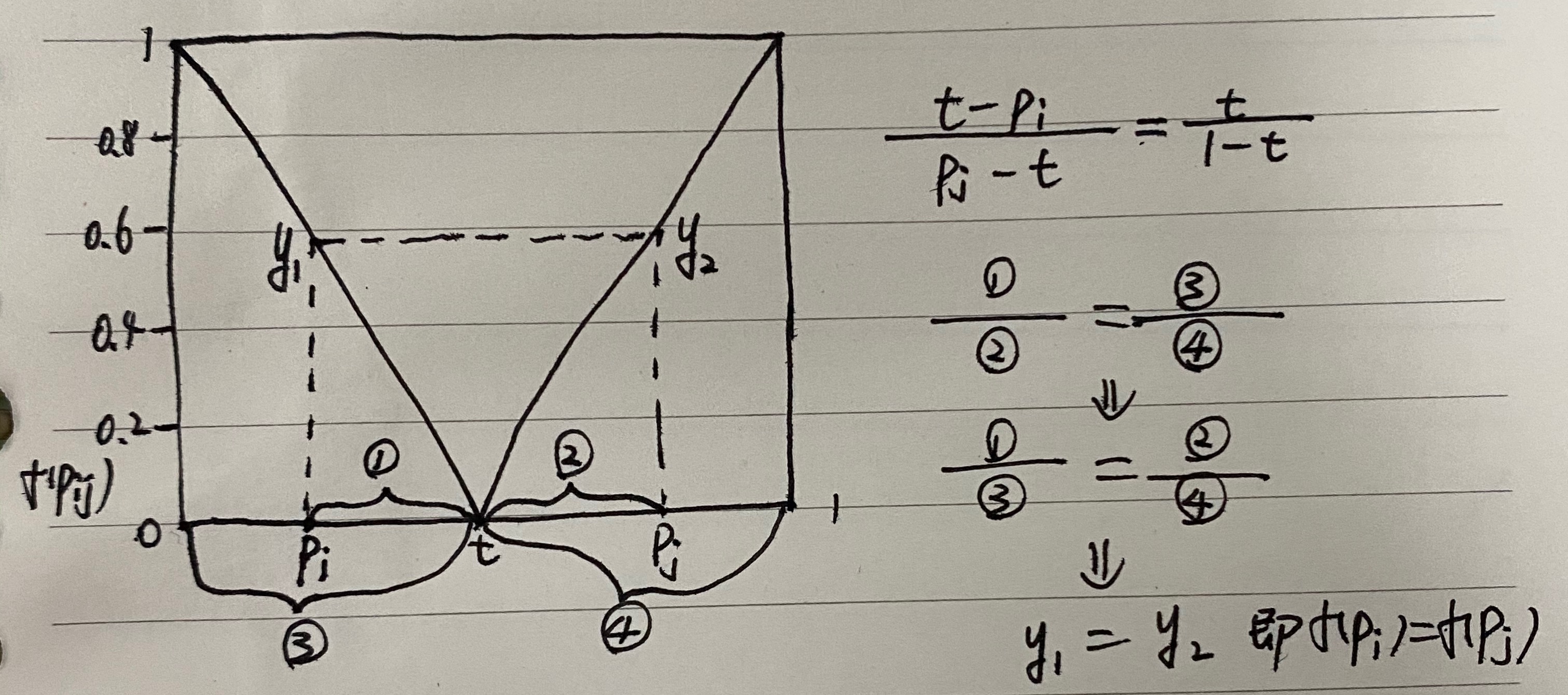
图5 fl曲线

//观察图5，图的两端函数f在两端应具有较高的值（//在两端两个样本之间关系的确定性越高），但在t处应具有较低的值。

定义1.（确定性函数）：如果对于参数p∈[0，1]且参数t∈（0，1），（//p是协关联频率即C矩阵中的数字，t是表示样本稳定性的阈值）则f是确定性函数，它满足：



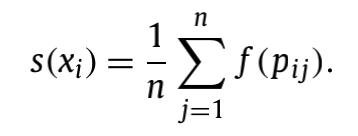
//先看（1），意思是参照图5，如果p值<t值，即协关联频率小于衡量样本稳定性的阈值，则斜率<0，参考图5，t左半边的曲线斜率是<0的，同理可得p值>t值时，斜率是大于0的。对于（2）的理解如下图



第一个条件要求确定性函数应在t上获得最小值，并获得高值离t很远 第二个条件要求确定性函数对于协关联应该是无偏的t两侧的值。 具体来说，如果两个协关联值pi和p j位于t的同一侧，则关联值越远，确定性值越高。形式上，该特性表示为（pi − t）（p j − t）> 0（//说明位于t的同一侧）且| pi − t | > | p j − t |（//绝对值越大表示和t的距离越大，离t越远），则f（pi）> f（p j）（//在t的同一侧离t越远表示两个样本之间关系的确定性越高）。基于第一条件，容易获得该特性。另外，位于t的不同侧的两个关联值具有以下关系：如果pi <t <p j且（1-t）（t-pi）> t（p j-t），则f（pi）> f（p j）。 该关系直接反映了基于两个条件成立的确定性函数。 以上讨论表明，确定性函数可以反映两个样本之间的稳定关系的两个方面。

在聚类集成中，对于一个样本，如果其与其他样本的关系稳定，则此样本应具有较高的稳定性。因此，利用确定性函数f，可以从具有n个样本的数据库X中获得样本xi的稳定性。

通过关于其与其他样本的关系的决策的平均协议来量化。然后，基于确定性函数f，xi的稳定性，s（xi）由下式计算：

  **（2）**

//例如计算s(x1)=1/6\*[f(pi1)+f(pi2)+f(pi3)+f(pi4)+f(pi5)+f(pi6)]=1/6\*[1+1+1/3+0+0+0]=7/18

以上f(pij)的值参考C矩阵中的值

基于这种方法，我们提出了两种稳定性测度：基于线性函数的稳定性和基于二次函数的稳定性。

### 3.2 基于线性函数的稳定性

通过定义1满足确定性函数的条件，可以将线性函数设计为

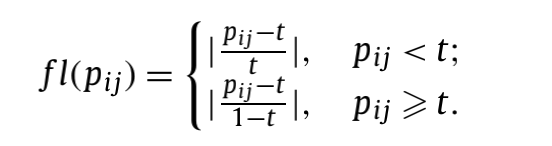
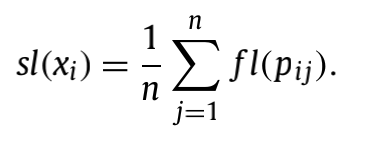
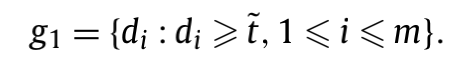
**（3）**

图5示出了f l（pij）的曲线。样本xi的基于线性的稳定性可以被基于该样本与其他样本的关联程度量化，可以如下计算：

 **（4）**

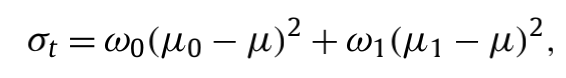
为了学习阈值t，在本文中，我们使用Otsu算法[53]解决了线性判别问题。 假设D = {d1，d1，...，dm}是一个包含m个元素的向量。 阈值t可以通过以下方式将D分为两组g0和g1：（如何分组？）

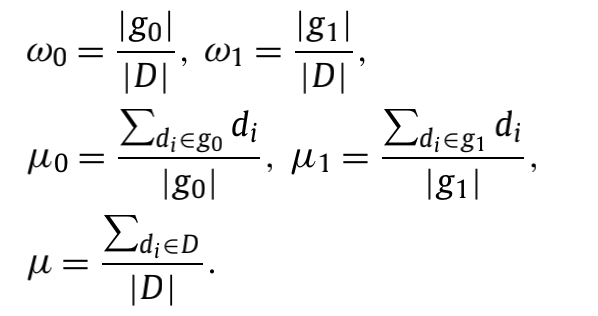
 **（5）**

 **（6）**



以下任务是评估每个阈值的性能并选择最合适的一个。 利用了判别标准，该判别标准测量类之间的方差。 对于两个类别g0和g1，类别间方差的计算依据为：（类别间方差反映什么）

**（7）**



在此场景中，最合适的阈值t \*应使公式（7）最大化，即：

**（8）**

公式（8）可用算法1求解。提出了Otsu算法，从灰度图像中选择阈值。 在[53]，作者指出，Otsu算法也可用于选择场景的直方图阈值提供了一些用于对对象进行分类的区分特征。 事实上，Otsu算法仍然是一种到现在为止大多数参考的阈值学习方法[54]。

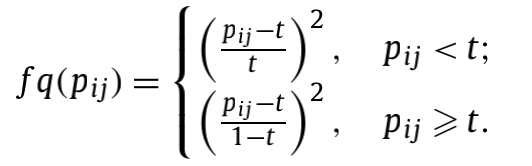
以协关联矩阵作为Otsu的输入，我们可以学习阈值t。 然后，我们计算出样本的稳定度，直观地讲，不稳定样本应该是第3个样本，最稳定的样本应该是第1个样本和第二个样本。 根据公式（4），图1中每个样本的稳定性由下式给出：



稳定性值与视觉感受一致。

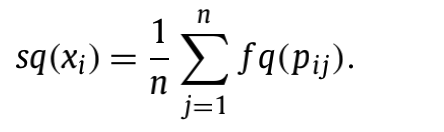
### 3.3 基于二次函数的稳定性

在本小节中，我们考虑一个简单的确定性映射函数，即二次函数，其定义如下：

 **（9）**

二次函数f q的曲线如图6所示。

对于f q，样本xi的稳定性可通过以下公式计算：

**（10）**

对于图1中的示例，基于二次函数的样本稳定性值为：

sq1 = 0.2130，sq2 = 0.2130，sq3 = 0.0648，

sq4 = 0.1389，sq5 = 0.1759，sq6 = 0.1759。

在该示例中，通过平方获得的稳定性值与视觉感知一致。

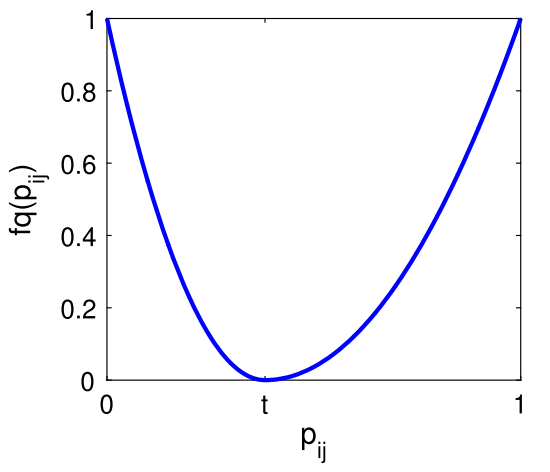


图6 fq曲线

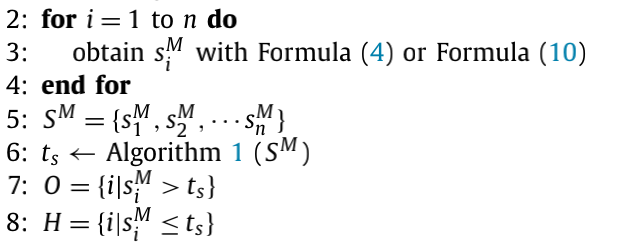
**Algorithm 2** Finding a cluster core and a cluster halo. 找到簇核和簇环

**INPUT:** clustering results set *(n* × *L) 输入结果集(n* × *L) 即a矩阵*

**OUTPUT:** indices of cluster core samples *O*, 簇核样本指标O

indices of cluster halo samples *H*. 簇环样本指标H

1: calculating the co-association matrix *M(n* × *n) 计算协关联矩阵M(n* × *n)即C矩阵*



2-4:利用公式（4）或（10）计算样本的稳定性

5:保存每个样本的稳定性

6:利用算法1求得阈值ts

7:当样本的稳定度大于阈值ts时将其设定为簇核（ts说明他足够稳定了可以当作聚类核）

8:当样本的稳定度小于阈值ts时将其设定为簇环（<ts说明它不稳定可以作为聚类环中的样本）

可以基于公式（4）或公式（10）计算每个样本的稳定性。 有了这些稳定性的度量，聚类核和聚类环可以通过阈值确定。 很容易发现清晰的预结构，这个结构仅基于聚类核的样本。 该预结构可用于指导确定的聚类环的分配。基于以上结果，本文设计了基于样本稳定性的聚类集成算法。（聚类环中的样本如何分配？）

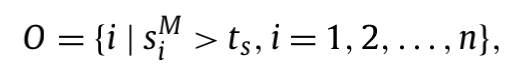
## 基于样本稳定性的聚类集成算法

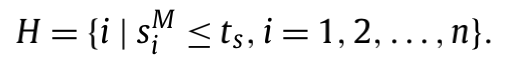
在本节中，我们将基于样本的稳定性提出一种聚类集成算法（以下简称CEs2）。 CEs2的主要想法是使用样本的稳定性测度来找到聚类核和聚类环，并分别处理聚类核和聚类环中的样本，以发现其底层结构。 CEs2包含四个部分：（1）查找基于样本稳定性的聚类核和聚类环（2）基于聚类核中的样本发现底层结构；（3）将聚类环中的样本分配给该结构； （4）调整结构以获得聚类解。

### 4.1 使用稳定性找到聚类核

通常，在聚类分析中，将聚类核定义为具有可靠分配的部分。提高聚类算法鲁棒性的有效的方法是找到合适的聚类核并使用聚类核来指导聚类环中的样本分配。通常根据数据集的分布来找聚类核心。但是，在聚类集成中，由于未知数据的原始特征，很难确定数据集的分布。为了找到聚类核，在这里，我们将利用第3节中介绍的示例的稳定性。

假设聚类核中的样本分配是可靠的，然后对于聚类核样本不同的基本聚类结果可能产生相似的分配，那么聚类核中的样本就会有很高的稳定度，因此可以有效的利用稳定性来确定聚类核，换言之，假设聚类核中的样本比聚类环中的样本稳定性高可以利用稳定性来确定聚类核，所有样本的稳定性S M = {sM1 , sM2 ,..., sMn } 可以利用公式（4）或公式（10）来计算，有了这些稳定性，一个数据集可以根据算法1中确定的阈值t被分为两个组，即聚类核和聚类环，由下给出：

**（11）**

**（12）**

//O代表充当聚类核的样本，当样本的稳定性>ts时说明稳定性高，因此当作聚类核

//H代表聚类环中的样本，当样本的稳定性<ts时说明稳定性不高，因此位于聚类环

**Algorithm 3** Discovering core structure. 算法3:寻找核结构

**INPUT:** indices of cluster core samples *O*, 输入：聚类核样本O

co-association matrix *M*. 协关联矩阵M（c矩阵）

**OUTPUT:** core structure C0\*输出：核结构

1: extracting co-association matrix of cluster core *MO*

2: C0\*<- HC-algorithm *(MO )*

//1:提取协关联矩阵中的聚类核*MO* （通过计算稳定性）

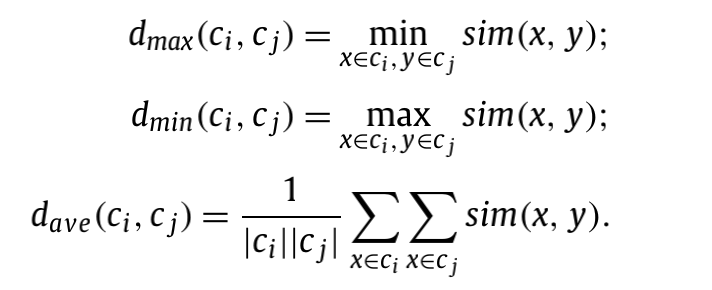
//2:将*MO*传进HC-algorithm算法中得到核结构

在确定了聚类核和聚类环之后，处理这两部分中的样本的过程相当不同。对于聚类核，希望找到一个清晰的底层结构。 通过这种结构，可以分配聚类环中的样本。 接下来，我们讨论一下处理这两种类型的样本的建议。

### 4.2发现聚类核的结构

样本的稳定性由协关联矩阵确定。为了减少计算量，应该使用协关联矩阵来发现聚类核的结构。我们将使用行和列来对应于聚类核中的样本，以形成聚类核的关联矩阵。除此之外协关联矩阵反映了每对样本之间的关系，这有助于发现共识聚类。任何基于协关联矩阵的聚类算法都可以应用于发现聚类核的底层结构。在这里，我们使用层次聚类（HC）算法。

对于具有n个样本的数据集，HC算法以n个簇开始，其中每个簇对应一个样本。HC算法迭代合并两个最相似的簇，直到簇的数量达到预期。两个聚类之间的相似性度量对于聚类结果的质量很重要。 经常用于两个聚类之间的相似性度量是最大相似度，最小相似度和平均相似度，其中分别对应于三种HC算法，分别称为single-linkage（单链接），complete-linkage（完全链接）和average-linkage（平均链接）。 三种相似性测量如下：

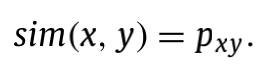


//dmax(ci,cj)计算的是ci类和cj类间的最大距离，因此需要求出ci簇中的某个点x,cj簇中的某个点y，x与y的最大距离即为ci类和cj类间的最大距离，此处中的距离利用两个点的相似度来代替（两个点的相似度最小所以两个点的距离就最大，故为最大类间距离），而在本文中相似度又是利用两个样本之间的稳定性来代替的。

//同理，dmin(ci,cj)计算的是ci类和cj类间的最小距离，因此需要求出ci簇中的某个点x,cj簇中的某个点y，x与y的最小距离即为ci类和cj类间的最小距离（两个点的相似度最大所以两个点的距离就最小，故为最小类间距离），此处中的距离利用两个点的相似度来代替，而在本文中相似度又是利用两个样本之间的稳定性来代替的。

//dave(ci,cj)计算的是ci类和cj类间的平均距离，因此计算来自与ci簇中的每一个点x与来自cj簇中的每一个点y的距离，然后求均值。1/|ci||cj|？

在聚类集成中，样本特征未知。作为妥协，可以使用pij代替两个样本xi和x j之间的相似性。然后，在聚类集成中，两个样本之间的相似度将是：



HC算法在发现聚类核的底层结构方面有两个主要优点。首先，协关联矩阵可以作为聚类集成问题中的相似矩阵。 这意味着HC算法中的输入是可用的，这是该算法的主要计算量。其次，HC算法可以通过合并簇期间的最大跳跃确定簇的数量。 在基于协关联矩阵的合并过程中，当过程遇到相似性最低的一对簇时，它将终止。

聚类核中的样本与其邻居的相似性非常高，与其他簇中的样本的相似性很低。 因此，每个聚类核的关联矩阵中的元素接近于0或1（越接近于1说明与聚类核中的样本相似性越高可以继续合并，接近于0说明无法进行合并）。因为聚类核的底层结构清晰，这三种HC算法将发现非常相似的结构。 在本文中，我们利用单链接算法来发现聚类核的结构。

### 4.3 将簇环中的样本分配给结构

此步骤的任务是将簇环中的样本分配给发现的簇核结构。 簇核结构表示为核样本的聚类结果。另外，我们获得了成对相似矩阵（d矩阵）。然后，分配簇环中样本的直接方法可以基于簇环样本和发现的簇之间的关系。可以将一个簇环中样本分配给最近的簇。应该注意的是，簇环中的一些样本距离所有的簇核都很远，因此，仅基于簇核样本确定其分配可能无效。 在这里，我们建议一种非参数迭代方法将簇环样本分配给所发现结构。

此方法描述了扩展过程，该过程使用簇环中的样本扩展簇核的大小逐步处理整个数据。这种方法是通过两阶段的迭代方法实现的：首先找到离簇核最近的一些样本，然后，通过这些样本扩展簇核。 这两个阶段是相继执行直到所有样本都分配给簇核为止。

**Algorithm 4** Assigning halo samples. 算法4:分配簇环中的样本

**INPUT:** indices of cluster core samples *O*, 输入：簇核样本集O

indices of cluster halo samples *H*, 簇环样本集 H

co-association matrix *M*, 协关联矩阵M（c矩阵）

core structure *C*0 *\** 核结构 *C*0 *\**

**OUTPUT:** pre-structure *C*’\*  输出：预结构*C*’\*

1: while |O| = n do

2: calculate O\_x0005\_’ with Formula (15)

3: extract MOO\_x0005\_’ with M, O and O\_x0005\_’

4: obtain *C*’\*  based on Formula (16)

5: O = O U O\_x0005\_’

6: end while

1: 当簇核中的样本个数<总样本数时，分配簇环中的样本，直到簇环中的所有样本都分配到了簇核中，退出循环

2:利用公式（15）计算O’（等待被分给簇核的簇环中的样本）

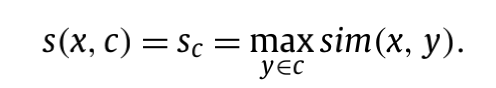
3:从矩阵M(协关联矩阵),O(簇核样本集),O’(等待被分给簇核的簇环中的样本)中提取MOO’

4:从公式（16）中获得*C*’\* (将选择好的簇环中的样本分给离他最近的簇，即簇核里的小簇)

5:O = O U O’即将簇环中的样本分配到簇核

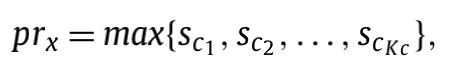
6:直到所有样本都分配给簇核

在第一阶段中，要找到簇核集附近的样本，计算簇环中的样本与发现的簇核集的相似度。 样本x和聚类c（c即为簇核集）之间的相似性可以简单地通过被测样本与其聚类中最近的样本之间的相似性：

**（13）**

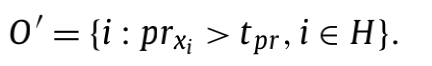
//上式计算的是簇环中某个样本x与聚类（簇核集）中最近的样本y之间的相似性，选出相似度最大的

样本与簇核集的接近程度是由样本与其簇核集（聚类）中最近的样本之间的相似性定义的。对于样本x，其接近度为：

**（14）**

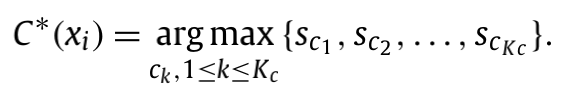
//上式选出了簇环中样本与簇核集中每个小簇接近程度最大的样本

其中Kc是簇核集中已发现簇的数量（也就是说簇核集中装着的都是稳定性高的数据，这些稳定性高的数据已经分成了几个类，这个是怎么计算的？簇核集中应用层次聚类计算得到）。通过公式（14），我们可以获得簇环中的每个样本与簇核集中每个小簇（这个小簇就是稳定高的数据分成的类，比如说x,y,z,m,n在簇核集中，x,y,z是一个类的，m,n是一个类的，则Kc=2）的接近度,利用这些接近度，接近与簇核的样本可以被选择合并到簇核集中，选择的方法是利用算法1中确定的阈值t，（比如说这个小簇就是稳定高的数据分成的类，比如说x,y,z,m,n在簇核集中，x,y,z是一个类的，m,n是一个类的，经过计算簇环中每个样本的接近度，最后选择了p合并进簇核集中，且p与m的接近度最大，所以p在簇核集中和m,n是一类的），被选择的样本指标如下：

**（15）**

//当簇环中的样本xi的接近度>tpr这个阈值时，将xi存入到O’中

接下来，通过标记选定的样本将它们分配给簇核。最直接的方法是分配每个选择的样本到其最近的簇（这个簇指的是上面提到的簇核集中的小簇）。簇环中样本xi的分配可以通过下式得到：

**（16）**

此过程基于以下事实：簇核中的聚类彼此远离（意思就是说小簇分的明确）。在此分配过程完成后，簇核已更改（因为合并了新的样本）。我们需要选择一组新的测试样本。 分配所有簇环样本后，数据集的预结构将被发现。算法4显示了分配簇环样本的详细步骤。 由于簇环样本是逐渐分配的，因此簇环中的关系信息也被充分利用来构造结构。在本文中，关系信息是协关联概率。实际上，以上过程是全部基于成对关系矩阵（d矩阵？）。 因此，如果有一些半监督信息可用，例如必须链接（协关联矩阵中接近于1的一对样本）和必须禁止链接对（协关联矩阵中接近于0的一对样本），通过添加一些约束很容易将上述过程扩展为半监督版本。（d矩阵到底做什么用了？）

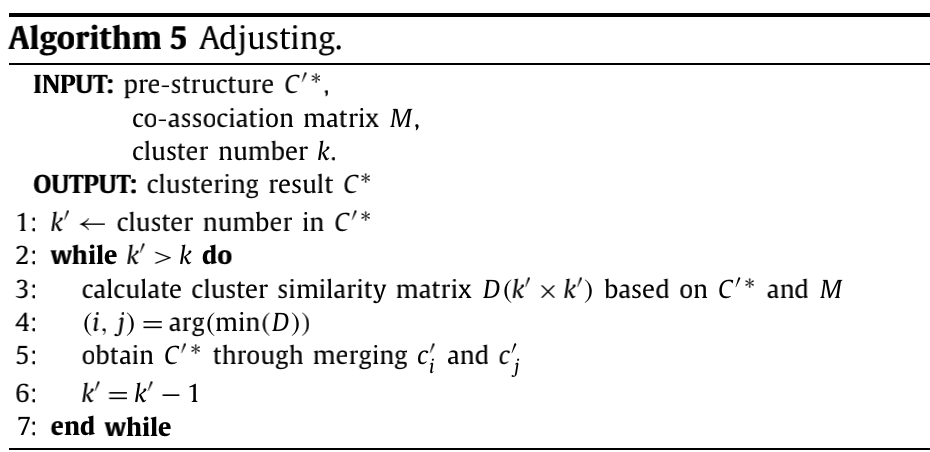
### 4.4 最终聚类的生成

发现的聚类数量与预期的数量不同这是很正常的。 拿走聚类环中的样本可以将簇分为几个部分。 因此，发现了更多的簇。为了产生最终结果，一些紧密的簇将被合并。HC算法也可以执行此任务。合并簇，直到聚类数达到预定义的聚类数。最后，将生成描述数据分布的结构。 调整步骤由算法5表示。按顺序执行算法2至算法5构成了算法的框架CEs2，显示为算法6。

## 实验分析

在本节中，我们验证CEs2的性能。基于这两种稳定性测度，提出了两种CEs2算法，分别是使用基于线性函数的稳定性的CEs2-L和使用基于二次函数的稳定性的CEs2-Q。

实验包括三个部分。首先，我们采用图像分割任务来验证样本的合理性通过sl和sq来衡量稳定性，然后使用八个综合数据集来显示CEs2的工作机理和鲁棒性。最后，我们将两种CEs2算法与十二种最新的聚类集成算法进行比较，得出十种CLUTO聚类工具包[57]附带了来自UCI的基准数据集[56]和六个文档数据集。



算法5:调整

输入：预结构C’\*

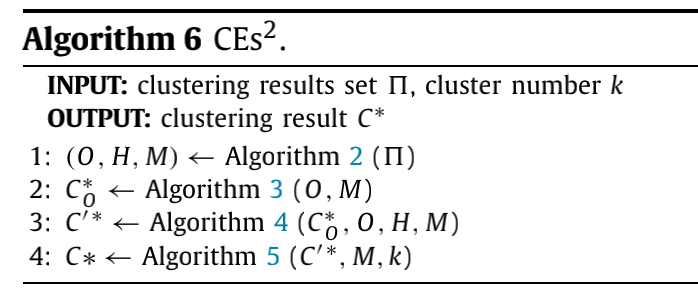
协关联矩阵M

簇数（类别数）k

输出：聚类结果C\*

1. k’← C’\*中簇的数目
2. 当k’>k即预结构C’\*中簇的数目比预先设定的类别数大的时候
3. 通过基于C’\*和M矩阵得到的D矩阵(k’ × k’)计算簇的相似度
4. D矩阵表示的是两个簇中共有的样本是多少，只有共有的样本数越小才说明簇分的越 明确，如何计算的D的最小取值？
5. 通过合并ci’ 和cj’获得C’\*
6. k’=k’-1
7. 结束

//算法5的功能是：当把簇环中的样本分配完以后，发现此时的整体样本聚类后的数目比规定的类别数多时，簇和簇之间进行合并，利用D矩阵进行簇和簇之间的合并，直到簇的数目等于预先规定的聚类数。例如项目要求把100个数据分成10类，而利用算法2-4（计算数据的稳定性，选取聚类核，再把聚类环中的样本分给聚类核），此时聚类数是12，因此需要进行簇和簇之间的合并。



算法6 CEs2

输入：聚类结果集，聚类数目k

输出：聚类结果C\*

1. （O,H,M）← 算法2（）
2. CO\*←算法3（O，M）
3. C’\*←算法4（CO\*,O,H,M）
4. C\* ←算法5（C’\*,M,k）

//算法6实际上就是算法2-5的一个集合，也就是聚类集成的整体框架

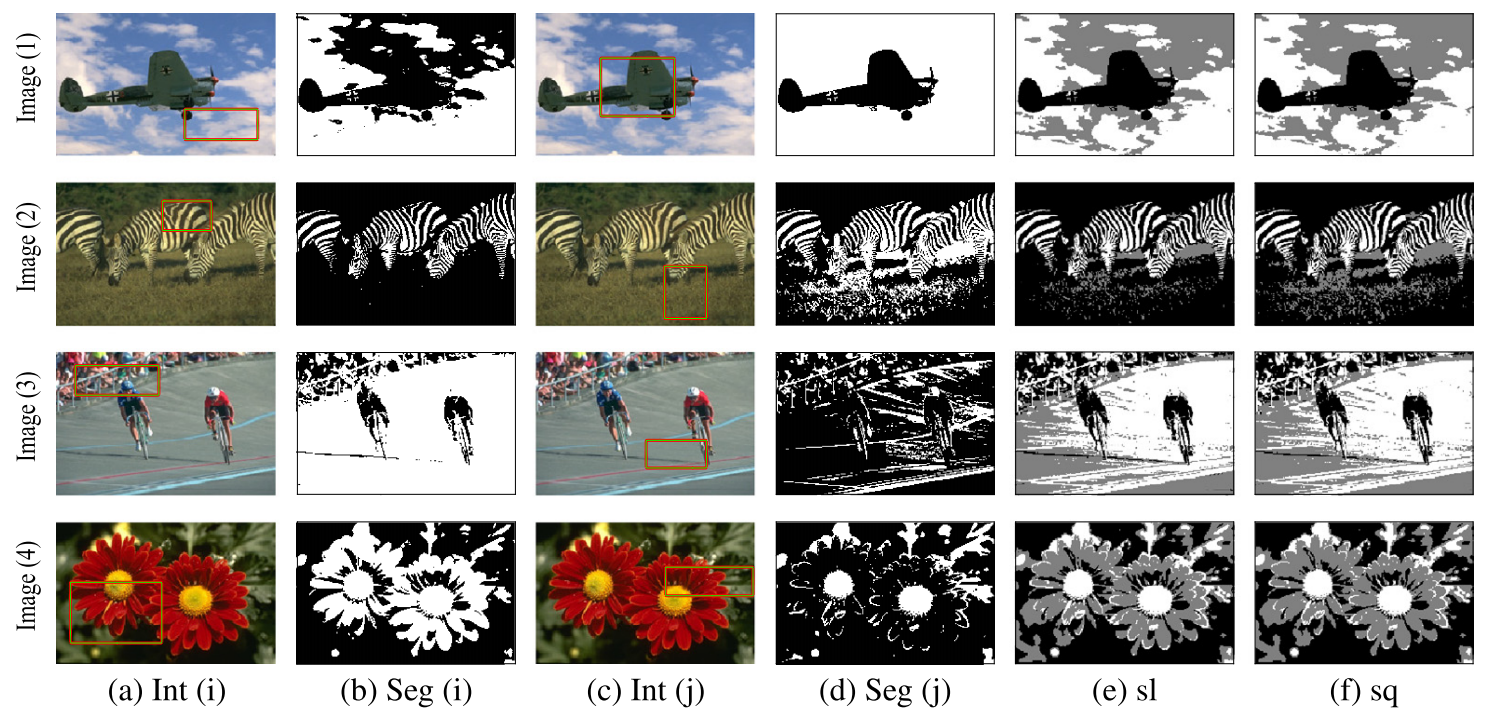


图7.图像分割实验的结果（1）

### 5.1样本稳定性的合理性

为了直观地说明拟议的两个样本的稳定性测度sl（公式4）和sq（公式10）的合理性，我们利用图像分割的情况。图像分割任务是将图像分割成不相交的均匀区域，通常通过发现不同区域的轮廓来实现。在[58]中，Chan and Vese提出了一种最流行的基于两相能级集[59]的方法，它被称为CV方法。CV方法是最小分区问题的一种特殊情况，它最小化了分区的能量。在CV模型中，水平设置功能用于表示轮廓。 CV算法的主要步骤是：（1）轮廓初始化，（2）相对于两个区域的平均强度使能量最小化，（3）相对于两个区域的能量最小化到设定的水平，（4）重复步骤（2）和（3），直到收敛为止。很自然，不同的初始轮廓可能会导致图像的不同分割。因此，我们在图像上多次运行CV方法以生成各种基础分割并根据sl和sq发现不稳定区域。

对于图像，我们首先利用CV方法生成具有随机初始轮廓的50个分割，具有随机大小和随机位置的矩形。然后，我们分别评估图像中每个像素的稳定性与公式（4）和（10）。 基于算法1，我们将图像分为稳定区域和不稳定区域。最后，我们只需使用k = 2的HC算法将稳定区域的分割结果集成在一起，并用灰色绘制不稳定区域。通过上述过程，以三种颜色显示图像，其中白色区域和黑色区域是稳定区域的分割结果，灰色区域是不稳定区域。

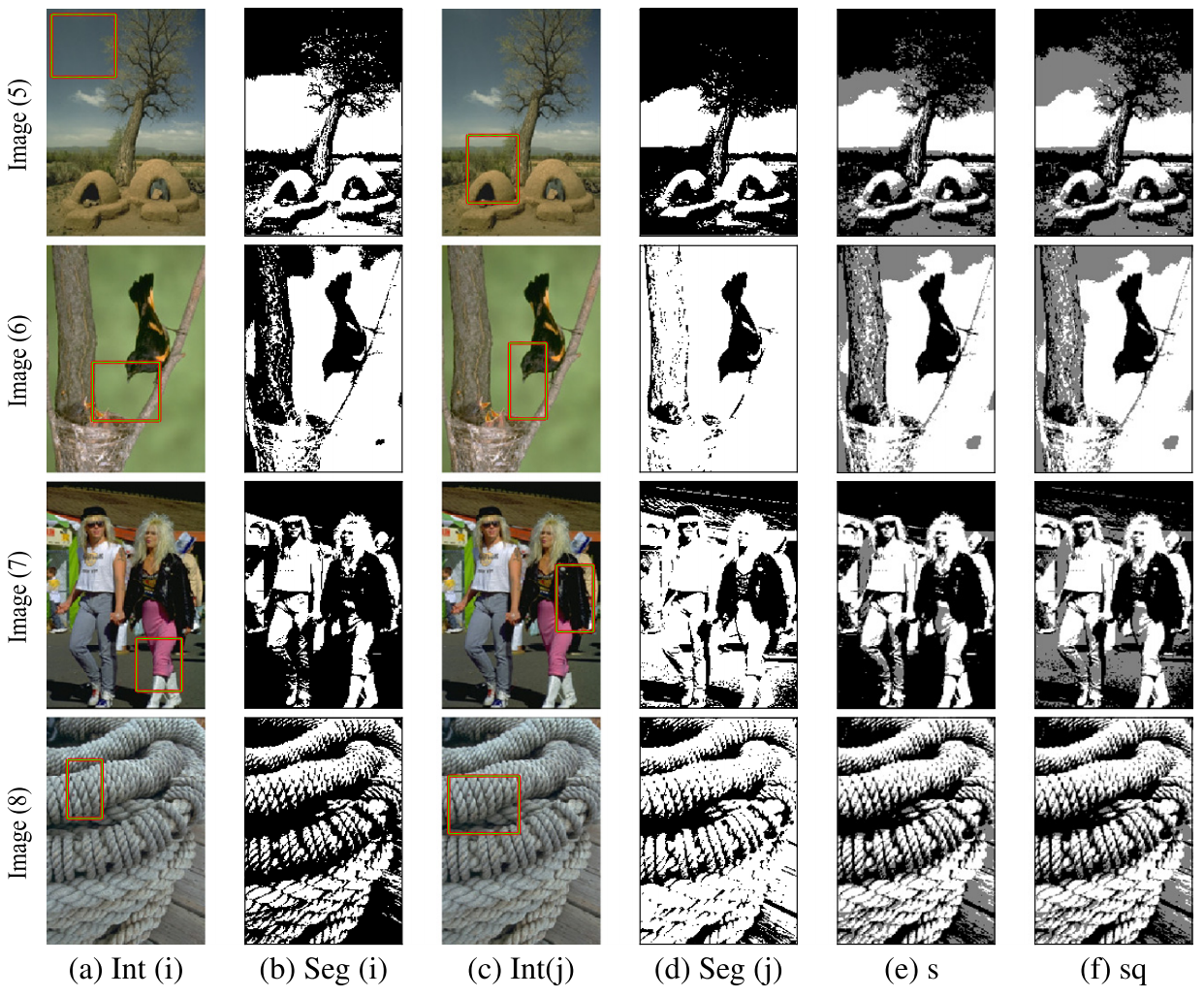
我们采用伯克利细分数据集（500）（BSDS500）[60]进行此实验。图7和图8示出八个示例图像的实验结果。在图7和图8中，（a）和（c）示出了随机初始轮廓的两个例子。（b）和（d）分别显示（a）和（c）的相应分段。 在（e）和（f）中分别显示出了通过基于线性的稳定性sl和基于二次的稳定性sq发现的不稳定区域。

图8图像分割实验的结果（2）

从图7、图8可以看出，CV模型基于不同的初始轮廓生成了不同的分割。 比较子图（e）和（f）以及子图（b）和（d），可以证明（e）和（f）中的灰色区域大致等于（b）和（d）之间的差异。在视觉上，子图（e）和（f）比（b）和（b）更真实，并且可以更清晰地反映原始图像。原因主要包括两个方面。首先，将难以定义的区域识别为不稳定区域。以图像（1）为例，天空被视为不稳定区域，可以将其分割为带有云（图7（b））或带有飞机（图7（j））的同一分区。 同样的事实发生在这样的图像中，作为图（3）中的道路，图（4）中的花瓣和图（6）中的树干。 其次，物体的边界通常是被识别为不稳定区域，从而使对象在视觉上更具立体感。 例如图像（2）中的斑马和图片（8）中的绳索。此外，结果表明，由sl和sq识别的不稳定区域在大多数情况下是相似的。而对于图像（2）和图像（7），sq比sl发现更多不稳定区域。

### 5.2 综合数据集实验

为了可视化CEs2的工作原理，我们在八个综合数据集上展示了CEs2的每个步骤的结果。 表1总结这些综合数据集的详细信息。 图9显示了这些综合数据集的分布。

为了生成一组基本的聚类结果，利用了具有随机初始中心的多个k均值算法。 至于每个聚类中的聚类数k，我们遵循以下建议：k应该大于期望的聚类数，在文献[34,35,31]中进行聚类，并在实验中将每个基本聚类中的聚类数设置为k = min {，50}。实验中每个集成的大小设置为50，即L = 50。

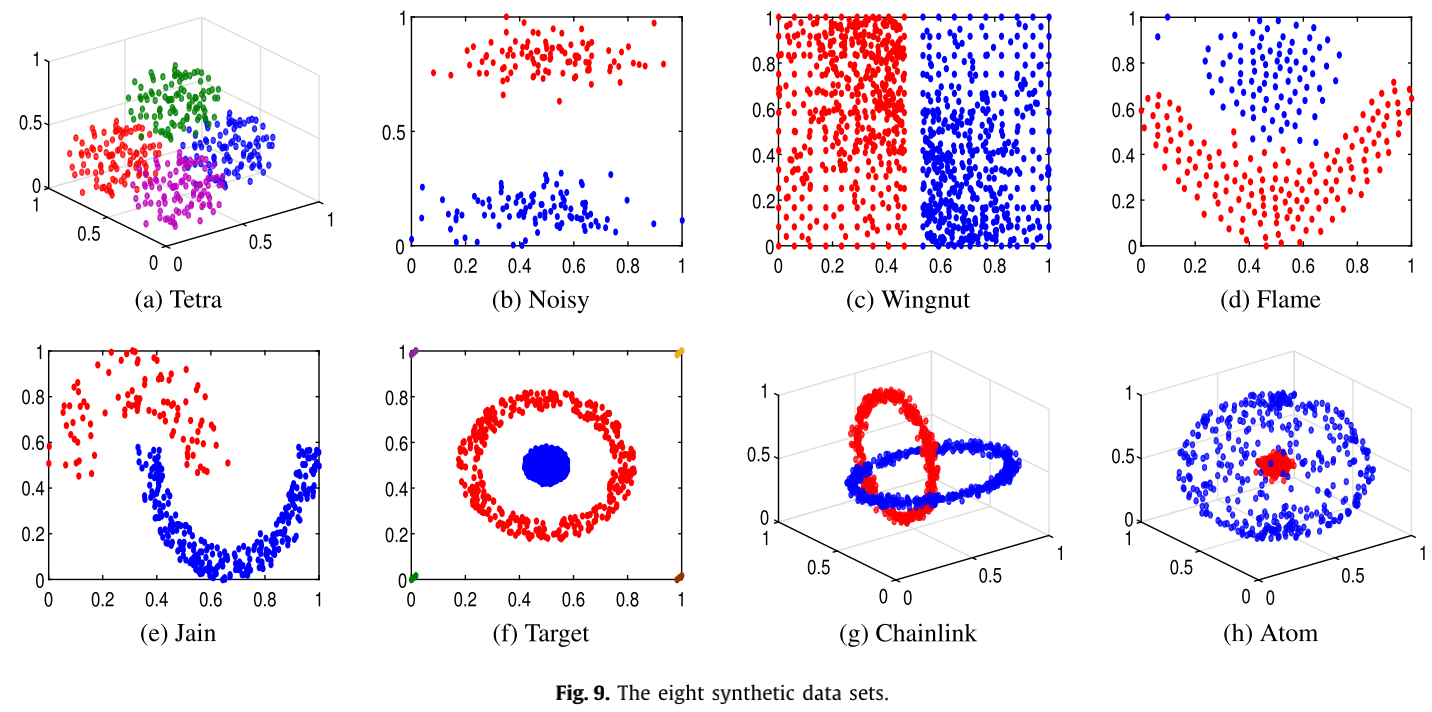
CEs2的每个步骤的结果如图10至图17所示，其中（a）和（d）是基于聚类核的聚类结果，（b）和（e）是分配聚类环中的样本的结果，而（c）和（f）是调整步骤的最终聚类结果。

如图9所示，T etra中的样本是通过混合高斯分布生成的。 该数据集中的聚类为球形。 对于此数据集，基于平方欧几里德距离的传统k均值算法可以生成有效的结果。图10显示CEs2还能够识别这两个数据集中的球形簇。 如图10所示（d），有趣的是，位于聚类中心区域中的样本被识别为属于聚类环。 其原因是当簇数k大于期望值时，中心区域被其周围的簇所分割。在这种情况下，中心区域中的样本将具有较低的稳定性，并且属于簇环。

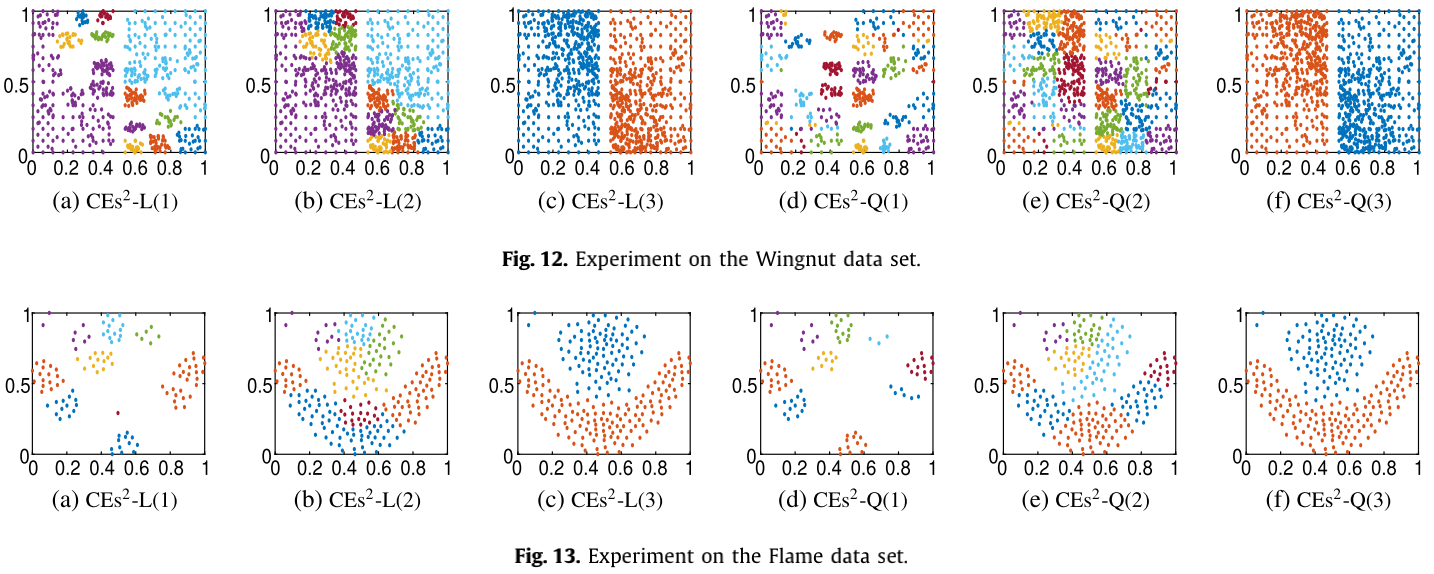
对于图9（b）到图9（h）所示的数据集，生成步骤中的k-means算法无法很好的处理这些数据。 如图11至图17所示CEs2如何将这些低质量的基础聚类结果集成到高质量的聚类结果中。从图11至图17中的（a）和（d）可以很容易地看出，对于高稳定性的样品生成的结果效果很好。 给定（a）和（d）中的结果，低稳定性样本的预测显示在图11至图17的（b）和（e）中。可以观察到，CEs2中的分配步骤提供了更多的聚类，进行了比预期的可以视为目标聚类结果的细分。为了获得更好的聚类，调整步骤将合并这些聚类以达到预期的聚类数量，其结果显示在图11至图17的（c）和（f）中。算法CEs2-L和CEs2-Q有效识别这些合成数据集的基础结构。

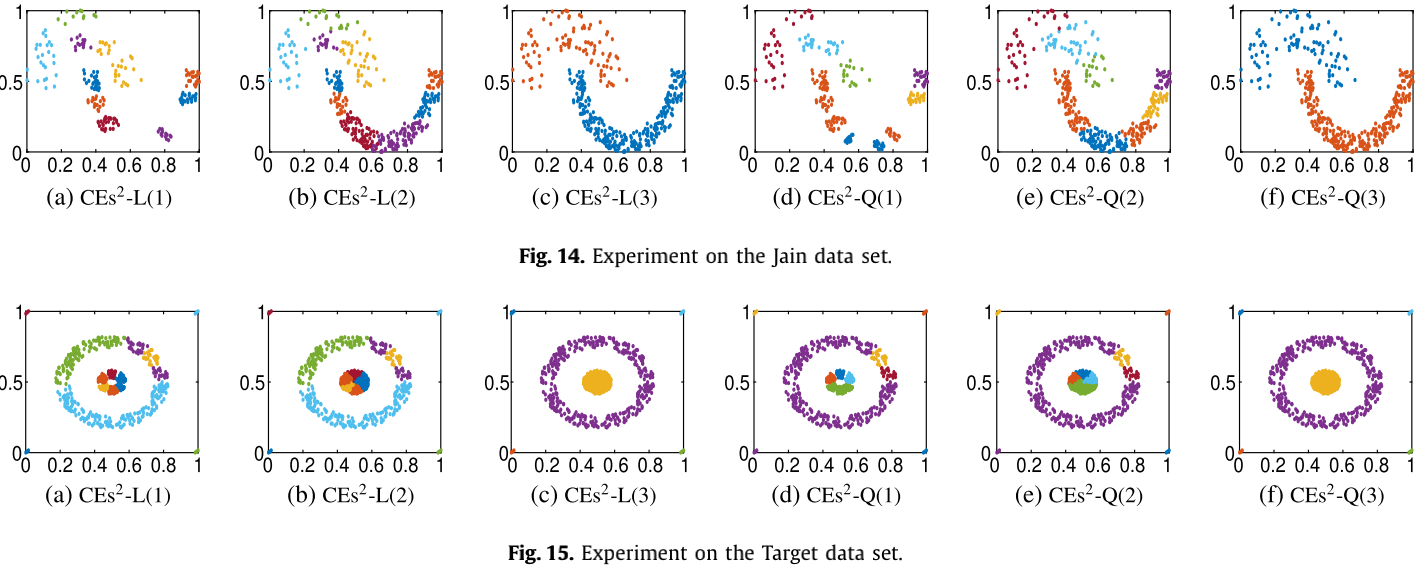
我们还展示了其他十二种聚类集成算法在处理八个合成数据集方面的能力。十二种聚类集成算法包括一种基于特征的方法（[47]），三种基于聚类的方法（WCT，WTQ和CSM [50,4]），一种基于集群关联的方法（MCLA [9]）和七种基于样本关联的方法方法（CSPA [9]，EAC [14]，HGPA [9]，PTA [22]，PTGP [22]，SCCE [16]和NCUT [64]）。投票方法查找聚类对应关系，并利用投票策略生成最终结果。 基于簇矩阵稀疏的事实可能会限制数据分组的质量，因此WCT，WTQ和CSM利用不同的基于链接的相似性度量来完善聚类矩阵，然后通过分区策略生成聚类结果。在实验中，我们使用k均值算法作为分区策略。MCLA基于二元Jaccard度量构建聚类关联矩阵并找到通过簇分组找到聚类对应关系。 在MCLA中，每个样本都分配给与其最相关的聚类。 基于协关联矩阵中，CSPA利用图划分方法METIS生成最终结果，而EAC利用分层聚类算法生成最终结果。HGPA基于n×n矩阵，该矩阵的元素表示连接两个样本的超边缘。 在HGPA中，通过切割最少数量的超边以使超图断开连接来发现聚类结构。 基于协关联矩阵，PTA和PTGP首先建立了基于概率轨迹的相似矩阵。 然后，PTA利用树状图生成最终结果，而PTGP利用Tcut图分区方法。 在SCCE中，频谱算法被用于解决基于协关联矩阵的聚类集成问题。 NCUT是一种图像分割方法，已广泛用于聚类集成算法[52,6]。 表2显示了十二种聚类集成算法处理八个合成数据集的能力。在表2中，符号表示聚类集成算法可以从相应的合成数据中正确发现组结构，而符号×则不能。从表2可以很容易地发现，十二种比较算法中没有一种能够有效处理所有八种综合数据集。表2显示，SCCE算法仅在处理目标数据时失败。 但是，在下一节中，SCCE在处理基准数据集方面表现不佳。

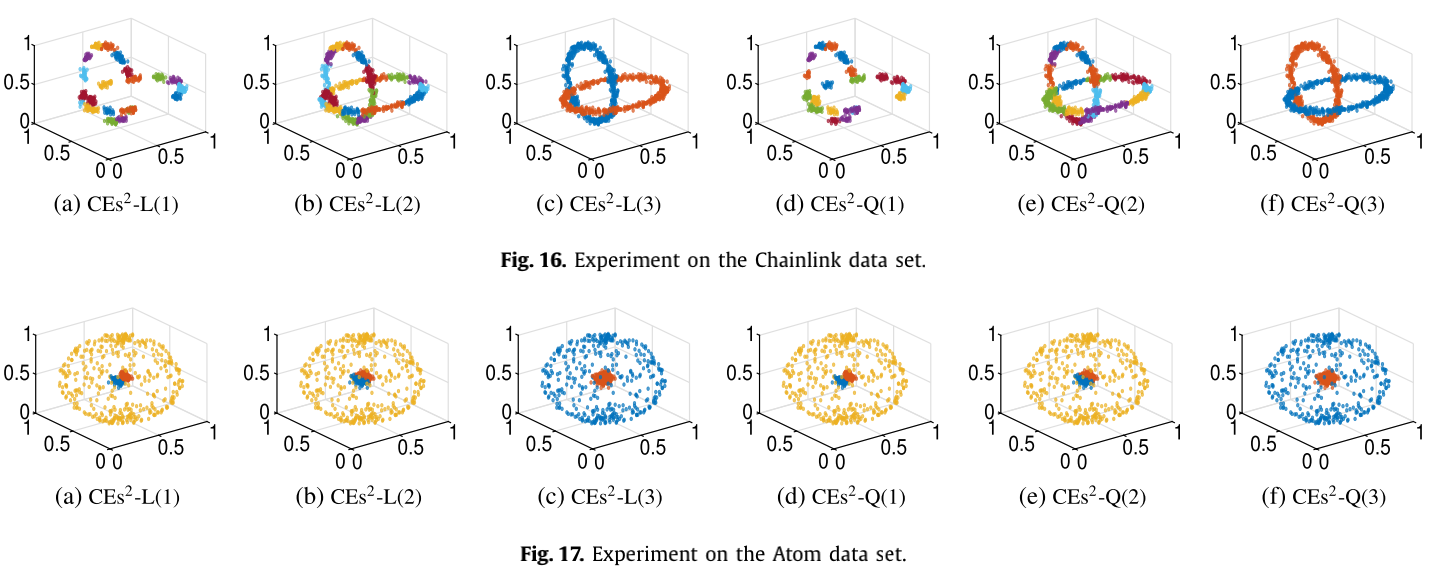


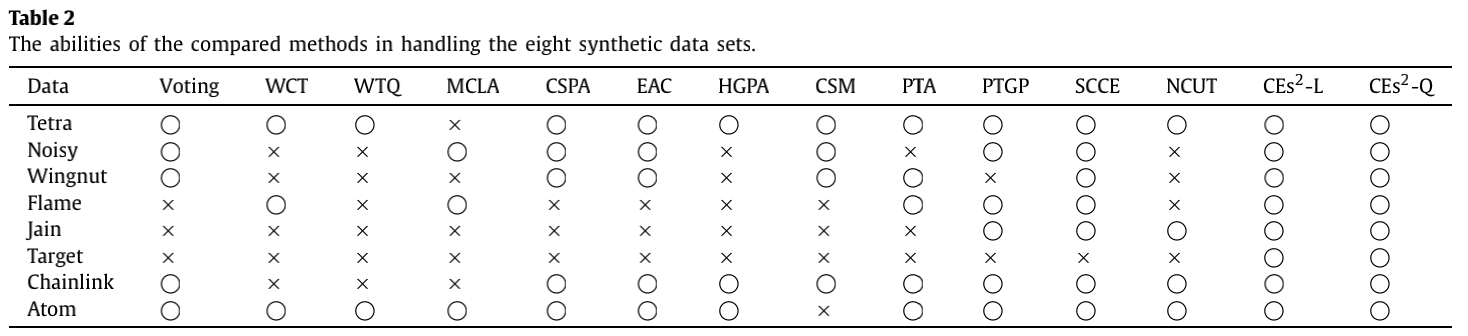


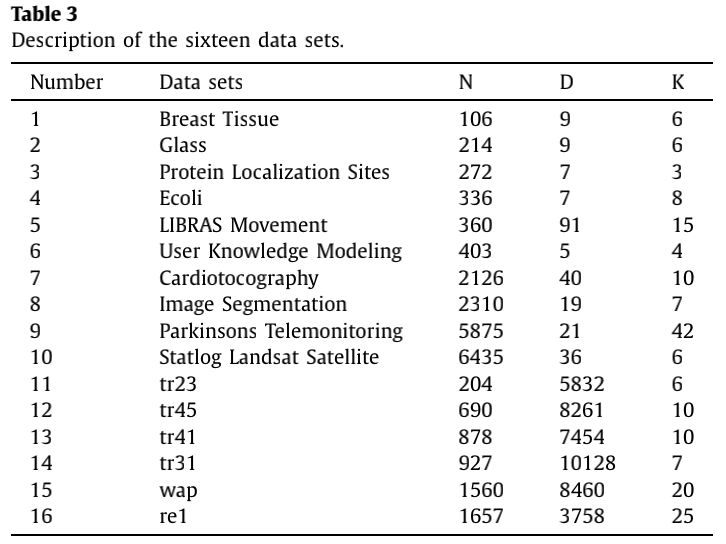












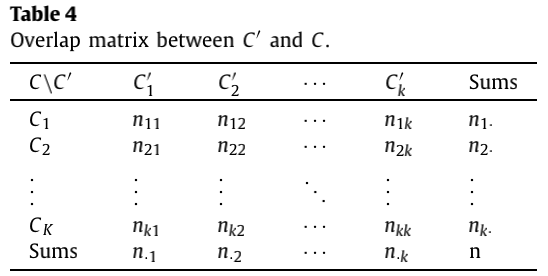
### 5.3 实验基准数据集

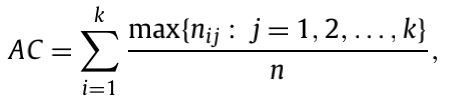
在此比较实验中，使用了来自UCI的十个数字基准数据集和六个文本基准数据集。表3显示了有关这些数据集的详细信息。为了验证CEs2的性能，我们将两种CEs2算法（CEs2-L和CEs2-Q）与十二种聚类进行了比较，集成算法已在5.2节中介绍。

在此实验中，集合大小仍设置为L = 50，每个基本分区中的簇数设置为k = min {，50}。 为了消除由不确定性引起的随机性，对50个集成体进行了每次比较，并报告了平均估计指标值。

为了评估聚类结果的性能，我们利用了两个广泛使用的聚类估计指标，即聚类准确性[65]和归一化互信息[9]。 这两个指标是通过计算聚类算法结果与参考聚类结果之间的相似性来衡量聚类算法性能的外部标准。 在以下实验中，将地面真实情况用作参考聚类。然后，我们仅在比较分区具有相同簇数的环境中引入两个指标，这是每个数据集中的真实数k。

聚类精度（AC）与比较结果中的相应聚类匹配，并报告其常见样本的分数。 根据表4中所示的聚类结果C’和基本事实C之间的重叠矩阵，可以通过以下公式计算AC：



**（17）**

//AC不仅要求每一行唯一，而且要求每一列唯一，也就是一个预测簇只能与一个真实簇对应，一个真实簇也只能与一个预测簇对应。也就是得到的最优解 X={xij}k×sX={xij}k×s是一个正交阵（当k=s时成立）

其中nij是C中簇Ci和C’中簇C j’的公共样本数。

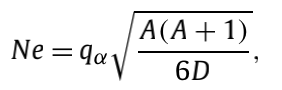
标准化互信息（NMI）计算两个分区之间共享的信息，定义为如下：

 **(18)**

两个索引的范围都在0和1之间，其中较高的值表示更好的性能。

比较实验中的AC和NMI值分别记录在表5和表6中。 在表5和表6中，最后一行显示了每种算法在16个数据集上的平均排名。对于每个数据集，最高索引值都用双下划线标出，第二高值用下划线标出。从表5可以很容易地看出，建议的CEs2算法获得了14个数据集的最高AC值。 其中有十二个数据集，CEs2算法的两个版本赢得了前两个位置。 对于许多数据集，两种CEs2算法可以显着提高许多数据集的AC值。 表6显示，在11个数据集上，CEs2算法获得的NMI值高于其他8种算法。 从表5和表6的最后一行，很明显，两种CEs2算法获得了前两位。 两种CEs2算法的平均等级在2左右，这表明两种CEs2算法在大多数数据集上始终优于其他算法。

为了进一步分析表5和表6中报告的结果，我们利用Friedman检验来检测比较的算法是否显着不同。为了进行此测试，我们使用Matlab函数“ friedman”。根据表5和表6，测试返回的p值分别为1.3026×10-18和2.1224×10-16。两个p值都足够小，这表明至少一对算法存在显着差异。为了直观地显示比较算法的差异，我们应用Nemenyi事后检验[66]。Nemenyi检验的临界值计算如下：

**(19)**

其中A是算法数量，D是数据集数量，当置信度α= 0.05时qα= 4.7427。如果一种算法的平均等级与另一种算法的平均等级不同，则可以认为这两种算法明显不同。在该实验中，利用公式（19），我们获得Ne = 7.0145。 图18和图19示出了Nemenyi检验的结果。在图18和图19中，横轴对应于十四种算法，纵轴对应于平均等级的值。对于每种算法，其平均等级由红点显示，其置信区间由长度为7.0145的蓝线显示。 黑色虚线表示CEs2-L的向上置信度，在两个CEs2算法中获得较高的排名。 从图18和图19可以很容易地看出，两个CEs2算法获得的平均等级比其他算法高得多。 具体来说，CEs2-L和CEs2-Q与CSPA，EAC，SCCE和NCUT显着不同。

## 结论

聚类集成是解决数据聚类问题的有效方法。在过去的十年中，已经提出了许多聚类集成算法，其中大多数算法均等地对待每个数据样本。但是，给定一组聚类结果，样本在聚类之间变化的频率是不同的，这意味着不同样本对基础结构检测的贡献应该不同。在本文中，我们介绍了样本的稳定性以反映这种差异，并提出了一种计算这种稳定性的方法。基于样本的稳定性，我们提出了一种新颖的聚类集成算法（CEs2）。该算法采用不同的过程来处理簇核心中的样本和簇晕中的样本，这些样本根据样本的稳定性进行划分。为了验证样本稳定性的合理性，我们将其应用于图像分割的情况。结果直观地表明，识别不稳定区域，分割结果非常令人鼓舞。对8个综合数据集的实验分析显示了CEs2的工作原理，对10个UCI数据集和6个文档数据集的实验分析证明了CEs2的优越性能。此外，样本的稳定性对于衡量一组基本聚类结果的质量可能是有效的。 因此，也可以利用稳定性来选择聚类结果，这被称为选择性聚类集成。 通常，选择性聚类集成算法仅集成所选聚类结果。 但是，丢弃的聚类结果可能会提供有用的信息。 设计一种区分选定聚类结果和未选定聚类结果的方法可能会很有趣。