深圳大学

本科毕业论文（设计）

题目：基于中心连通性的聚类方法分析

姓名：钟俊鹏

专业：计算机科学与技术

学院：计算机与软件

学号：2015180094

指导教师：贾森

职称：教授

2021 年 4 月 19 日

深圳大学本科毕业论文（设计）诚信声明

本人郑重声明：所呈交的毕业论文（设计），题目《基于中心连通性的聚类方法分析》 是本人在指导教师的指导下，独立进行研究工作所取得的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明。除此之外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。本人完全意识到本声明的法律结果。

毕业论文（设计）作者签名：钟俊鹏

日期：　2021　年 4 月 19 日

目录

基于中心连通性的聚类方法分析

【摘要】随着经济时代的进步和科学计算技术的发展，信息化大潮席卷全球。海量的数据从未与我们的日常生活有过如此密切的联系，也从未像今天那么活跃过。复杂繁多的数据将人与人、人与世界连接起来。每个人都在制造数据、使用数据。正是因此，如何进行数据处理以及选择什么方法来更有效地处理数据显得尤为重要。聚类方法作为一种无监督学习的算法，在近年来发展迅速，取得了许多的进步，并应用在许多的场景下。

本文围绕着基于中心连通性的聚类方法进行研究。基于中心连通性的聚类方法是图类的聚类算法，此算法基于图理论。首先要通过某种相似性度量来构造样本之间的相似性关系，再通过迭代的方式寻找聚类中心，以及非聚类中心样本的分配。

本文主要的研究工作如下：

1.简要地介绍聚类的起源、基本思想和内容，解释了表示样本之间相关程度的距离、相关系数等相似性度量，同时说明了常见的聚类算法。

2.详细说明基于中心连通性的聚类方法的有关定义、聚类过程。通过对通用数据集聚类，并且与常见的聚类算法相比较，说明基于中心连通性聚类方法的可行性。

3.在前面说明基于中心连通性的聚类算法的基础上，将此算法应用在高光谱的波段选择上。同时，通过边缘探测与此算法的结合，改进算法的性能。

【关键词】聚类；图；基于中心连通性的聚类算法；高光谱；边缘探测

1. 引言

1.1 研究背景及意义

1.1.1 研究背景

分类学是聚类分析方法的源头。在原始分类学当中，人们对物品进行分类几乎都是利用自身或者是别人的经验，只有少部分人会使用数学工具对物品进行定性或者定量的分类。因此，主观性、任意性是原始分类学的缺点。它不能够很好地表现事务之间内在本质的关联性和差异性。对于特征多、分类指标多的事物或数据，原始的分类学更是难以用客观的准则去进行准确地分类。随着人们对各种事物深入地探究，人类在各领域对分类地需求推动了分类的高要求。因此，人们逐渐引入数学工具来改进原始的分类学。经过不断地研究和改进，传统的分类学也在不断地演化与进步，形成了数值分类学。再通过引入多变量分析，最终形成了聚类分析。

如今，聚类分析方法应用在了许多领域如学术研究、生产等领域，并且取得了丰硕的成果。

1.1.2 研究意义

聚类分析方法作为无监督的分类方法，在进行数据分类、信息提取中有着广泛的应用。当前，我们正处于大数据时代，我们所接触到的数据无论是数量上还是更新速度上，都要远远的超过前人。对于我们不了解的数据，我们无法凭借已有的经验快速地、准确地提炼出数据当中有用的信息。而通过聚类分析方法，通过某种度量去判断数据中各部分的相似相异程度，我们能够快速的了解到数据的基本结构、类型以及相关的内容。研究和发展聚类分析方法可以有力的帮助人们发现数据的内在特点、解决数据带来的有关问题。

1.2 聚类分析的基本思想

聚类分析认为，研究的样本里，样本和样本之间存在着不同程度的相似性或者相关联性。根据多个样本的多个特点，首先观察分析样本的特性，找出能够表示样本之间的相似性的度量，并根据该度量，采用聚类方法将所有的样品归为不同的类别，使同一种类别的样本具有较大的相似性，而不同类别的样本具有较小的相似性或相关联系。我们将这种分类方法称为聚类分析。

1.3 聚类分析的内容

聚类中心和聚类数目的确定是聚类分析的关键。许多聚类方法已经被广泛探索过。常见的聚类方法有 K-Means 聚类[1]、均值漂移聚类[2]、基于密度的聚类[3]、层次聚类等聚类方法、基于图论的聚类等聚类方法（如谱聚类）。本文主要介绍基于中心连通性的聚类方法。基于中心连通性的聚类方法是基于图论的方法。

事物之间的相关联性是聚类分析方法的核心。聚类分析方法关键就是确定一个能反应样本之间相关联的，客观的度量，将这种度量称为相似性度量。再根据这种反应样本相关性的相似性度量，将样本分成多个类。目前在聚类分析方法中，距离和相似系数是常用的相似性度量。

1.3.1 相似性度量——距离

用距离来衡量样本之间的相似程度。假设$d(x_i, x_j)$是样本x_i和样本x_j的距离。常用距离:

欧式距离:

$$d(x_i, x_j) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}} \quad (1\text{-}1)$$

绝对距离:

$$d(x_i, x_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \quad (1\text{-}2)$$

Minkowski 距离:

$$d(x_i, x_j) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^m\right]^{\frac{1}{m}} \quad (1\text{-}3)$$

1.3.2 相似性度量——相关系数

对 n 个样本进行聚类的时候，可以使用相关系数作为相似性度量。样本之间的相关联程度反映在相关系数的大小上。用$c_{\alpha\beta}$表示样本x_α和样本x_β之间的相似系数，应当满足以下条件：

$$|c_{\alpha\beta}| \le 1 \text{ 且} c_{\alpha\alpha} = 1 \quad (1\text{-}4)$$

$$c_{\alpha\beta} = c_{\beta\alpha} \quad (1\text{-}5)$$

$|c_{\alpha\beta}|$越接近 1，说明x_α和x_β越相关。夹角余弦和相关系数通常被用作为表征样本数据的相关系数。在基于图的聚类中，我们通常使用高斯核函数。鉴于本文研究内容，下面介绍基于中心连通性的聚类方法相关内容。

1.4 常见的聚类算法

1.4.1 K-Means 聚类算法

给定样本集$D = \{x_1, x_2, \ldots, x_m\}$，通过计算出样本之间的距离大小（这里的距离可以选择不同的距离度量），划分成k个类$C = \{C_1, C_2, \ldots, C_k\}$，经过$N$次计算迭代，让簇类里的数据点之间的距离尽可能的小，不同的簇类之间的距离必须要尽可能大。

$K - Means$聚类算法步骤如下：

步骤 1 从样本集D中随机选取k个样本作为初始的k个质心向量$\{\mu_1, \mu_2, \ldots, \mu_k\}$

步骤 2 对于每一次迭代$n = 1,2,\ldots,N$,首先将簇C初始化为$C_t = \emptyset, t = 1,2,\ldots,k$。对于样本集$D$中的每一个样本数据点$x_i (i = 1,2,\ldots,m)$,计算他们与各质心向量$\mu_j (j = 1,2,\ldots,k)$的距离$d_{ij} = \| x_i - \mu_j \|_2^2$。将$x_i$标记最小的$d_{ij}$所对应的类别$\lambda_i$,更新$C_{\lambda i} = C_{\lambda i} \cup \{x_i\}$。对于$j = 1,2,\ldots,k$，对$C_j$中每一个样本数据点重新计算质心$\mu_j = \frac{1}{|C_j|}\sum_{x \in C_j} x$。经过计算后，如果上一次计算的 k 个质心向量都和原来相同，则输出簇划分$C = \{C_1, C_2, \ldots, C_k\}$，否则继续迭代。

K-Means 聚类算法的优点是计算速度相对比较快，计算思路简洁，但是必须要提前将数据划分为指定的类别数，而且聚类结果受初始选择点的影响。

1.4.2 DBSCAN

DBSCAN （density-based clustering method with noise）是不同于 K-Means 聚类算法的另外一种算法，它是基于样本空间密度的聚类方法。这类型的算法通过对样本分布的紧密程度进行聚类，把密度足够大的区域归为一类，而且能在有噪声的空间数据中寻找各种形状的簇类。

相关定义 首先将给定样本集$D = \{x_1, x_2, \ldots, x_m\}$通过半径$Eps$和样本个数阈值$MinPts$将数据分为三类

核心点：如果样本x_i的半径Eps区域内的样本多于或者等于$MinPts$个样本，则该样本点x_i为核心点。

边界点：如果样本x_i的半径Eps区域内包含的样本数目不多于或者不等于$MinPts$个样本，并且它位于其他核心点的半径区域范围内，那么我们认为样本x_i为边界点。

噪声点：如果样本x_i的半径Eps区域内包含的样本数目不多于或者不等于$MinPts$个样本，同时x_i也不处于其他核心点的半径区域范围内，那么我们认为样本x_i为噪声点。

$DBSCAN$算法步骤如下：

步骤 1 给定数据样本集合$D = \{x_1, x_2, \dots, x_m\}$,半径区域范围参数$(Eps, MinPts)$,对核心点集合进行初始化$\Omega = \emptyset$,对簇类数目进行初始化$k = 0$,初始化还没有访问过的样本集合$\Gamma = D$,簇划分$C = \emptyset$。

步骤 2 对于$j = 1,2,\dots, m$，寻找核心点。根据核心点的定义，将符合核心点定义的样本点x_j加入核心点集合$\Omega = \Omega \cup \{x_j\}$.

步骤 3 如果核心点集合$\Omega = \emptyset$,说明该数据样本集合的样本都已经划分到对应的簇类当中，则算法结束。否则，说明还有簇类和样本未被划分，进入步骤 4。

步骤 4 在核心点集合Ω中，选择一个核心点o,并对当前簇核心点集合$\Omega_{cur} = \{o\}$进行调整和初始化,同时改变簇类序号$k = k + 1$,对当前样本集合进行初始化$C_k = \{o\}$,更新还没有被访问的样本集合$\Gamma = \Gamma - \{o\}$.

步骤 5 如果当前簇核心点集合$\Omega_{cur} = \emptyset$,说明当前进行聚类的簇当中已经没有未被访问的核心点，则当前聚类簇C_k生成完毕，更新簇划分$C = \{C_1, C_2, \dots, C_k\}$,更新核心点集合$\Omega = \Omega - C_k$,到步骤 3，否则到步骤 6

步骤 6 在当前簇所包含的核心点集合Ω_{cur}中取出一个核心点o',通过Eps找出半径区域范围内所有样本集$N(o')$,令$\Delta = N(o') \cap \Gamma$,并且更新当前簇样本集合$C_k = C_k \cup \Delta$,更新还没有被访问的样本集合$\Gamma = \Gamma - \Delta$,更新$\Omega = \Omega \cup (\Delta \cap \Omega) - o'$,继续执行步骤 5

最终输出簇划分结果$C = \{C_1, C_2, \dots, C_k\}$

DBSCAN 相较于一些传统的聚类方法的优点是不需要预先声明需要划分簇类的数目k,并且可以对任意类型的密集数据集进行分类；在进行聚类的同时，可以找到噪声点。从 DBSCAN 的聚类过程可以知道，它对数据集的异常点(噪声点)不敏感；不仅如此，相较于 K-Means 聚类算法，DBSCAN 不受初始值选取的影响。

然而，在空间聚类的密度不均匀的条件下，参数Eps和$MinPts$选取困难；当数据集较大的时候，DBSCAN 聚类收敛的时间会相对较长；由于需要调整参数Eps和$MinPts$,不同的参数组合对实际的聚类结果影响较大，相对于传统的 K-Means 聚类算法只需要调整一个簇类划分参数显得更加复杂。

1.4.3 均值漂移聚类

均值漂移聚类算法是结合了均值漂移思想和滑动窗口思想的聚类算法。均值漂移背后的主要思想是将多维特征空间中的点视作为一个经验概率密度函数。其中，特征空间中的密集区域对应于总体分布的局部极大值或者众数。对于特征空间的每一个数据点，它都会沿着梯度上升的方向朝局部密度最大点移动直至收敛。在这个过程中，固定的数据点代表分布的集中点。除此以外，与固定数据点相关的其他数据点可以被认为是属于同一簇的成员。均值漂移聚类相关定义如下

$Mean\ Shift$向量：对于给定的d维空间R^d中的n个样本点$x_i = 1,2,\ldots,n$,对于x点，其$Mean\ Shift$向量的基本形式定义为：

$$M_h(x) = \frac{1}{k}(x_i - x) \tag{1-6}$$

半径区域范围S_h：半径区域范围指的是以x为中心点，一个半径为h的高维球区域中的样本点，即样本集中到样本点x的距离小于高维球半径h的样本点，定义为：

$$S_h(x) = \{y: (y - x_i)^T(y - x_i) < h^2\} \tag{1-7}$$

k表示在S_h范围内样本点的个数。

中心更新：将中心点通过偏移向量移动到偏移均值的位置。

$$x^{t+1} = M^t + x^t \tag{1-8}$$

其中M^t为在状态t下求得的偏移均值，x^t为状态t下的中心点。

核函数：为了计算低维不可分的数据，我们常将数据映射到高维空间。然而低维空间映射到高维空间中会出现维数灾难问题。通过核函数变换技巧，不仅能够计算低维不可分的数据，也避免了维数灾难问题。利用核函数这种变化技巧，可以通过忽略低维空间映射到高维空间的关系，在低维空间中完成相关数据的计算。为了实现样本与被偏移点的距离不同，偏移量对$Mean\ Shift$向量的贡献不同的目的，我们在$Mean\ Shift$聚类方法中使用核函数。核函数是在一些计算当中经常使用的一种变换技巧，也是机器学习中常用的方法，核函数的定义如下：

X表示一个d维的欧式空间，x是该空间的一个点，$x = x_1, x_2, x_3, \ldots, x_d$，其中，$R$ 表示为实数域，如果一个函数 $K: X \to R$ 存在一个剖面函数，$k: [0, \infty] \to R$，即

$$K(x) = k(\| x\|^2) \tag{1-9}$$

并且核函数需要同时满足以下条件：k是非负的、非增的和k是分段连续的。如果满足上述条件，那么函数$K(x)$就称为核函数。核函数有很多，例如经常使用的核函数有线性核函数、多项式核函数和径向核函数(高斯核函数)等。高斯核函数定义如下：

$$K(x) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{x^2}{2h^2}} \tag{1-10}$$

其中，h称为带宽(bandwith)。在高斯核函数中，将带宽h的值固定时，高斯核函数的值和样本点之间的距离成反比。如果样本点之间的距离相同，随着高斯核函数的带宽h的增加，高斯核函数的值会减小。

引入核函数的$Mean\ Shift$向量：为了在计算中反映离中心点的距离越近，其权值越大的特性，我们在均值漂移中引入核函数，将原来的偏移向量进行改进。改进的$Mean\ Shift$向量定义如下：

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i G\left(\frac{x_i - x}{h_i}\right)}{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h_i}\right)} - x \tag{1-11}$$

其中，x为中心点，x_i为带宽范围内的点；n为带宽范围内的点的数量；$G\left(\frac{x_i - x}{h_i}\right)$为核函数。

均值漂移聚类算法步骤如下：

步骤 1 在未被分类的样本点中随机选择一个点作为中心点v

步骤 2 以v为中心点，计算$Mean\ Shift$向量$m_h(v)$。

步骤 3 沿着$Mean\ Shift$向量移动中心点v。

步骤 4 重复步骤 2，3 直至收敛($\| m_h(v) \|$小于阈值)并记录下此时中心点的位置

步骤 5 重复步骤 1，2，3，4，收敛到相同点的样被认为是同一簇类的成员。

均值漂移聚类相较于$K-Means$聚类算法，不需要预先设定好最终要划分的聚类个数，并且对需要处理的聚类的形状没有要求；同时，算法的结果相对稳定，对样本初始点的选择没有要求。

然而，带宽的设置决定了均值漂移聚类方法的聚类结果。如果带宽的数值设置得过于小，最终我们得到的簇类个数会比较多，同时该聚类过程收敛速度会比较慢；如果我们将带宽的数值设置得过于大，最终可能会损失一些簇类，使簇类数目减少，导致结果的准确率降低。如果数据集处于较大的特征空间中，使用$Mean\ Shift$聚类算法进行聚类的计算时间会变得很大。

1.5 本文研究内容

本文主要利用基于中心连通性的聚类算法在通用数据集上进行聚类研究，并与传统的聚类算法比较他们之间的性能。通过结合边缘探测算法，增强基于中心连通性的聚类算法在高光谱数据集波段选择的性能。

1.6 论文总体结构

本论文主要通过 3 个章节进行阐述，下面为每一个章节的简要叙述：

第一章为引言，主要介绍了聚类分析方法的研究背景以及意义、聚类方法的基本内容以及常见聚类方法。

第二章详细的介绍了基于中心连通性的聚类算法。

第三章详细的介绍了基于中心连通性的聚类算法应用在高光谱数据的波段选择，并通过与边缘探测方法相结合，达到提高算法性能的目的。

2.基于中心连通性的聚类方法及其研究

2.1 基于中心连通性的聚类方法

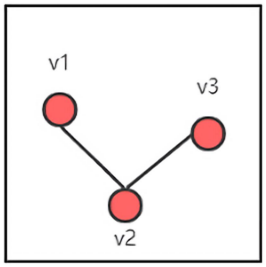
基于中心连通性的聚类方法是基于图论游走次数理论的。唯一不同的是，在基于中心连通性的聚类方法当中，在无向图中的每一个顶点都是自连的。

我们认为聚类是一种演化的、动态的形式。一个数据点是否是聚类中心取决于我们怎样去观察它。我们将需要聚类的数据映射到无向图的顶点中。基于中心连通性的聚类方法拓展图论游走次数理论，通过简单地比较相似矩阵的元素，动态智能地确定聚类中心以及数据的分类。不仅如此，我们可以通过相似矩阵不同次数的幂矩阵来适应我们想要的观察规模。

假设我们现在有一个无向图G，它的自连邻接矩阵为A。$a_{ij}^{(k)}$表示A的k次幂矩阵A^k中第i个顶点v_i到第j个顶点步长为k的数目。

下面举一个简单的例子。在图 1 中，在我们的三次幂邻接矩阵A^3中，$a_{22}^{(3)}=7$ 表示从v_2到v_2步长为 3 的数目为 7

然而对于真实的数据集，我们无法构造出相对应的邻接矩阵。我们只能构造数据的成对相似矩阵。因此我们有必要对上面的关于邻接矩阵的理论拓展到相似矩阵。下面的关于连接性概念只是对上面理论的一个拓展。这两个概念不同之处在于，邻接矩阵是由无向图计算出来的，邻接矩阵中所有的元素都是整数。而连接性是由数据相似矩阵所定义的，所以仅是实数就可以。



$$A=\begin{pmatrix}1&1&0\\1&1&1\\0&1&1\end{pmatrix}\qquad A^3=\begin{pmatrix}4&5&3\\5&7&5\\3&5&4\end{pmatrix}$$

图 1.一个简单的例子说明一个自连通无向图的步行次数

2.2 相关概念及定义

定义 1(k 阶连接性) 在相似矩阵S中，k阶相似矩阵S^k中的$s_{ij}^{(k)}$定义为v_i和v_j的k阶连接性，用$con^{(k)}(v_i,v_j)$来表示。显然$con^{(k)}(v_i,v_j)$能够近似地代表v_i和v_j之间步长为k的数目。$con^{(k)}(v_i,v_j)$能够表示v_i和v_j之间的关联程度。

定义 2(聚类中心) 如果顶点v_i满足以下条件，那么v_i就定义为连接中心和数据的k阶聚类中心

$$con^{(k)}(v_i,v_i)>con^{(k)}(v_i,v_j),j=1,\ldots,n(j\neq i) \tag{2-1}$$

定义 3(k阶相对连接性) 对于任意数据点v_i和v_j，k阶相对连接性定义为

$$rcon^{(k)}(v_i,v_j)=con^{(k)}(v_i,v_j)/con^{(k)}(v_i,v_i) \tag{2-2}$$

k阶相对连接性相较于k阶连接性，能够消除自身连接性的影响，使得数据点分配到聚类中心的时候更加准确。

定义 4(无向图切图) 对于无向图$G = \{v_1, v_2, \ldots, v_n\}$，我们将其分割为$m$个子图$A_1, A_2, \ldots, A_m$，并满足以下条件

$$A_1 \cup A_2 \cup \ldots \cup A_m = G \tag{2-3}$$

$$A_i \cap A_j = \emptyset\left(i, j = 1,2,\ldots,m\text{且}i \neq j\right) \tag{2-4}$$

定义 5(无向图切图权重) 对于任意两个子图A，B（A，$B \subset G$且$A \cap B = \emptyset$）,我们定义其权重为

$$W(A,B) = \sum_{i\in A, j\in B} w_{ij} \tag{2-5}$$

其中w_{ij}为邻接矩阵中的元素。那么对于m个子图，我们定义切图 cut 为

$$cut(A_1, A_2, \ldots, A_m) = \sum_1^m W\left(A_i, \bar{A_i}\right) \tag{2-6}$$

$\bar{A_i}$是除了A_i子图外其他子图的并集。

定义 6(正规化切图) 对于定义 5 的切图，我们做进一步的处理

$$Ncut(A_1, A_2, \ldots, A_m) = \sum_1^m \frac{W(A_i, \bar{A_i})}{Vol(A_i)} \tag{2-7}$$

其中$Vol(A) = \sum_{i\in A}\sum_{j=1}^{n} w_{ij}$

分类规则 假设现在我们有m个聚类中心$v_{c_i}\left(c_i \in \{1,2,\ldots,n\}\text{同时}i = 1,2,\ldots,m\right)$，对于任意数据点$v_j$,它将会被分配到聚类中心$v^*$,$v^*$满足由下面等式得到：

$$v^* = \text{argmax}_{v_{c_i}}\left(rcon^{(k)}\left(v_{c_i}, v_j\right)\right) \tag{2-8}$$

分类规则表示v_j会分类到与它相对连接性最强的聚类中心，十分的直观。

2.3 聚类过程

对于每一次迭代，通过(2-1)可以得到聚类中心。通过分类规则可以将其余的数据点分配到对应的聚类中心。

当$k = 1$，时，所有的数据点都可以看作是聚类中心，初始的类别数目和数据点的数目相同。随着迭代次数k的增加，聚类中心和聚类数目反映了数据点之间的连通性。当迭代次数k继续增加趋于无穷的时候，聚类中心数目会缩减到一个点，只有一个聚类。这时候所有的数据点就属于同一类。因此我们可以将迭代次数k作为我们的观测规模。当观测规模k小的时候，数据点只会被他们邻近的数据点所影响，这时候的类别数相对较多；当观测规模k较大的时候，数据点直接的连接性会传播的更加广泛，这时候的类别数会相对较少。

例如在图 2 中，当$k = 1$ 的时候，数据点v_1,v_2和v_3都可以看作是聚类中心，当$k \geq 2$ 时，只剩v_2这个聚类中心，其余数据点v_1和v_3都被分配到这个v_2这个聚类中心。通过(2-1)和(2-2)我们可以得到在观察规模为k时的聚类中心以及聚类结果。然而，对于一些数据集来说，他们在不同的观察规模k下有相同的类别数，但是他们的聚类中心和其余数据点分类却不完全相同。这时候我们引入正规化切图$Ncut$(定义 6)来确定在这种情况下更好的聚类结果。正规化切图$Ncut$能够表示各子图之间相互关联的程度。子图数目确定的情况下，$Ncut$越小，各子图相互的关联性就越小。

显然，根据上述定义，在类别固定的情况下，对于我们所得到的多个聚类中心以及其他点的分类结果，$Ncut$越小，结果就更加合理。

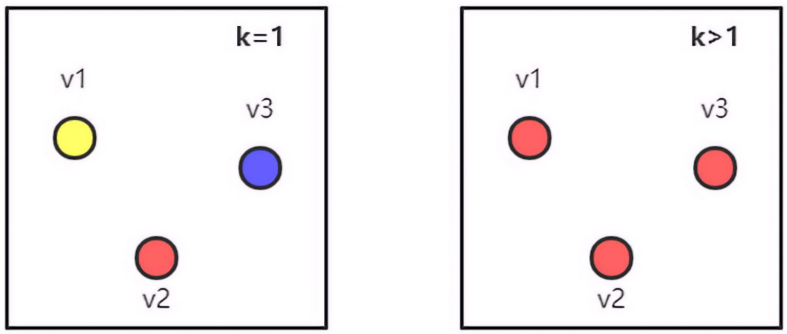


图 2.聚类中心选择与分配

2.4 可行性

2.4.1 鸢尾属植物数据集

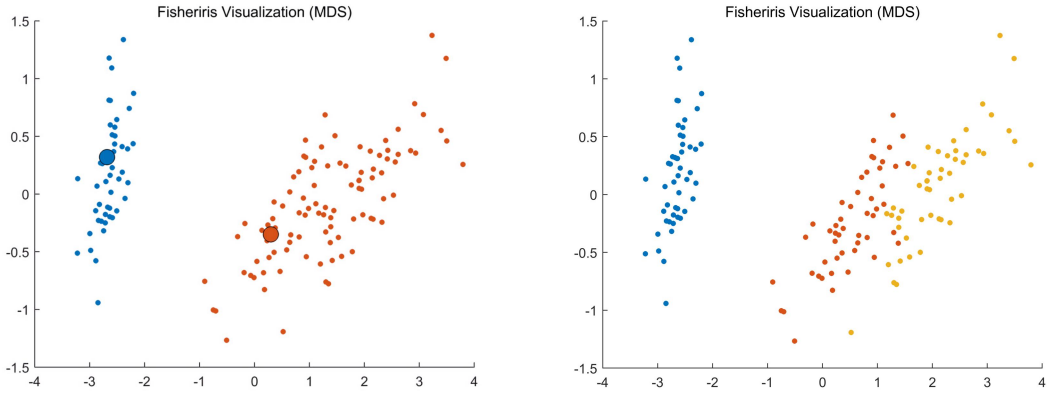
鸢尾属植物数据集[4](Iris Data Set)是著名的数据集。在鸢尾属植物数据集中，包括了三类不同的鸢尾属植物：山鸢尾(Iris Setosa)、杂色鸢尾(Iris Versicolour)和维吉尼亚鸢尾(Iris Virginica)。此数据集中一共包含了 150 个样本，每个样本包含了四个特征，分别是：花萼长度(sepal length)、花萼宽度(sepel width)、花瓣长度(petal length)以及花瓣宽度(petal width)。以上四个特征的单位都是厘米(cm)。

2.4.2 使用基于中心连通性聚类方法对鸢尾属植物数据集进行分类

对于鸢尾属植物数据集,总共有$m = 150$ 个数据，每个数据都有$n = 4$ 个特征。

在鸢尾属植物数据集$D = \{v_1, v_2, \ldots, v_{150}\}$中，$v_k \in D$代表鸢尾属植物数据集中第$k$个样本。这里我们通过高斯核函数来构造数据点之间的相似矩阵。

$$H_{Gauss}: S_{ij} = exp\left(-\| v_i - v_j \|^2 / \sigma^2\right) \tag{2-9}$$



(A) (B)

图 3.鸢尾属植物数据集分类可视化

通过 **2.3 节的聚类过程**，我们可以得到不同迭代次数时鸢尾属植物数据集的聚类中心和聚类数目。在聚类过程中，基于中心连通性的聚类方法认为，鸢尾属植物数据集应该分为两类，而数据集本身是有三类的。如图 3(A)所示，通过MDS方法可视化鸢尾属植物数据集，我们可以发现基于中心连通性的聚类方法将数据集分成两类是合理的。在图 3(B)中，我们发现，鸢尾属植物数据集的杂色鸢尾(Iris Versicolour)和维吉尼亚鸢尾(Iris Virginica)十分接近，甚至于有些混在了一起。

2.4.3 调整兰德系数

调整兰德系数[5](Adjusted Rand index)用于聚类模型的性能评估。并不是所有数据集都能够使用调整兰德系数来进行聚类方法的性能评估，只有数据集本身具有标记，才能够使用这个度量。调整兰德系数是一个标量，其表示的范围在$[-1,1]$之间，反映了数据标记实际划分与经过聚类方法得到的划分的重叠程度。它的数值越大，则在此数据集上该聚类模型的性能越好。

定义 1(兰德系数) 给定一个有n个样本的集合$S=\{o_1,o_2,\ldots,o_n\}$,$X=\{X_1,X_2,\ldots,X_r\}$和$Y=\{Y_1,Y_2,\ldots,Y_s\}$是对集合S的两个不同划分。给出如下定义

a为在X中为同一类，在Y中也为同一类的对象对数。

b为在X中为同一类，在Y中不为同一类的对象对数。

c为在X中不为同一类，在Y中为同一类的对象对数。

d为在X中不为同一类，在Y中不为同一类的对象对数。

兰德系数(Rand Index)的计算公式定义为

$$RI=\frac{a+d}{a+b+c+d} \tag{2-10}$$

定义 2(调整兰德系数) 兰德系数是调整兰德系数的基础，然而该评价方法不能确保随机划分的聚类结果接近 0。因此，在原始的兰德系数的基础上提出了调整兰德系数（Adjusted Rand Index）。

调整兰德系数计算公式定义为

$$ARI=\frac{RI-E[RI]}{max(RI)-E[RI]} \tag{2-11}$$

为了更方便地计算ARI的值，我们使用列联表(contingency table),反映两类别划分的重叠程度。

如表 1 所示，实际的划分 X 有 m 类，用聚类方法得到的划分 Y 有 n 类。$x_i(i\in 1,\ldots,m)$表示实际划分 X 中的第 i 类；$y_j(i\in 1,\ldots,n)$表示用聚类方法得到的划分 Y 中的第 j 类,n_{ij}表示上述两类划分之间重叠实例的数量；α_i代表实际划分中第 i 类的样本数；b_j代表用聚类方法得到的划分中第 j 类的样本数。

根据上述的列联表可以通过下面的公式计算ARI

$$ARI=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_i\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_i\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_i\binom{b_j}{2}\right]/\binom{n}{2}} \tag{2-12}$$

表 1. 计算 ARI 的列联表.

x/y	y_1	...	y_n	a_i
x_1	n_{11}	...	n_{1n}	$\sum n_{1j}$
...	...	...	...	...
x_{m}	n_{m1}	...	n_{3n}	$\sum n_{mj}$
b_j	$\sum n_{i1}$	...	$\sum n_{in}$	

如图 4 所示，用基于中心连通性的聚类方法将鸢尾属植物数据集分成三类所计算出的ARI的值大于用$K-Means$聚类方法所计算出的ARI。调整兰德系数的比较说明基于中心连通性的聚类算法是可行的。

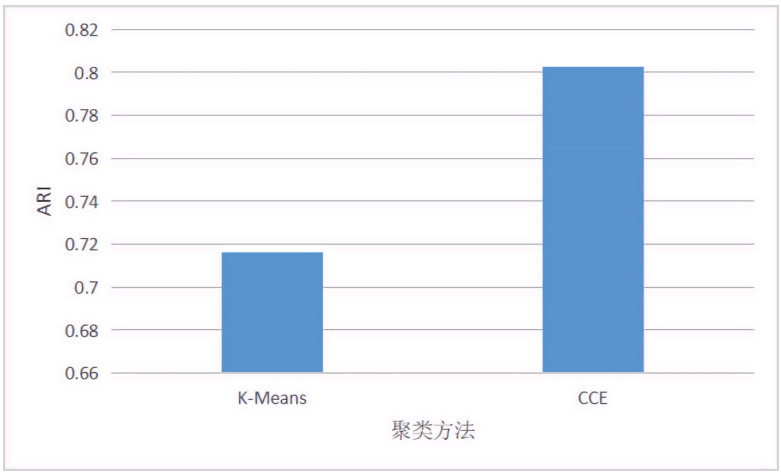


图 4.基于中心连通性的聚类算法($\sigma = 0.04$)与 K-Means 聚类算法性能对比

3. 基于中心连通性聚类方法的应用

3.1 高光谱以及波段选择

3.1.1 高光谱遥感

高光谱遥感[6]即高光谱分辨率遥感，是指使用很多非常窄的电磁波波段从感兴趣的物体获取相关数据。传统空间成像技术能够形成丰富的空间信息；目前先进的光谱测量技术能够获取连续的光谱数据。高光谱遥感将这两种技术相结合，使数据包含了许多精确的信息。在各个领域中，高光谱遥感已经取得了长足进步，并获得了许多成果。

高光谱遥感具有许多特点。在近红外光谱、可见光谱内成像光谱仪可以形成数十上百个波段，使高光谱遥感有多个波段；由于成像光谱仪的采样间隔在 $10nm$左右，间隔很小，能够使光谱的分辨率变得精细，从而提供给高光谱遥感细微的地物特征；高光谱遥感的数据量在随着波段数逐渐增加的情况下，呈指数增加；由于地物和成像特性，相邻的波段相关性很高，造成高光谱遥感信息冗余；高光谱遥感可提供光谱与空间的相应信息，由成像光谱仪得到的光谱曲线可以与地面实测的同类地物光谱曲线相类比。

我们可以用数据立方体（三维的数据图像）来表示高光谱数据集，相当于在采集到的普通二维图像以外增加了不同于图像两个维度的光谱信息。传统空间成像技术可以采集地物的二维图像能够反映出其二维空间的特征。在此基础上通过与光谱测量技术相结合，对二维图像增加光谱维的描述，共同构成高光谱数据。大部分地物都有相应的光谱特征，特别是光谱的吸收特征。根据上述地物对光谱测量所反映出的良好特性，我们可以通过将地物光谱与光谱数据库中的相匹配，从而实现地物识别。

高光谱图像将确定物质或者地物性质的光谱与表征其空间几何特征的图像结合在一起。并非所有物质的特征都在大的光谱范围内，与之相反，许多物质的特征集中体现在某些狭窄的光谱范围。高光谱遥感很好的保持了地物光谱特征、整体形态以及和周围地物的相关性。

3.1.2 高光谱波段选择

尽管高光谱图像相较于传统的空间图像能更准确的体现地表特征和地表之间的相互联系，但也存在一些技术难点。高光谱的波段数多，数据量大，并随着波段数逐渐增多而爆炸式的增长，非常容易出现棘手的维数灾难和修斯现象，导致高光谱图像的分类、识别等需要花费更多的时间和空间。由于相邻波段之间的相关性很强，从而导致高光谱信息冗余，数据存储需要花费很大的空间以及较长的时间。针对上述问题，通过降维处理高光谱数据来减少数据量和节省资源。在对高光谱图像进行降维处理时，常用到特征提取和波段选择这两类降维的方法。

利用特征提取的方法对高光谱图像进行降维处理时，使用的算法比较复杂，计算量相对较大，并且是通过某种对数据的变化来达到降维的目的，改变了高光谱图像原始数据，损失了数据之间的一些相关性。

与高光谱图像特征提取的方法相比较，波段选择的方法更加合理。波段选择是指从高光谱图像所有的波段中选择起主要作用的波段，有效的降低了高光谱图像的维度。因为是从原来的波段集合筛选出的波段子集，所以不会改变高光谱图像的原始数据。

高光谱图像的波段选择不同于特征提取，是一个复杂的组合优化的问题。波段选择需要从原来的波段集中，筛选出包含大量信息、波段之间具有较小的相关性和类别之间易于区分的波段组合。

3.1.3 高光谱聚类分析

高光谱聚类分析[7]广泛应用于高光谱图像的解释和信息的提取中。由于聚类方法本身的特点，高光谱聚类可以以无监督的方式来揭示像素的自然分割模式。

高光谱图像的解释通常依赖于大量的高质量的标记样本，以免由于训练样本不足而导致的休斯现象。在具体实践中，样本的采集通常十分消耗时间和人力物力，除此以外，在一些偏远的或者是没有人居住的地方，样本很难采集得到，这极大地限制了高光谱遥感的应用。因此，发展无监督的对地物识别的理论和方法来解决样本和先验知识的限制是十分有必要的。

聚类是一种十分有效的无监督信息提取和模式识别的方法。高光谱聚类是处理高光谱图像的常用手段，它用某些相似性度量例如距离、相关性、光谱角度等指标来表示高光谱图像的结构特性，将不相似的像素分离开，同时将其分配到某个对应的类。由于高光谱聚类是一种无监督的有效识别地物的方法，而且不需要标记的样本，与监督方法相比较，高光谱聚类更加节省样本标记所需要的成本，在很大程度上提高了高光谱遥感的应用潜力。

然而，高光谱图像的结构相较于手写图形、文本、自然图像等要复杂许多。不仅如此，考虑到复杂的成像环境，高光谱图像有光谱变异性，即出现“同物异谱”（相同的地物但光谱不同）或是“同谱异物”（相同的光谱但地物不同）。一般来说，在高维特征空间中，像素的分布相对均匀和稀疏，没有明确的规律。因此，高光谱聚类是一个十分有挑战性的任务。

3.1.4 常见高光谱聚类方法

常见的高光谱聚类方法有几类。基于质心的聚类方法，假设聚类在特征空间中具有球状结构，通过迭代最小化整体分区误差来实现高光谱聚类，如$K-Means$和FCM；基于密度的聚类方法，假设聚类是由特征空间中的稀疏区域分隔的密度点集，高光谱聚类是基于局部密度和相对像素距离的，如$CFSFDP$；基于概率的聚类方法，假设同一类的像素满足概率分布模型，高光谱图像基于相应的概率准则，如GMM；[8]基于仿生学的聚类方法，用一定的生物模型模拟高光谱图像的复杂内部结构，通过一些生物进化算法实现高光谱图像聚类，如SOM；基于智能计算技术的聚类方法，基于其他的聚类模型，利用先进的智能计算算法寻找聚类模式的全局最优解来对高光谱图像聚类，如$FCIDE$；基于图的聚类算法，利用邻接矩阵对像素之间的相似性进行建模，用图切的算法对高光谱图像进行聚类，如SC；基于子空间的聚类，通过子空间的并集来对高光谱图像内部复杂结构进行建模，通过自适应学习来探索像素之间的潜在邻接关系，利用得到的邻接关系通过谱聚类来对高光谱图像进行聚类；基于深度学习的聚类方法，依赖于深度神经网络来学习更多的特征，更准确的模拟高光谱数据非线性关系来对高光谱图像进行聚类。

接下来讨论将基于中心连通性的聚类方法应用在高光谱图像的波段选择上。

3.2 将基于中心连通性聚类算法应用于高光谱图像的波段选择

首先构造高光谱数据的特征相似矩阵，再利用基于中心连通性的聚类算法对得到的特征相似矩阵进行迭代，从不同的观测规模得到不同的聚类结果以及聚类中心。

3.2.1 构造特征相似矩阵

利用高光谱的所有波段构建一个二维矩阵$W=[w_1,w_2,\dots,w_N]\in R^{M\times N}$，其中$M$代表高光谱数据的像素个数，$N$代表高光谱数据的波段数量，$w_k$为第$k$个波段所对应的向量。利用欧氏距离计算出距离矩阵$D$。

$$DI=\begin{pmatrix} di_{11} & \cdots & di_{1N} \\ \vdots & \ddots & \vdots \\ di_{N1} & \cdots & di_{NN} \end{pmatrix} \tag{3-1}$$

其中d_{ij}代表第i个波段到第j个波段的欧氏距离。

使用高斯核函数(H_{Gauss}: $s_{ij} = exp\left(-\| w_i - w_j\|^2/\sigma^2\right)$)构造相似性矩阵$S$:

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1N} \\ \vdots & \ddots & \vdots \\ s_{N1} & \cdots & s_{NN} \end{pmatrix} \tag{3-2}$$

在许多情况下对数据集聚类，每个类的样本的数量都是不相同的。基于这一点，在不同类别间，聚类数目相差非常大的时候十分容易出现“大簇吞小簇”(聚类数量较少的簇可能会更快地合并到聚类数量较多地簇)的现象。为了解决”大簇吞小簇”这个问题，我们可以对相似性矩阵S进行正规化:

$$\tilde{S} = D^{-1/2} S D^{-1/2} \tag{3-3}$$

其中矩阵$D = diag(d_1, d_2, \ldots, d_N)$是相似性矩阵$S$的度矩阵，$d_i = \sum_{j=1}^{n} s_{ij}$是第$i$个点的度。

3.2.2 对相似性矩阵进行迭代

计算出高光谱数据集的相似性矩阵后，使用基于中心连通性的聚类算法，对相似性矩阵进行迭代。

$$\tilde{S}^{(k)} = \tilde{S}^{(k-1)} \tilde{S} \tag{3-4}$$

$\tilde{S}^{(k)}$表示经过第k次迭代的相似性矩阵。根据相似性矩阵，筛选出聚类中心同时将其他数据点分配给聚类中心。

3.2.3 选择聚类中心和分配聚类

对于每一次的迭代，根据基于中心连通性的聚类方法，将满足

$$con^{(k)}(w_i, w_i) > con^{(k)}\left(w_i, w_j\right), j = 1, \ldots, N(j \neq i) \tag{3-5}$$

的数据点i确定为聚类中心。对于剩余其他的数据点，通过计算他们到各聚类中心的k阶相对连通性，来决定分配到哪一个聚类。

假设在第k次迭代的时候，选择出了m个聚类中心$w_{c_i}\left(c_i \in \{1,2,\ldots,N\}同时 i = 1,2,\ldots,m\right)$，对于非聚类中心的数据点$w_j$，它将会被分配到聚类中心$w^*$,$w^*$满足下面的等式：

$$w^* = \text{argmax}_{w_{c_i}}\left(rcon^{(k)}\right)\left(w_{c_i}, w_j\right) \tag{3-6}$$

每一次选择出的聚类中心w_{c_i}就认为是波段选择的集合。

3.2.4 印度松数据集(Indiana Pines Dataset)

印度松树数据集的场景是由位于印第安纳州西北部的印度松树试验场上方的 AVIRIS 传感器收集的。145×145 像素和波长范围为 $0.4 - 2.5 \times 10^{-6}$的 224 个光谱反射带共同组成了原始的印度松数据集。此场景是更大场景的一个子集。印度松场景包括了农田、森林和其他天然植物。除此以外，该场景还包括了四车道的高速公路和一条铁路，一些低密度的房屋以及其他建筑和小路。本文所使用的数据集是去除一些噪声带的印度松数据集。

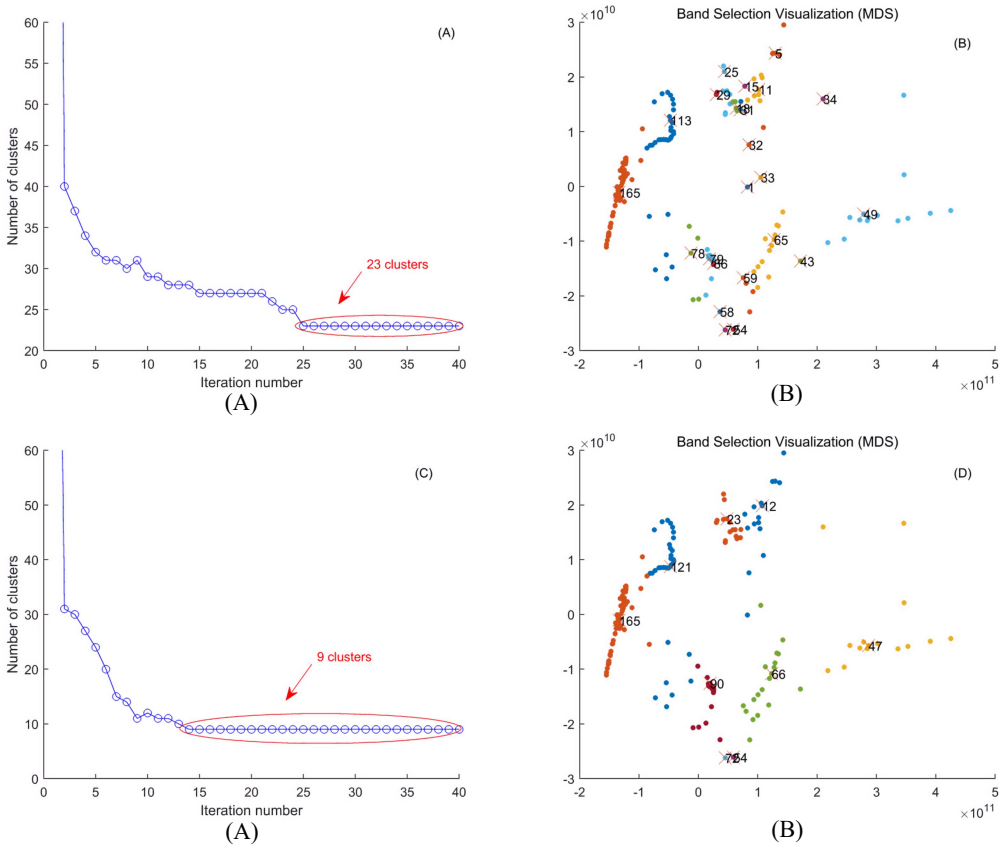


图 5.选取不同的参数大小对印度松数据集的波段选择.

如图 5 所示，在印度松数据集上使用基于中心连通性的聚类方法来进行波段选择。图 5(A)在参数σ = 0.03 下不同迭代次数得到的聚类中心，并且稳定在 23 个聚类中心的迭代次数最多。图 5(B)展示了所选择出的 23 个波段。图 5(C)在参数σ = 0.05 下不同迭代次数得到的聚类中心，并且稳定在 9 个聚类中心的迭代次数最多。图 5(D)展示了所选择出的 9 个波段。由于通过高斯核函数来构造，参数σ越大，样本在同等距离度量下，相似性表现得更强，表现在聚类收敛得更快。

3.3 使用基于中心连通性聚类方法进行波段选择方法的改进

3.3.1 边缘检测

边缘检测[9]是图像处理中一种重要的方法。它通过一些数学方法来标记等待处理的图像中明亮变化明显的点，这是边缘检测算法的内在核心。通常来说，数据样本或者是事物的属性发生的重大改变会反映在图像属性的明显变化上。这些重要的变化通常包括了深度的连续性（如物体在不同的平面上）、表面方向的连续性（如立方体不同的表面）、物质的属性（如不同地物对光的反射不同）变化或者场景照明的变化（如被建筑物遮挡的地面）等。

对待处理得图像进行边缘检测处理可以减少原有复杂繁重的数据量，并且减少了一些边缘检测算法认为不相关的信息，最大限度地保证了图像的结构和图像数据相关属性，类似于对图像进行预处理，使得后续对图像得处理变得轻松。

利用边缘检测的特性，可以先对高光谱图像进行分组，再利用基于中心连通性的聚类方法分别对高光谱图像的每一组数据进行聚类。

3.3.2 对高光谱数据集进行边缘检测

对高光谱数据集进行边缘检测的步骤和有关概念解释如下：

步骤 1 利用高光谱的所有波段构建一个二维矩阵$W = [w_1, w_2, \ldots, w_N] \in R^{M\times N}$，其中$M$代表高光谱数据中，每一个二维图像的像素个数，$N$代表高光谱数据的波段数量，$w_k$为第$k$个波段所对应的向量。利用欧氏距离计算出高光谱波段之间的距离矩阵D。

$$DI = \begin{pmatrix} di_{11} & \cdots & di_{1N} \\ \vdots & \ddots & \vdots \\ di_{N1} & \cdots & di_{NN} \end{pmatrix} \tag{3-7}$$

其中d_{ij}代表第i个波段到第j个波段的欧氏距离。

得到高光谱距离矩阵的缩放颜色显示图像(RGB 图像)$P \in R^{N\times N\times 3}$。其中$P_{ijk}$表示第$i$行第$j$列像素点的第$k$个颜色通道的数值。$k = 1$ 为 RGB 颜色通道中的R(红色)通道，$k = 2$ 为 RGB 颜色通道中的G(绿色)通道,$k = 3$ 为 RGB 颜色通道中的B(蓝色)通道。

步骤 2 对高光谱距离矩阵的缩放颜色显示图像进行灰度处理,得到灰度图像G.

$$G_{ijk} = P_{ij1} \times 0.299 + P_{ij2} \times 0.587 + P_{ij3} \times 0.114 \tag{3-8}$$

在图像的生成、收集和传输的过程中，要避免引入图像噪声是十分困难的。一些在图像中冗余或者干扰原数据的干扰信息被称作是图像噪声。在我们对图像数据进行信息提取或特征提取的过程中，图像噪声会影响我们的提取过程。因此，我们需要通过一些算法来消除图像噪声的影响。常见的去除图像噪声方法有许多，例如高斯滤波、均值滤波、中值滤波[10]等。

滤波是通过滤波器(3 × 3 或者 5 × 5 的矩阵)对图像进行从上到下，从左至右地扫描，通过滤波器计算与其对应像素点的值，根据不同的滤波器进行不同的计算，然后将计算结果赋值回当前的像素点。本文使用中值滤波处理高光谱距离矩阵的缩放颜色显示图像的灰度图像。

步骤 3 在滤波器遍历图像时，当前待处理的像素点为$I_{r,c}$，它的周围像素所组成的矩阵为

$$\begin{pmatrix} I_{r-1,c-1} & I_{r-1,c} & I_{r-1,c+1} \\ I_{r,c-1} & I_{r,c} & I_{r,c+1} \\ I_{r+1,c-1} & I_{r+1,c} & I_{r+1,c+1} \end{pmatrix} \tag{3-9}$$

将九个像素进行排序，选出中值I_{median}，并赋值给$I_{r,c}$。对所有像素点(除边缘像素点之外)进行中值滤波，最终得到中值滤波后的灰度图像I。

高光谱距离矩阵的缩放颜色显示图像P经过灰度处理和中值滤波处理之后，得到图像I。接下来对图像I进行边缘检。通过一阶导数或者是二阶导数可以实现灰度变化的检测，从而找到图像的边缘。在本文中，使用梯度来寻找边缘的强度和方向。

图像是二维的，二维函数的一阶偏微分方程：

$$\frac{\partial f(x,y)}{\partial x} = \lim_{\varepsilon\to 0}\frac{f(x+\varepsilon,y)-f(x,y)}{\varepsilon} \tag{3-10}$$

$$\frac{\partial f(x,y)}{\partial y} = \lim_{\varepsilon\to 0}\frac{f(x,y+\varepsilon)-f(x,y)}{\varepsilon} \tag{3-11}$$

在二维图像中，图像的梯度(gradient)是当前像素点对X轴和对Y轴的偏导数。梯度的模则表示$f(x, y)$在其最大变化率方向上的单位距离所增加的量:

$$G[f(x,y)] = \left[\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]^{\frac{1}{2}} \tag{3-12}$$

根据上述描述，如果要得到一幅图像的梯度，就要获得图像中每一个像素的梯度,即计算图像每一个像素点$f(x,y)$在(x,y)位置处的x方向上的梯度大小和y方向上的梯度大小。为了更方便地表示和计算梯度，我们通常使用梯度算子。梯度算子是用来计算梯度偏导数的一种滤波器模板，也有人称其为边缘算子或者边缘检测算子。

图像的边缘检测算法也是利用了滤波器的原理，与上面描述的处理灰度图像的中值滤波算法的过程类似，不同的是对需要处理的图像像素遍历时所使用的滤波器不相同。同时，设置一个阈值，如果一个像素点可以通过 Prewitt 算子进行计算结果出来的值大于阈值，则可认为是边缘点。本文使用 Prewitt 算子对处理后的灰度图像进行边缘探测。

用于对灰度图像边缘检测的滤波器模板 Prewitt 算子如下所示：

$$d_y = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \; d_x = \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \tag{3-13}$$

步骤 4 使用上述 Prewitt 算子的模板对需要处理的像素点$I_{r,c}$进行不同方向的梯度计算：

$$\frac{\partial f(x,y)}{\partial y} = \left(I_{r-1,c+1} + I_{r,c+1} + I_{r+1,c+1}\right) - \left(I_{r-1,c-1} + I_{r,c-1} + I_{r+1,c-1}\right) \tag{3-14}$$

$$\frac{\partial f(x,y)}{\partial x} = \left(I_{r-1,c+1} + I_{r,c+1} + I_{r+1,c+1}\right) - \left(I_{r-1,c-1} + I_{r,c-1} + I_{r+1,c-1}\right) \tag{3-15}$$

得到$PR_{r,c} = G[f(x,y)]$,让后将此结果与阈值T比较，得到边缘。对所有像素点(除边缘像素点外)使用 Prewitt 算子进行计算，最终得到边缘检测后的图像。

通过边缘检测，可以首先将高光谱数据进行分组。

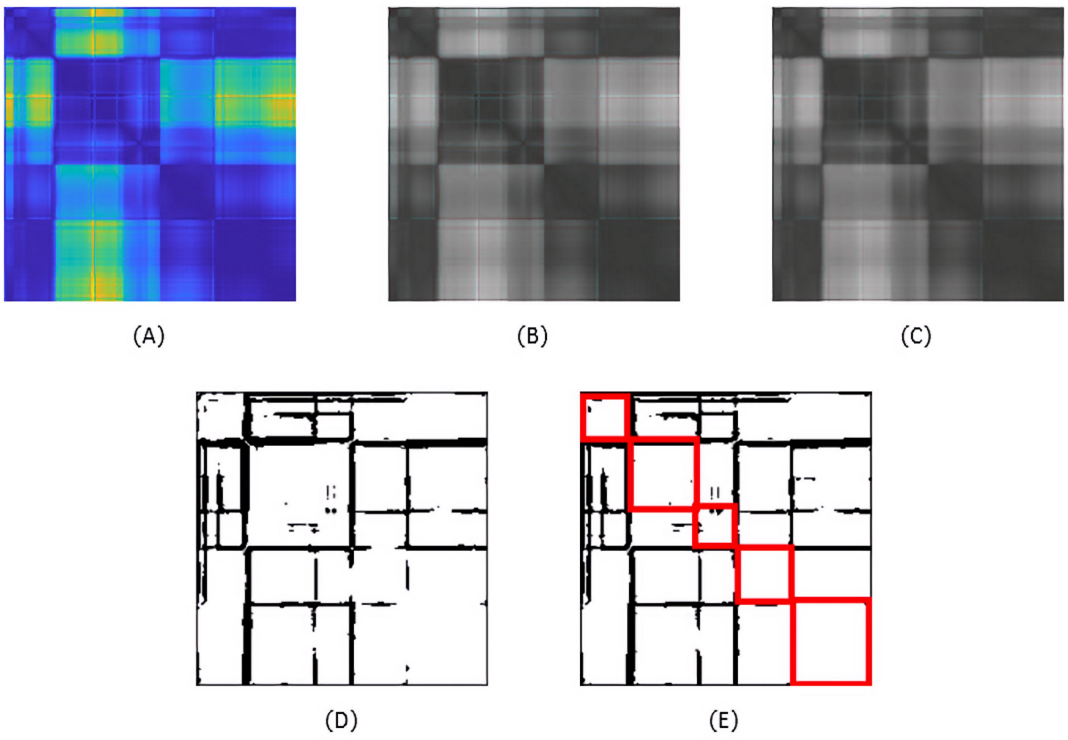


图 6.对高光谱数据进行边缘探测.

如图 6 所示，首先对印度松数据集依次进行获取距离矩阵的缩放颜色显示图像(图 6(A))，其次得到灰度图像(图 6(B))。然后对灰度图像进行中值滤波（图 6(C)）。最后对中值滤波后的灰度图像使用 Prewitt 算子进行边缘探测(图 6(D))，得到波段分组(图 6(E))。

3.3.3 将改进后的基于中心连通性的聚类算法应用在不同的数据集

接下来我们在 Indian Pines，PaviaU 和 Salinas 数据集上应用改进后的算法。各数据集的影像像素和波段数如表 2 所示。

表 2.三个高光谱数据集的相关信息

数据集	影像像素	波段数
Indian Pines	145×145	185
PaviaU	610×340	103
Salinas	512×217	204

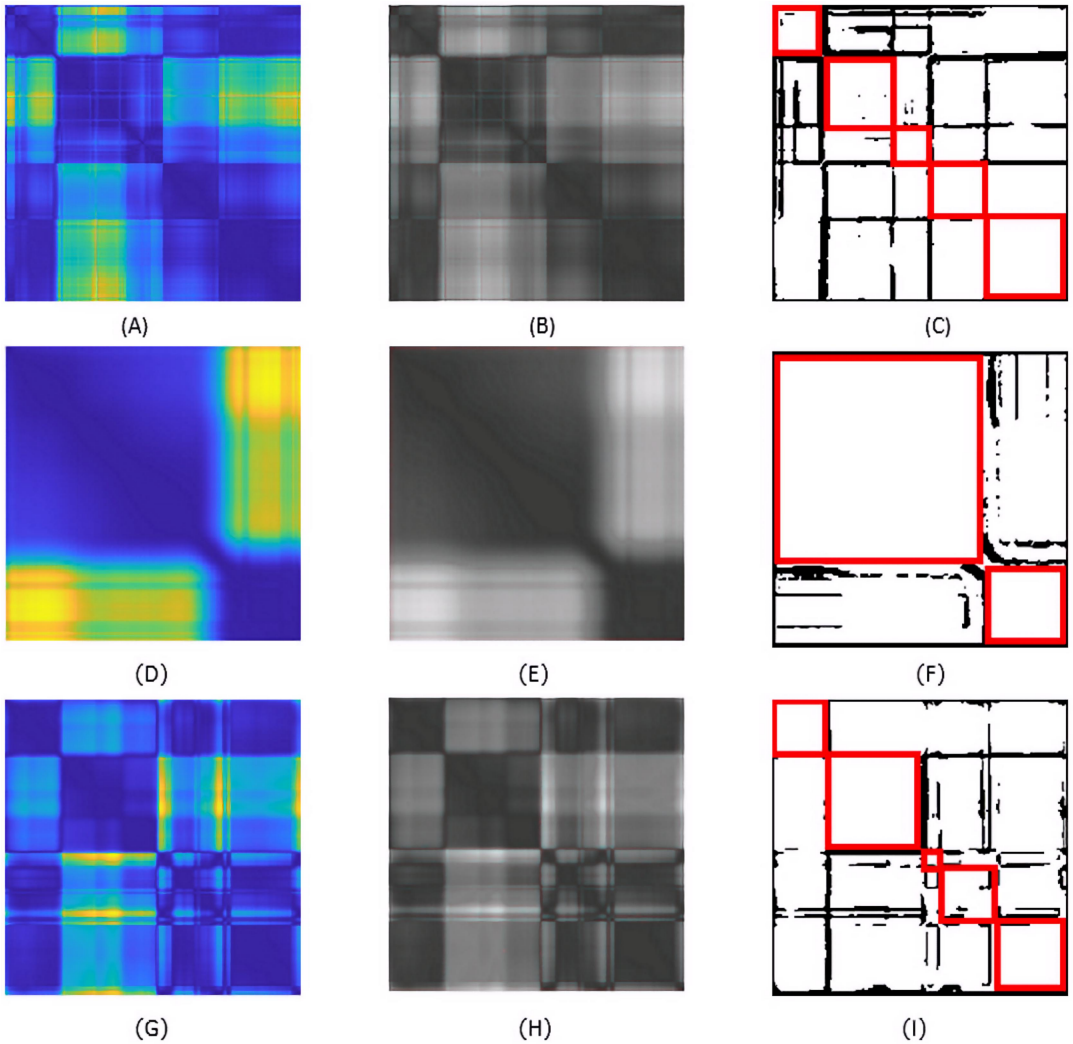


图 7. 对各高光谱数据边缘探测.

如图 7 所示，图 8(A)、(B)、(C)表示 Indian Pines 数据集边缘探测过程；图 8(D)、(E)、(F)表示 PaviaU 数据集边缘探测过程；图 8(G)、(H)、(I)表示 Salinas 数据集边缘探测过程；

对高光谱数据进行边缘探测之后，可以获取相应的波段分组。分组情况如表 3 所示。

表 3. 各高光谱经边缘探测后分分组情况

数据集	分组情况
Indian Pines	[1,34][35,78][79,100][101,134][135,185]
PaviaU	[1,74][75,103]
Salinas	[1,39][40,105][106,115][107,158][159,204]

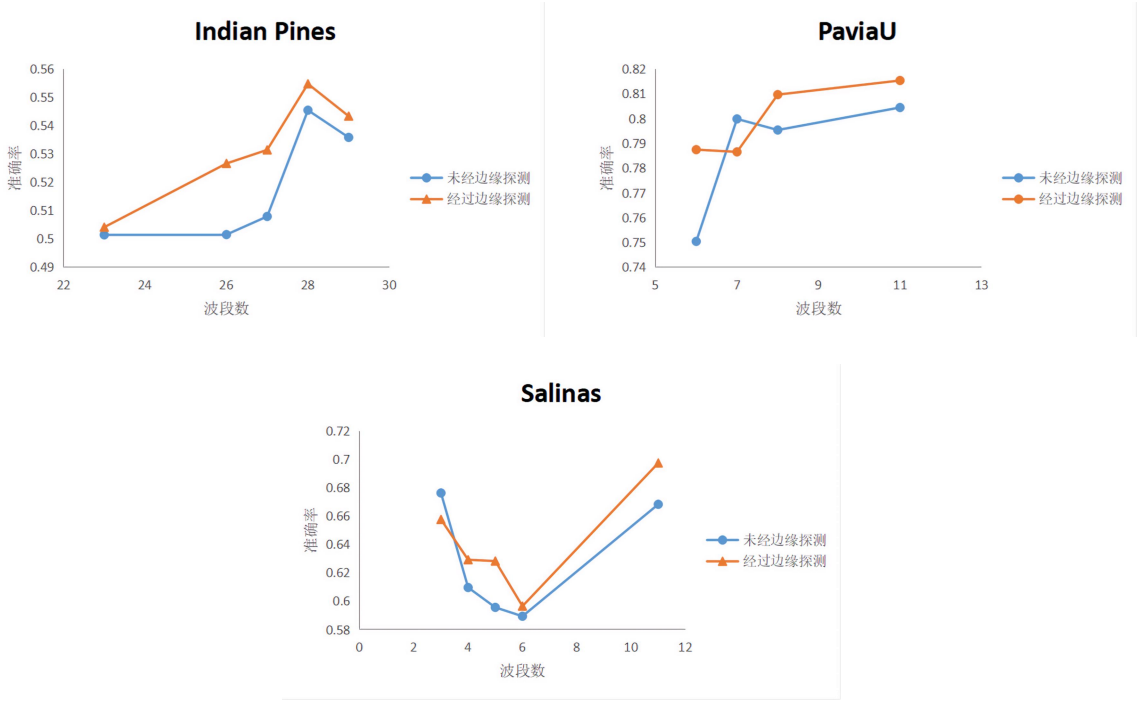


图 8. 各高光谱数据经边缘探测和未经边缘探测对比

如图 8 所示，分别对各高光谱数据使用边缘探测后的基于中心连通性的聚类算法和直接使用基于中心连通性聚类算法。其中固定各数据集的参数:Indian Pines,$\sigma = 0.04$; PaviaU,$\sigma = 0.1$; Salinas,$\sigma = 0.05$;利用 svm 分类器，通过对比可以发现，在选择到相同数量的波段时，使用边缘探测后的基于中心连通性的聚类算法总体的准确率要高于直接使用基于中心连通性的聚类算法。

4.结论

基于中心连通性的聚类算法是一种基于图论的聚类算法，目标是对于具有n个样本的集合$X \in R^{n\times d}$,首先通过某种相似性度量，构造样本之间的相似性矩阵。例如在本文中，在处理鸢尾属植物数据集时，使用高斯核函数来构造样本之间的相似性矩阵。通过对相似性矩阵的迭代，可以动态的得到不同观察规模时的聚类。基于中心连通性的聚类算法的优点是算法简单、算法时空复杂度不高(仅需对矩阵进行迭代相乘)，相比于$K-Means$等传统聚类算法，此算法可以动态观察不同规模的聚类，不需要预先规定簇数。然而，基于中心连通性的聚类算法也存在几个缺点。此算法是基于图论的，所以受相似性矩阵的影响十分大。如果相似性度量等选择不够好，就会使该算法的使用大打折扣。

除此之外，在对高光谱使用基于中心连通性的聚类方法时，通过与边缘探测相结合，能够提高波段选择的质量。

【参考文献】

[1]J.A. Hartigan, M.A. Wong. Algorithm as 136: a k-means clustering algorithm[J]. Appl.Stat, 1979:100-108.

[2]Y. Cheng. Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal Mach. Intell, 1995(17): 790－799.

[3] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996:226－231.

[4] Edgar Anderson. The irises of the Gaspé Peninsula. Bulletin of the American Iris Society, 1935(59): 2－5.

[5] L. Hubert, P. Arabie. Comparing partitions. J. Classif, 1985(2):193－218.

[6] 杜培军, 夏俊士, 薛朝辉, 谭琨, 苏红军, 鲍蕊. 高光谱遥感影像分类研究进展. 遥感学报, 2016(020) :236-256.

[7] HAN ZHAI, HONGYAN ZHANG, PINGXIANG LI, AND LIANGPEI ZHANG. Hyperspectral Image Clustering.

[8] N. Acito, G. Corsini, and M. Diani. An unsupervised algorithm for hyperspectral image segmentation based on the Gaussian mixture model. Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2003(6):3745－3747.

[9]段瑞玲, 李庆祥, 李玉和. 图像边缘检测方法研究综述. 光学技术 ,2005,31(003):415-419.

[10]高浩军, 杜宇人. 中值滤波在图像处理中的应用[J]. 信息化研究, 2004, 30(008):35-36.

致谢

大学的时光已经接近尾声，回想这段经历，觉得感慨万分。首先要感谢的是我的指导老师，贾森老师。从论文的选题、实践再到论文的撰写这整个过程中，贾森老师给了我很多建议，耐心指导我如何去实践、如何去写好一篇论文。在这段时间里，我学会了如何更有效的查阅书籍、独立思考以及平衡时间的冲突。

此外，还要感谢我的家人在我编写论文时给我包容、关爱和鼓励，感谢我的舍友和朋友给我提供浓厚的学习氛围。

再次感谢所有给予我帮助的老师、家人和朋友们，你们的帮助让我获益良多，受益终生。

Clustering by connection center evolution

【Abstract】With the progress of the times and the development of science and technology, the tide of informatization has swept the world. The massive amount of data has never been so closely related to our daily lives, and has never been as active as it is today. Data connects people and the world, and everyone is creating and using data. Therefore, the processing of data is particularly important. As a kind of unsupervised learning algorithm, clustering method has developed rapidly in recent years and has made great progress, and it is applied in many scenarios.

This article focuses on clustering by connection center evolution. The clustering method based on center connectivity is a clustering algorithm based on graph theory. It first constructs the similarity relationship between samples, and then finds the cluster centers in an iterative manner, as well as the allocation of non-cluster center samples.

The main research work of this paper is as follows:

1. Briefly introduce the origin, basic ideas and content of clustering, explain the similarity measurement in the clustering method, and give a detailed explanation of common clustering algorithms.

2. Explain in detail the definition and clustering process of the clustering method named clustering by connection center evolution(CCE). By clustering the general data set and comparing it with common clustering algorithms, it shows the feasibility.

3. Based on the previous description of CCE, this algorithm is applied to the hyperspectral bands selection. At the same time, through the combination of edge detection and this algorithm, the performance of the algorithm is improved.

【Keywords】Clustering; Graph; Clustering by connection center evolution; Hyperspectral image; Edge detect