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】随着时代的进步和科学技术的发展，信息化大潮席卷全球。海量的数据同我们日常生活的联系从未如此紧密过，也从没有像今天那么活跃过。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。因此，对数据的处理显得尤为重要。聚类方法作为一种无监督学习的算法，在近年来发展迅速并取得了长足的进步，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基于图论的聚类算法，首先要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.简单地介绍聚类的起源、基本思想以及内容，解释了聚类方法中的相似性度量，并对常见的聚类算法进行细致说明。

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. 绪论

聚类分析起源于分类学，在古老的分类学中，人们主要依靠经验和专业知识来实现分类，很少利用数学工具进行定量的分类,致使许多分类带有主观性和任意性，不能很好地揭示客观事物内在的本质区别与联系，特别是对于多个特征、多指标的分类问题更难以实现客观的分类。随着人类科学技术的发展，对分类的要求越来越高，于是人们逐渐地把数学工具引用到了分类学中，形成了数值分类学，之后又将多元分析的技术引入到数值分类学形成了聚类分析。

聚类分析在许多领域中都得到了广泛地应用，取得了许多成果。

1.1 聚类分析的基本思想

聚类分析认为，所研究的样本之间存在不同程度的相似性。根据多个样本的多个特征，找出能够表示样本之间的相似性的度量，并且根据这种度量采用某些聚类方法，将所有的样品分配到不同的种类当中去，使同一种类中的样本具有较大的相似性，不同类中的样本相似性小。我们将这种分类方法称为聚类分析。

1.2 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。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。

每种聚类分析方法都涉及事物之间的相似性。聚类分析方法的本质就是寻找一个客观能反应样本之间相关联系的统计量，然后根据这种统计量把样本分成若干类，常用的统计量有距离和相似系数。

1.2.1 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：

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

绝对距离：

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

Minkowski 距离：

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

1.2.2 相似性度量——相关系数

对 n 个样本进行聚类的时候，用相似系数来衡量变量之间相关联程度。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：

$$|c_{\alpha\beta}| \le 1 \text{ 且} c_{\alpha\alpha} = 1$$

$$c_{\alpha\beta} = c_{\beta\alpha}$$

$|c_{\alpha\beta}|$越接近 1，说明x_α和x_β越相关，相似系数中最常用的是相关系数和夹角余弦。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。

1.3 常见的聚类算法

1.3.1 K-Means 聚类算法

给定样本集$D=\{x_1,x_2,\ldots,x_m\}$，按照样本之间的距离大小，划分成k个簇$C=\{C_1,C_2,\ldots,C_k\}$，经过N次迭代，让簇内的点紧密联系，而簇间的距离尽可能大。

$K-Means$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$D$中的各样本点$x_i(i=1,2,\ldots,m)$,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$C_j$中所有的样本点重新计算质心$\mu_j=\frac{1}{|C_j|}\sum_{x\in C_j}x$,如果所有的 k 个质心向量都没有发生变化，则输出簇划分$C=\{C_1,C_2,\ldots,C_k\}$，否则继续迭代。

K-Means 聚类算法的优点是速度快，计算简便，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。

1.3.2 DBSCAN

DBSCAN(具有噪声的基于密度的聚类方法)是一种基于密度的空间聚类算法。这类密度算法通过样本分布的紧密程度进行分类，即将具有足够密度的区域划分为簇，并在具有噪声的空间数据样本中发现任意形状的簇，它将簇定义为密度相连的点的最大集合。

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：样本x_i的半径Eps的邻域内至少包含$MinPts$个样本，则称x_i为核心点。

边界点：样本x_i的半径Eps的邻域内包含的样本点数目小于$MinPts$，但是它在其他核心点的邻域内，则称样本x_i为边界点。

噪声点：样本x_i的半径Eps的邻域内包含的样本点数目小于$MinPts$，同时x_i也不在其他核心点的邻域内，则称样本x_i为噪声点。

$DBSCAN$算法步骤如下：

步骤 1 给定样本集$D=\{x_1,x_2,\ldots,x_m\}$,邻域参数$(Eps,MinPts)$,初始化核心点集合$\Omega=\emptyset$,初始化聚类簇数$k=0$,初始化未访问样本集合$\Gamma=D$,簇划分$C=\emptyset$。

步骤 2 对于$j=1,2,\ldots,m$，寻找核心点。根据核心点的定义，如果x_j是核心点，则将x_j加入核心点集合$\Omega=\Omega\cup\{x_j\}$.

步骤 3 如果核心点集合$\Omega=\emptyset$,则算法结束，否则进入步骤 4。

步骤 4 在核心点集合Ω中，选择一个核心点o,初始化当前簇核心点集合$\Omega_{cur} = \{o\}$,初始化簇类序号$k = k + 1$,初始化当前样本集合$C_k = \{o\}$,更新未访问样本集合$\Gamma = \Gamma - \{o\}$.

步骤 5 如果当前簇核心点集合$\Omega_{cur} = \emptyset$,则当前聚类簇C_k生成完毕，更新簇划分$C = \{C_1, C_2, \ldots, C_k\}$,更新核心点集合$\Omega = \Omega - C_k$,转入步骤 3，否则到步骤 6

步骤 6 在当前簇核心点集合Ω_{cur}中取出一个核心点o',通过Eps找出邻域内所有样本集$N(o')$,令$\Delta = N(o') \cap \Gamma$,更新当前簇样本集合$C_k = C_k \cup \Delta$,更新未访问样本集合$\Gamma = \Gamma - \Delta$,更新$\Omega = \Omega \cup (\Delta \cap \Omega) - o'$,转入步骤 5

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

DBSCAN 的优点是不需要预先声明需要划分簇类的数目k,并且可以对任意形状稠密的数据集进行分类；在聚类的同时，可以发现异常点（噪声点），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，当空间聚类的密度不均匀，参数Eps和$MinPts$选取困难；当数据集较大的时候，聚类收敛的时间会相对较长；相对于传统的 K-Means 聚类算法稍显复杂，因为需要调整参数Eps和$MinPts$,不同的参数组合对实际的聚类结果影响较大。

1.3.3 均值漂移聚类

均值漂移聚类算法是以滑动窗口为基础的聚类算法，来找到数据点密集的区域。这是一个基于质心的算法，通过移动中心点(将中心点的候选点更新为滑动窗口里面点的均值)，定位每一个簇类的中心点。对这些候选的窗口进行筛选，最终形成中心点集以及相应的分组。

相关定义

$Mean\ Shift$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$Mean\ Shift$向量的基本形式为：

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

其中，S_h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：

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

k表示在S_h范围内样本点的个数。

中心更新：将中心点移动到偏移均值的位置。

$$x^{t+1} = M^t + x^t$$

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：核函数是用来计算映射到高维空间之后的内积的一种简便方法，目的是为了让低维不可分的数据变成高维可分。利用核函数，可以忽略映射关系，在低维空间中完成计算。在$Mean\ Shift$聚类算法中引入核函数的目的是使随着样本与被偏移点的距离不同，其偏移量对$Mean\ Shift$向量的贡献也不同。核函数是机器学习中常用的一种方式，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，$x = x_1, x_2, x_3, \ldots, x_d$，$R$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

$$K(x) = k(\| x\|^2)$$

并且同时满足k是非负的、非增的和k是分段连续的，那么函数$K(x)$就称为核函数。核函数有很多，常见的核函数有线性核函数、多项式核函数和高斯核函数等。高斯核函数定义如下：

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

其中，h称为带宽(bandwith)。在高斯核函数中，当带宽h一定时，样本点之间的距离越近，其核函数的值越大，当样本点之间的距离相等时，随着高斯核函数的带宽h的增加，核函数值减小。

引入核函数的$Mean\ Shift$向量：在均值漂移中引入核函数的概念，能够让计算中距离中心的点具有更大的权值，反映距离越短，权值越大的特性。改进的$Mean\ Shift$向量定义如下：

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

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，可以不需要设置簇类的个数，也可以处理任意形状的的簇类；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，均值漂移聚类的结果取决于带宽的设置。带宽设置得太小会导致收敛慢，簇类个数较多；带宽设置得太大，一些簇类可能会丢失。而且相较于较大的特征空间，计算量会十分大。

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

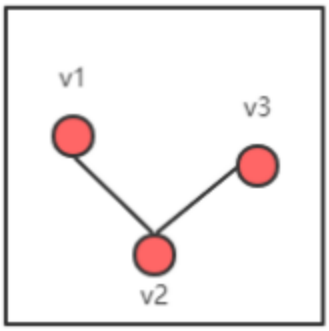
基于中心连通性的聚类方法是基于图论游走次数理论的。唯一不同的是，在基于中心连通性的聚类方法当中，在无向图中的每一个顶点都是自连的。

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。

假设有无向图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，详情可见**表 1**

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，A_k中的$a_{ij}^{(k)}$代表v_i到v_j步长为k的路径数。因为这里的每个顶点都是自连通的，所以邻接矩阵A的对角元素值都为 *1*

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

$$con^{(k)}(v_i, v_i) > con^{(k)}(v_i, v_j), j = 1,\dots, n(j \neq i)$$

在**图 1** 中，经过简单计算，可以得出A^3的聚类中心为v_2

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

$$rcon^{(k)}(v_i, v_j) = con^{(k)}(v_i, v_j)/con^{(k)}(v_i, v_i)$$

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$G = \{v_1, v_2,\dots, v_n\}$，我们将其分割为$m$个子图$A_1, A_2,\dots, A_m$，并满足以下条件

$$A_1 \cup A_2 \cup \dots \cup A_m = G$$

$$A_i \cap A_j = \emptyset(i, j = 1,2,\dots, m\text{且}i \neq j)$$

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

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

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

$$cut(A_1, A_2,\dots, A_m) = \sum_{1}^{m} W(A_i, \bar{A_i})$$

$\bar{A_i}$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

$$Ncut(A_1, A_2, \dots, A_m) = \sum_{1}^{m} \frac{W(A_i, \bar{A}_i)}{Vol(A_i)}$$

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$v_{c_i}(c_i \in \{1,2,\dots,n\}$同时$i = 1,2,\dots,m)$，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：

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

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，通过定义 2 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**(1)**和**(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。

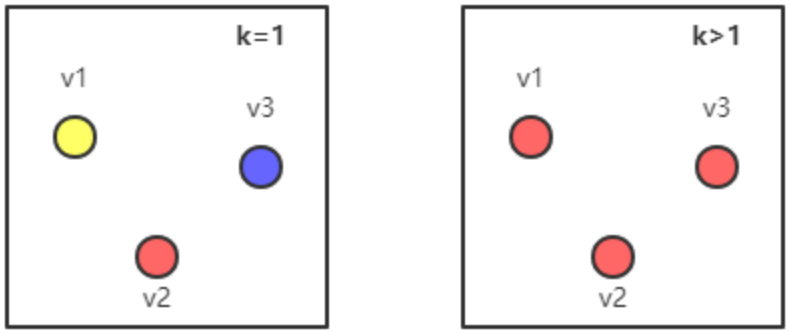


图 2.

2.4 可行性

我们使用基于中心连通性的聚类方法对**鸢尾属植物数据集**[4]进行聚类

2.4.1 鸢尾属植物数据集

鸢尾属植物数据集(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。

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

所得到的相似矩阵是实对称矩阵。**图 3** 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**图 3** 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。在**图 4** 中，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 4** 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。

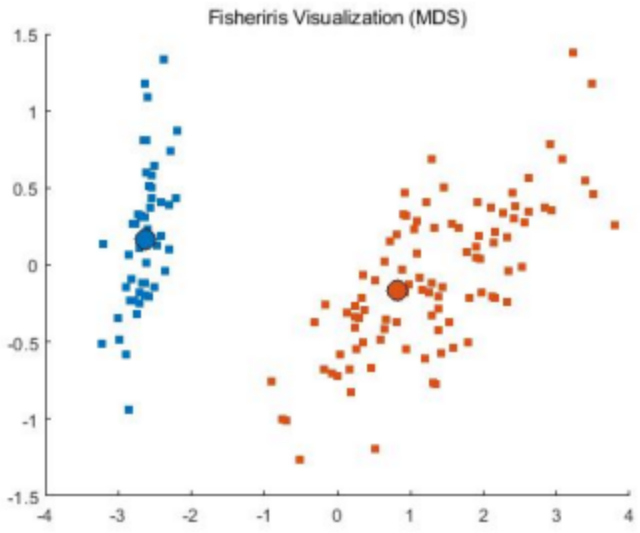


图 3.基于中心连通性的聚类算法将鸢尾属植物数据集分成两类

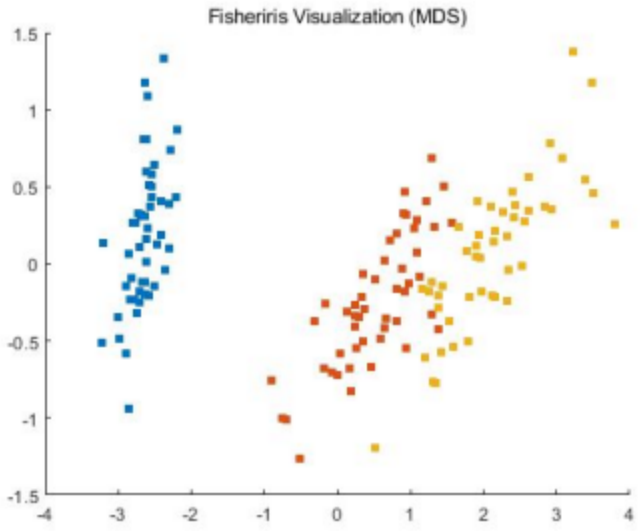


图 4.鸢尾属植物数据集根据样本标签分类

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。使用这个度量指标需要数据本身有标记类别。调整兰德系数是一个标量，范围在$[-1,1]$之间，反映了两种划分的重叠程度。它的数值越大，说明聚类的效果越好。

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)的计算公式为

$$RI = \frac{a + d}{a + b + c + d}$$

兰德系数无法保证随机划分的聚类结果的值接近 0，所以提出了调整兰德系数(Adjusted Rand Index)。

调整兰德系数计算公式为

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

为了计算ARI的值，我们引入列联表(contingency table),反映实际类别划分与聚类所得到的划分的重叠程度。表的行表示实际的类别划分，列表示聚类划分的簇标记,n_{ij}表示重叠实例的数量。

如**表 2**。

x/y	y_1	y_2	y_3	a_i
x_1	n_{11}	n_{12}	n_{13}	$\sum n_{1j}$
x_2	n_{21}	n_{22}	n_{23}	$\sum n_{2j}$
x_3	n_{31}	n_{32}	n_{33}	$\sum n_{3j}$
b_j	$\sum n_{i1}$	$\sum n_{i2}$	$\sum n_{i3}$	

通过列联表计算ARI

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

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，如**图 5**。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。

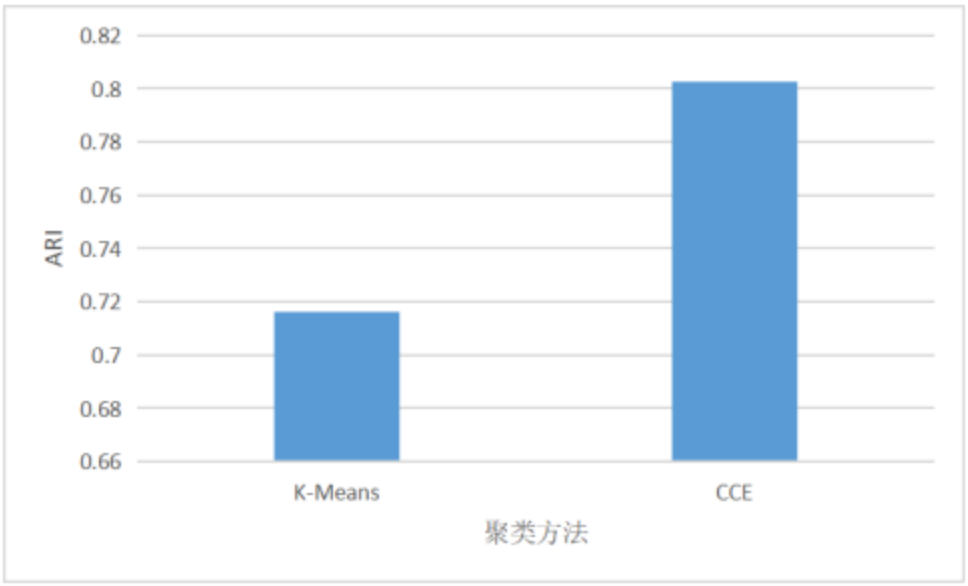


图 5.基于中心连通性的聚类算法应用于鸢尾属植物数据集的表现要比 K-Means 聚类算法好

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，指的是利用很多很窄的电磁波波段，从感兴趣的物体获取有关数据，将传统的空间成像与先进的光谱测量技术有机结合，具有连续光谱数据和丰富的空间信息。高光谱遥感目前已经广泛应用于各个领域。

高光谱遥感具有以下特点：波段多，成像光谱仪在可见光和近红外光谱区内有数十甚至数百个波段；光谱分辨率高，成像光谱仪采样的间隔小，一般为 $10nm$左右，精细的光谱分辨率反映了地物光谱的细微特征；数据量大，随着波段数的增加，数据量呈指数增加；信息冗余，由于相邻波段的相关性高，信息冗余度增加；可提供空间域信息和光谱域信息，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。

高光谱数据可以表示为高光谱数据立方体，可以看作是三维的数据图像。即在普通的二维图像以外多了一个维度的光谱信息。空间图像描述高光谱数据的地表二维空间特征，而光谱维这描述图像的光谱曲线特征，由此可以将传统的空间成像与现金的光谱测量技术有机结合。大多数地物都具有典型的光谱波形特征，尤其是光谱吸收特征。因此，从光谱数据库中将光谱匹配，可以实现地物识别的目标。

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。许多物质的特征往往表现在一些狭窄的光谱范围内，高光谱遥感实现了获取地物的光谱特征，同时又不会丢失其整体形态以及和周围地物之间关系的信息。

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，随着波段数的增加，数据量呈指数增长，容易出现维数灾难和修斯现象，使得高光谱图像的分类、识别等比较困难。由于相邻波段相关性高，导致信息冗余度高，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。在高光谱图像的降维处理中，特征提取和波段选择是两类常用的降维方法。

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是一种比较复杂的波段组合优化的问题。波段选择需要从原来的波段集中，选取出信息量大、波段间相关性小、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$D$。

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

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$S$:

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

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：

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

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。

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

$\tilde{S}^{(k)}$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$con^{(k)}(w_i, w_i) > con^{(k)}(w_i, w_j), j = 1, \ldots, N(j \neq i)$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$w_{c_i}(c_i \in \{1,2,\ldots,N\}$同时$i = 1,2,\ldots,m)$，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：

$$w^* = \mathrm{argmax}_{w_{c_i}} \left(rcon^{(k)}\right)\left(w_{c_i}, w_j\right)$$

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)

印度松数据集的场景是由 AVIRIS 传感器在印第安纳州西北部的印第安松树测试点上面收集的。原数据集由 145×145 像素和 224 个光谱反射带组成，波长范围为 $0.4 - 2.5 \times 10^{-6}$，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。有两条主要的双车道高速公路，一条铁路线以及一些低密度的房屋，其他建筑物和较小的道路。本文所使用的数据集是去除一些噪声带的印度松数据集。

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，聚类结果如**图 6**

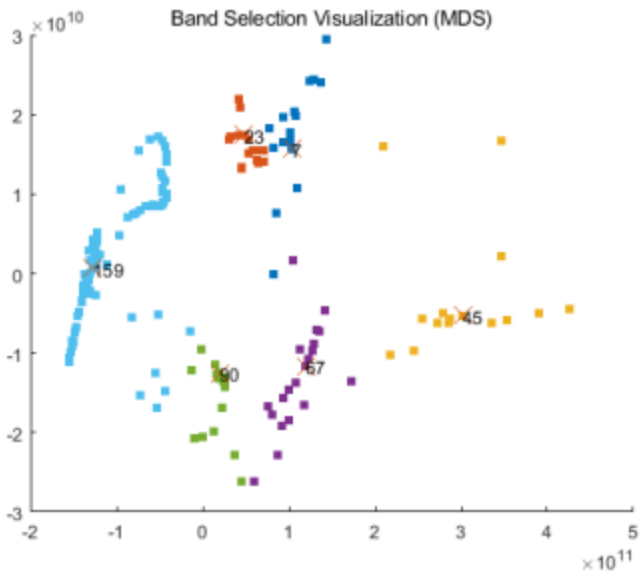


图 6.印度松数据集的波段选择

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，边缘检测的目的是标识数字图像中亮度变化明显的点。图像属性中的显著变化通常反映了事物属性的重要事件和变化。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。

图像的边缘检测可以减少原有复杂繁重的数据量，并且剔除了一些边缘检测认为不相关的信息，最大限度地保证了图像的结构和某些相关属性。

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

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$D$。

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

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)$P \in R^{N\times N\times 3}$。如**图 7**。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$R$(红色)通道，$k = 2$为$G$(绿色)通道,$k = 3$为$B$(蓝色)通道。

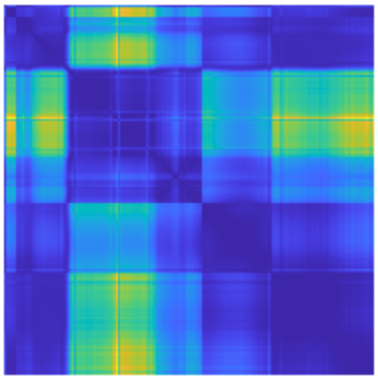


图 7

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，如**图 8**。

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

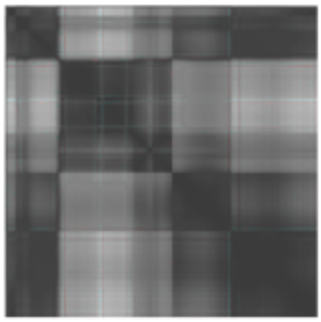


图 8

在图像的生成与收集中，难以避免引入噪声。图像噪声是指在图像数据中多余或者干扰原数据的干扰信息。图像噪声会干扰我们对图像信息的提取。因此，需要通过一定的方法去除图像噪声。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，从左至右地进行遍历，计算滤波器与对应像素点的值，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。

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

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

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，如**图 9**。

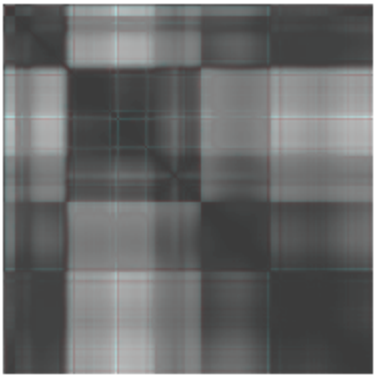


图 9

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。要达到边缘寻找的目的，灰度变化的检测可以用一阶导数或者二阶导数来实现。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：

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

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

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：

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

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。用于计算梯度偏导数的滤波器模板，通常称为梯度算子、边缘算子和边缘检测子等。

边缘检测实际上也是一种滤波算法，与上面处理灰度图像的中值滤波算法类似，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T，如果像素点通过 Prewitt 算子计算出来的值大于阈值，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。

Prewitt 算子模板如下：

$$d_y = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \quad d_x = \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

对待处理像素点$I_{r,c}$使用 Prewitt 算子进行计算：

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

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

得到$PR_{r,c} = G[f(x,y)]$,通过与阈值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，如**图 10** 所示。

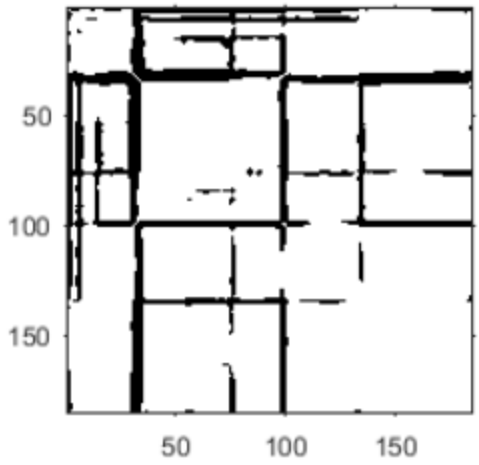


图 10

3.3.3 分别对不同组的高光谱数据使用基于中心连通性的聚类算法

用上述边缘检测的方法首先对高光谱数据集进行分组，如**图 11**。分别对边缘探测分组所得到的波段数据使用基于中心连通性的聚类方法，得到波段集合。

图 11.使用边缘探测对数据进行分组

通过 svm 分类器，将经过边缘探测处理的高光谱与未经过边缘探测处理的高光谱进行比较，得到结果如**图 12 所示。**

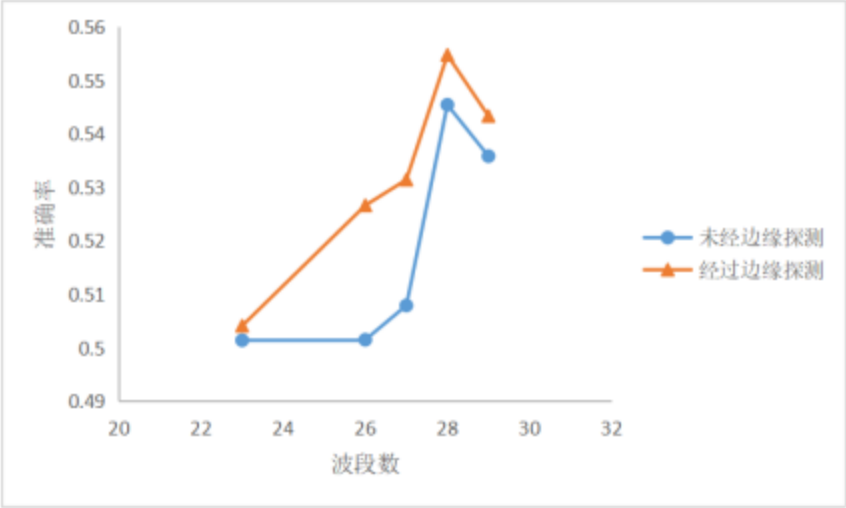


图 12

通过对比可以发现，经过边缘探测之后再使用基于中心连通性的聚类算法对高光谱进行波段选择的分类准确率要优于未经边缘探测。

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