

摘 要

多视角数据相对于单视角数据而言，每个视角之间可以相互提供互补信息，在聚类工作当中，这些互补信息可以提高聚类的性能。同时，多视角数据的出现使得数据样本规模变得十分庞大，容易出现计算困难的问题。另外，数据维度过高加深了维度灾难问题。谱聚类由于对任意形状样本表现出的良好性能，而且只需要简单的线性代数知识就可以求解，因此受到了广泛的关注。如何将谱方法运用到多视角聚类当中，成为了当前多视角聚类的热点问题。越来越多的学者对谱聚类和多视角聚类进行研究和总结。本文发现基于低秩稀疏分解思想的多视角聚类由于解决了噪声污染的问题，往往可以取得不错的效果。但是大多数多视角谱聚类构造相似矩阵时只考虑距离的因素，而不会去考虑数据是否处于同一密度区域，这不符合现实情况。针对此问题，本文设计出一种基于图卷积网络的鲁棒多视角谱聚类（GCN-RMSC）。

GCN-RMSC的思想比较简单。首先本文通过图卷积神经网络进行特征重构，得到数据集新的潜在特征表示，该潜在特征表示已经通过近邻信息的更新更有利于后续的聚类工作，它增强了位于同一密度区域的数据的近邻关系，而弱化了位于不同密度区域内的非近邻数据之间的关系。然后本文通过共享近邻思想利用潜在特征表示为每个视角构造出相似矩阵，利用该相似矩阵得到转移概率矩阵。利用共享近邻思想可以充分利用潜在特征表示中的近邻信息，为每个视角得到更符合实际情况的相似矩阵，利用该相似矩阵得到转移概率矩阵。最后，考虑到噪声的因素，可以采用低秩稀疏分解的思想，从噪声矩阵的影响中恢复出各个视角共享的转移概率矩阵，该共享转移概率矩阵已经十分符合实际情况，因此可以将此共享转移概率矩阵用于最后的谱聚类当中。

本文采用了五个基准算法，选取了六个数据集和四个评价指标进行验证，实验结果表明，本文所提出的GCN-RMSC算法优于原有的RMSC算法。并且本文的算法对于参数是不敏感的，而且在迭代很少的次数之内目标函数值就会达到满意的结果。

**关 键 词**：多视角谱聚类；低秩稀疏分解；GCN；共享近邻

ABSTRACT

Compared with the data from single view, the data from multi-view can provide complementary information to each other. In the clustering work, the complementary information can improve the performance of clustering. At the same time, the emergence of multi-view data makes the sample size of the data become very large, which is easy to appear the problem of computing difficulties. In addition, the high dimension of data deepens the dimension disaster problem. Spectral clustering has attracted extensive attention because of its good performance for arbitrary shape samples and a simple knowledge of linear algebra. How to apply spectral method to multi-view clustering has become a hot topic. More and more scholars have studied and summarized spectral clustering and multi-view clustering. It is found in this thesis that multi-view clustering based on low-rank and sparse decomposition can often achieve good results because it solves the problem of noise pollution. However, most of the multi-view spectral clustering constructs the similarity matrix only considers the factor of distance, without considering whether the data is in the same density region, which is not in line with the reality. To solve this problem, a robust multi-view spectral clustering based on graph convolutional network(GCN-RMSC) is designed in this thesis.

GCN-RMSC is sample. First in this thesis, graph convolutional network for feature reconstruction, from data set new potential feature, said the potential characteristics of said already through neighbor information update is more advantageous to the subsequent clustering work, it enhances the neighborhood relationship of data in the same area density, and weaken the are located in different density area of the relationship between the neighbor data. Then the similarity matrix is constructed for each perspective by using the idea of shared nearest neighbor and the transition probability matrix is obtained by using the similarity matrix. By using the shared nearest neighbor idea, we can make full use of the nearest neighbor information in the potential feature representation, get the similarity matrix which is more in line with the actual situation for each perspective, and use the similarity matrix to get the transition probability matrix. Finally, considering the factor of noise, the idea of low-rank sparse decomposition can be adopted to recover the transition probability matrix shared by all perspectives from the influence of noise matrix. The shared transition probability matrix is quite consistent with the actual situation, so this shared transition probability matrix can be used in the final spectral clustering.

In this thesis, five baseline algorithms, six datasets and four evaluation indexes are used. The experimental results show that the proposed GCN-RMSC algorithm is superior to the original RMSC algorithm.Moreover, the algorithm in this thesis is insensitive to parameters, and the objective function value can reach a satisfactory result within a few iterations.

**KEY WORDS**: Multi-view spectral clustering; Low-rank and sparse decomposition; GCN; Shared neighbor

目 录

[1 绪论 1](#_Toc74815468)

[1.1 研究背景与意义 1](#_Toc74815469)

[1.2 国内外研究现状 2](#_Toc74815470)

[1.3 论文主要工作内容 3](#_Toc74815471)

[1.4 论文组织结构 3](#_Toc74815472)

[2 相关背景知识介绍 5](#_Toc74815473)

[2.1 谱聚类 5](#_Toc74815474)

[2.1.1 相关概念 5](#_Toc74815475)

[2.1.2 拉普拉斯矩阵的基本特性 6](#_Toc74815476)

[2.1.3 标准化拉普拉斯矩阵 6](#_Toc74815477)

[2.1.4 马尔科夫随机游走谱聚类 7](#_Toc74815478)

[2.1.5 切图理论 8](#_Toc74815479)

[2.1.6 拉普拉斯谱聚类算法流程 11](#_Toc74815480)

[2.2 多视角学习 12](#_Toc74815481)

[2.2.1 协同训练 13](#_Toc74815482)

[2.2.2 多核学习 14](#_Toc74815483)

[2.2.3 子空间学习 15](#_Toc74815484)

[2.3 本章小结 16](#_Toc74815485)

[3 GCN-RMSC算法设计 17](#_Toc74815486)

[3.1 基于低秩稀疏分解的鲁棒多视角谱聚类 17](#_Toc74815487)

[3.2 基于GCN特征重构的鲁棒多视角谱聚类 18](#_Toc74815488)

[3.2.1 引入 18](#_Toc74815489)

[3.2.2 基于GCN的特征重构 19](#_Toc74815490)

[3.2.3 基于共享近邻的相似矩阵构造 22](#_Toc74815491)

[3.2.4 求解 23](#_Toc74815492)

[3.3 本章小结 25](#_Toc74815493)

[4 实验 27](#_Toc74815494)

[4.1 实验设置 27](#_Toc74815495)

[4.1.1 数据集 27](#_Toc74815496)

[4.1.2 聚类指标 28](#_Toc74815497)

[4.1.3 对比算法 28](#_Toc74815498)

[4.2 实验结果分析 29](#_Toc74815499)

[4.2.1 实验结果 29](#_Toc74815500)

[4.2.2 参数敏感性分析 32](#_Toc74815501)

[4.2.3 收敛性分析 33](#_Toc74815502)

[4.3 本章小结 34](#_Toc74815503)

[5 结论与展望 35](#_Toc74815504)

[5.1 结论 35](#_Toc74815505)

[5.2 展望 35](#_Toc74815506)

[参考文献 37](#_Toc74815507)

# 绪论

## 研究背景与意义

爬虫技术可以很好地用于数据挖掘中，近年来，随着网络技术水平的提高，人们通过爬虫获取数据的能力得到了极大的提高，同时，计算机的存储和运算能力的提高使得大数据的存储和运算成为可能[1]。数据可以从多个方面或多种源获得，从不同方面或不同源获得的数据称为多视角数据，其中不同的源得到的数据构成一个视角。现实生活中随处可见多视角数据，如图1-1所示，一篇文章可以由它上面的图片和周围的文本信息表示；一句话可以由不同的语言表示；一件事物可以从不同的角度描述；一场电影可以由视频帧和音频帧共同组成。单一视角的数据表达能力往往有限，并且在获取和存储数据过程中，可能会造成数据丢失或污染，因此聚类性能往往不高。单视角数据往往会存在视角数据缺失或者数据被噪声污染的问题，这些问题导致聚类或分类性能变差，而多视角数据的不同视角之间可以通过互补信息克服这些缺点，再加上可以利用不同视角独特的信息，运用多视角数据极大的提高聚类或分类算法的性能，因此对于多视角数据的研究引起了国内外学者广泛的关注。

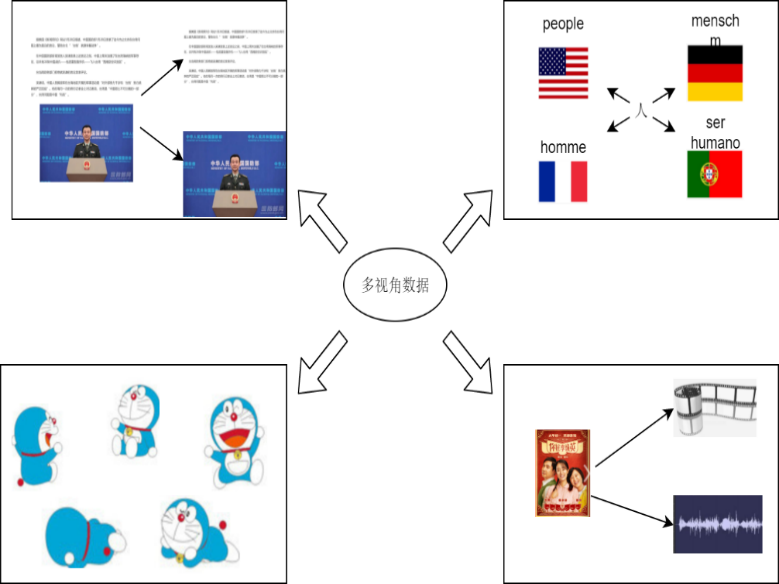


图1-1 多视角数据例子

多视角数据的每个视角来源于不同的源，因此它们之间既有联系又相互独立，对于多视角数据的学习需要遵循一定的规则以此来充分的挖掘数据信息，这促进了多视角学习机制的提出和广泛的研究。为了更好的利用多视角数据，一般需要遵循两个准则对数据进行处理：一致性准则和互补性准则[2, 3]。另外对于多视角数据的利用，主要有三种方法：第一种是协同训练，协同训练主要用于半监督学习，经过改造后可以用于无监督学习，它旨在设计一种规则使得各个视角之间互相提供互补信息；第二种是多核学习，多核学习旨在根据每个视角的重要性为每个视角设置不同的权重；第三种是子空间学习，子空间学习采用投影技术恢复出不同视角共享的子空间。

聚类是用来挖掘数据内部关系的，应用广泛。一旦通过聚类方法获得数据不同的簇，就可以执行许多后续的分析任务来实现不同的最终目标。传统的聚类学习，如k-means和GMM对于凸形数据的聚类效果不错，但是对于任意形状的数据聚类效果不佳，往往会陷入局部最优解。谱聚类运用拉普拉斯谱图原理和切图理论进行聚类，可以适用于任意形状的数据，而且可以得到全局最优解。另外谱聚类算法可以通过矩阵特征分解进行求解，并且有一定的理论支撑，因此也受到了国内外学者广泛的关注。

多视角聚类是一种机器学习范式，因为结合了多个视角的特征信息，不同视角之间信息可以相互补充，因此聚类性能常常优于单视角聚类。将谱聚类的方法运用到多视角聚类当中可以极大地提高聚类的性能，因此多视角谱聚类受到了国内外学者的广泛关注，在近二十年间得到了迅速的发展。

## 国内外研究现状

对于多视角数据，传统的聚类只是简单地将多个视角拼接在一起[2]，这种做法有很大的缺陷。首先这会造成数据冗余，降低聚类的性能。再者，不同视角之间是有特殊的统计意义的，简单的将视角拼接在一起，会丢失这些统计属性。最后，不同视角的数据在收集和存储的过程中可能会被噪声污染，将数据拼接在一起没有考虑到噪声的影响。因此将传统的聚类算法运用到多视角数据甚至可能会降低聚类性能。

由于谱聚类对非凸数据集的良好性能，谱聚类得到了广泛的关注，已经有学者对谱聚类进行了总结。邢洁清等人总结了谱聚类的研究进展[4]。李玲俐对谱聚类的应用情景进行总结[5]。Von L等人对谱聚类的理论和实验进行了详细的总结[6]。

多视角的不同视角之间可以提供互补性信息，因此可以得到比单视角聚类更好的性能，国内外越来越多的学者对其进行总结。孔阿栋在总结多视角子空间学习的基础上，提出了采用近邻对多视角保持嵌入的方法[2]。唐静静等人总结了多视角学习的主要方法[7]。Chao X等人对多视角的理论和不同的算法进行了详细的总结[3]。

多视角聚类的良好性能，使得多视角聚类成为当前研究的热点。为了提高多视角聚类的性能，国内外的学者展开了积极的研究。已经有论文将多视角聚类进行分类[8]，如图1-2所示。另外，Guo Q等人采用了一种新型的方法对多视角聚类进行总结[8]。贺艳芳等人在单视角谱聚类的基础上介绍了多视角谱聚类如何运用到大数据中[9]。黄宗超等人提出了一种基于子空间融合的多视角聚类算法[10]。Hao WY等人采用超图拉普拉斯正则化和自动加权策略来学习多视角谱聚类[11]。Kumar等人提出了一种用于多视角谱聚类的协同训练算法和协同正则化算法[12, 13]。Jie W等人提出了一种采用一致性规则解决不完全视角问题的多视角谱聚类算法[14]。宋艳等人提出了一种采用共享近邻思想改造每个视角相似矩阵的方法[15]。杨金鸿等人将度量学习运用到多视角谱聚类当中[16]。程士卿等人将低秩张量分解运用到多视角谱聚类[17]。胡世哲等人提出了一种双重加权的多视角聚类算法[18]。Chang Q等人提出了一种通过子空间学习构造潜在表示的多视角聚类算法[19]。Rong K等人提出了将低秩稀疏分解运用到多视角谱聚类，已解决噪声问题[20]。于心源提出了将RMSC运用到极化SAR图像分类[21]。

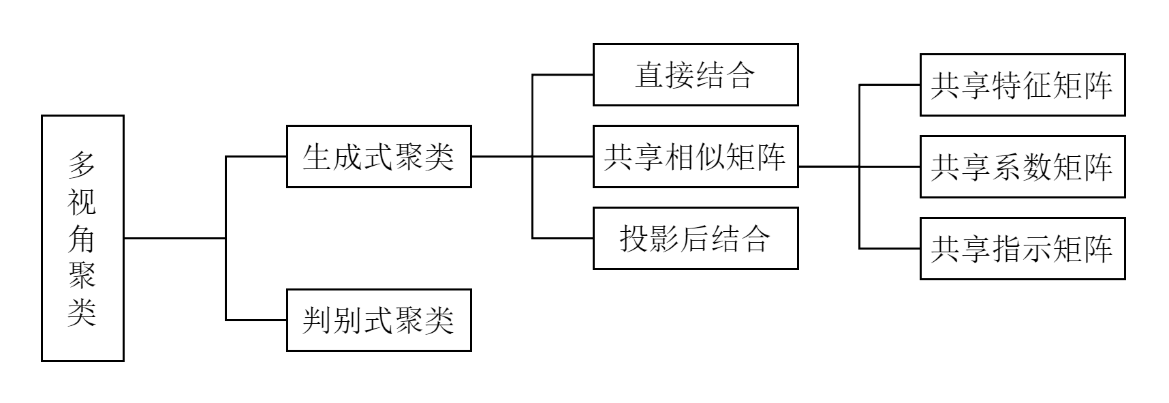


图1-2 多视角聚类分类

## 论文主要工作内容

由于多视角数据的出现和谱聚类算法的良好性能使得多视角谱聚类成为当前研究的热点，本文在总结多视角谱聚类的基础上，发现了基于低秩稀疏分解的谱聚类算法往往可以取得不错的效果，鉴于此，本文在深入学习了基于低秩稀疏分解的鲁棒多视角谱聚类（RMSC）的基础上，对此算法进行了改进，提出了一种新颖的多视角谱聚类算法（GCN-RMSC）。本文的主要工作如下：

（1）本文深入研究了当前多视角谱聚类的背景和相关研究内容，对谱聚类和多视角学习的基础进行了介绍。

（2）本文详细地介绍了基于低秩稀疏分解的鲁棒多视角谱聚类，它是一种基于马尔可夫随机游走的多视角谱聚类算法，其中心思想类似于低秩稀疏分解。

（3）本文提出了一种新颖的多视角谱聚类算法——GCN-RMSC，它采用GCN和共享近邻的思想对RMSC中相似矩阵的构造方法进行改进，以便得到更接近真实情况的相似矩阵。

（4）本文采用了五个baseline算法，并且在六个数据集上将它们与本文的GCN-RMSC进行比较，实验结果表明，本文所提出的GCN-RMSC优于原先的RMSC算法。

## 论文组织结构

本文具体展开方式如下：

第一章，介绍多视角谱聚类的研究现状和本文主要完成的工作。

第二章，介绍了谱聚类的相关基础概念。介绍了多视角学习的分类情况和每种类型的主要算法思想。

第三章，详细介绍了RMSC算法，它是一种基于低秩稀疏分解的算法。另外介绍了改进后的GCN-RMSC算法，它在RMSC的基础上采用了GCN和共享近邻思想进行改进。它的主要步骤为首先通过GCN重构数据集的潜在特征，然后通过共享近邻思想构造相似矩阵，最后将相似矩阵运用到RMSC当中。

第四章，详细介绍了本文的实验内容，本文采用了五个baseline算法，选取了六个数据集和四个聚类指标进行验证，实验结果表明本文所提出的算法GCN-RMSC优于原先的RMSC算法。另外，本文对GCN-RMSC进行了参数敏感性分析和收敛性分析。

第五章，总结了本文所完成的工作，介绍多视角谱聚类目前主要的缺陷以及未来需要做什么。

# 相关背景知识介绍

## 谱聚类

传统的聚类算法，例如GMM，k-means等对于凸型数据集的聚类效果比较好，但是对于任意形状的数据集的聚类效果不佳。另外，根据不同的初值设定，传统的聚类算法可能会陷入局部最优解。谱聚类可以很好地解决任意形状样本数据集的聚类问题，并且可以通过简单的线性代数知识进行求解而得到全局最优解，因此受到了国内外学者的广泛关注[9]。同时，谱聚类可以通过降维的方法解决k-means和核k-means等算法的最终聚类结果依赖于初值的问题，尤其当样本维度很高的时候特别有用。

谱聚类算法依赖于样本集构造的相似矩阵，不同的拉普拉斯矩阵对于最终的结果影响很大。谱聚类的本质思想是首先通过局部保持映射得到有利于聚类的信息，然后用k-means得到最终的聚类结果。

### 相关概念

首先给出本小节需要用到的相关概念和基本符号。给定数据集以及两个样本点和之间的相似度，由构造出相似矩阵。谱聚类划分直觉上的目标是将数据集中的不同样本划分到几个簇中，使得不同簇间的相似度小而同一簇内的相似度大。这与谱图划分十分相似，因此本文构造，其中是图中的样本点，表示顶点间的相似度。这样谱聚类对应于图的划分——将图进行切割之后，使得子图之间由很小的权重边连接，而同一子图内由很大的权重边连接。

将图中每一个顶点的度定义为：与相连的所有边的权重之和，即。

利用所有点的度可以得到度矩阵，它是一个的对角矩阵，其中。利用所有点的权重值得到图的邻接矩阵,它也是的矩阵，其中。对于两个不相交的子集,定义。最后定义为子集中点的个数，。

获得相似矩阵的思想是将距离比较近的两个点的权重设置的高一点，将距离比较远的两个点的权重设置的低一点，基本做法是用样本点之间的距离度量获取相似矩阵。有三种常见的方法来构造相似矩阵。

1. 阈值法：当相邻两个顶点的距离不大于时，将相似矩阵对应位置的值设置为相同的值，当相邻两个顶点的距离大于时，将相似矩阵对应位置的值设置为0。公式如下所示，其中:

（2-1）

1. 近邻法：取每个样本最近的个点作为它的近邻，只有这个样本点对应的，其他样本点对应的，但这会造成邻接矩阵非对称，为了解决非对称问题，采用下面两种形式：
2. 其中一个样本点是另一个样本点的近邻时就保留:

（2-2）

1. 必须两个样本点互为近邻时才能够保留:

（2-3）

1. 全连接法：相似矩阵的每一个值都大于0，我们可以选用不同的核函数来定义邻接矩阵的值，如高斯核、多项式核，其中最常用的是高斯核：

（2-4）

### 拉普拉斯矩阵的基本特性

拉普拉斯矩阵来源于拉普拉斯映射算法，它具有一些很好的性质。我们定义拉普拉斯矩阵，其中和是上文所说的度矩阵和相似矩阵。拉普拉斯矩阵具有以下性质，它可以很好的用于谱聚类。

**性质1**拉普拉斯矩阵满足以下性质：

1. 对于任意向量,满足条件：。
2. 是个对称矩阵，它的所有元素都大于等于零。
3. 的最小特征值是0，它所对应的特征向量是全1的特征向量。
4. 的所有特征值都大于等于0，即

**证明：**

1. 通过的定义

（2-5）

1. 因为和都是对称矩阵，所以也是对称矩阵，再由1）的结论得的所有元素都大于等于零。
2. 由的定义显然成立。
3. 由上面三个结论联合起来可以证明。

### 标准化拉普拉斯矩阵

标准拉普拉斯矩阵是非标准化拉普拉斯矩阵的正则化形式，主要可以采用两种方法，它们定义如下：

（2-6）

这样标注两个拉普拉斯矩阵是因为第一个拉普拉斯矩阵是一个对称矩阵，而第二个拉普拉斯矩阵可以通过随机游走得到。下面给出这两个标准拉普拉斯矩阵的一些性质。

**性质2**标准拉普拉斯矩阵和满足以下性质：

1. 对于每一个满足
2. 是对应于特征向量的特征值，同时也是对应于特征向量的特征值。
3. 通过求解可以得到的特征值和特征向量。
4. 对于特征值，对应的特征向量是，对应的特征向量是。
5. 和都是非负的半正定矩阵，且其特征值满足。

### 马尔科夫随机游走谱聚类

给定一组数据点，本文定义了一个相似矩阵。设是一个有顶点集和边集的带权图，其中每个顶点与数据点相关联，每条边与相关联。可以采用定义相似矩阵（本文后面会介绍改进的方法）。

马尔科夫随机游走和谱聚类之间有着天然的联系，随机游走是一个随机过程，经过有限步之后最终会达到稳定过程。谱聚类可以被解释为试图在上找到一个划分，使得数据点的随机游走会有很大的概率一直待在同一个簇内，只有在噪声扰动的情况下才会跳跃到不同簇中。

定义从一个顶点跳到另一个顶点的概率为，然后得到转移矩阵（转移矩阵满足行和为1）。对于一个非二分连通图，随机游走总会产生一个唯一的稳定分布，其中。可以很自然的发现，和存在紧密的关系，即。因此可以很容易的想到随机游走和谱聚类之间存在某种关系，下面给出证明。

**定理1**（通过转移矩阵得到Ncut）:令G为非二分连通图，假设在G上从开始随机游走并且最终到达稳定状态，对于节点集的两个不相交子集，令，然后得到。最后证明可以通过转移概率得到Ncut。

（2-7）

根据公式（2-7），我们得到：

（2-8）

因此，随机游走的收敛等同于谱聚类的收敛，最小化NCut实际上对应的是使得随机游走几乎不会从一个簇跑到另一个簇，反之亦然。

给定一个与转移矩阵连通的非二分图，存在满足，这是马尔科夫链的稳态。将马尔科夫运用到谱聚类的结论如下[20]：

**定义1**（马尔科夫谱聚类）：马尔科夫谱聚类的定义如下：

1. 给定带权图，定义一个转移概率矩阵的上的随机游走，使得它有一个平稳分布满足。
2. 令表示对角矩阵，其第个对角元素为平稳分布，构造矩阵。
3. 通过广义特征分解获得r个最小广义特征向量。

通过k-means对新的特征进行聚类。

### 切图理论

对于任意两个子图集合并且，定义之间的权重为以及的补集为。那么对于个子图的集合：我们定义切图。

谱聚类划分直觉上的目标是将数据集中的不同样本划分到几个簇中，使得不同簇间的相似度小而同一簇内的相似度大。切图的目标是使图内的点距离最小而图之间的点距离最大。这样，可以很容易的将聚类问题转化为图的切割问题。如何切割图使得图内的点距离最小而图之间的点距离最大的一种直观上的想法是：最小化，但此时会存在问题，如图2-1所示：

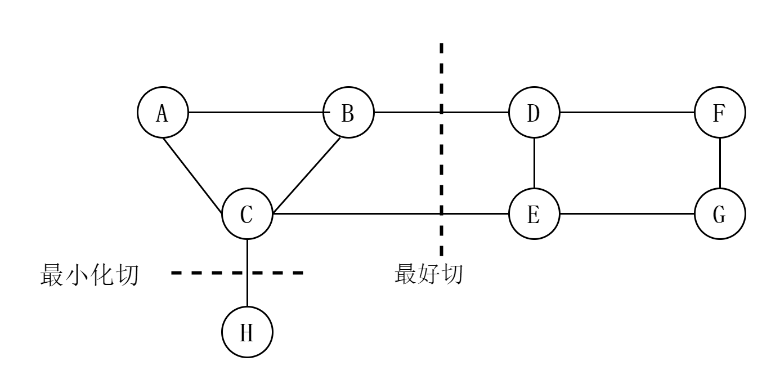


图2-1 最小切图的不合理性

上图得到了一个独立点，这种聚类效果是不佳的。为了避免出现图一所示的切图不佳的情况，本文采用了两种方法避免产生孤立点这种情况即比例割集准则和规范割集准则。

#### 比例割集准则

孤立点的存在使得划分效果也就是聚类效果不佳，为了克服此问题，Hagen和Kahng提出了比例割集准则（RatioCut）[4-6]，它不仅仅考虑到了最小化cut(A\_1,A\_2,…,A\_k )，同时考虑到了最大化每一个划分中数据点的个数。

（2-9）

这个优化问题的复杂度是指数级的，无法求解，为了更方便的求解，首先我们考虑的情况，它的最优化问题可以转换成。为了更方便的求解这个问题，我们定义一个指示向量，其中的定义为：

（2-10）

有了这个指示向量，我们可以推导出更好的便于求解的最优化问题。现在利用非标准化拉普拉斯矩阵的性质重写此最优化问题，我们得到下面的公式：

（2-11）

另外有以及，结合公式（2-11）的最优化问题可以得到新的优化问题：

（2-12）

其中为全为1的向量。公式（2-12）是离散最优化问题，仍然无法求解，通过放宽解的条件允许，可以得到一个松弛的优化问题：

（2-13）

公式（2-13）已经存在了相应的解法，即Rayleigh-Ritz定理，根据此定理可以得到公式（2-13）的最优解为的第二小特征值对应的特征向量[6]。

现在我们考虑任意的，引入指示向量，对于任意的，定义：

（2-14）

那么对于有：

（2-15）

令表示指示向量，可以得到，因此比例切的优化目标转换为：

（2-16）

对应的最优解仍然是找到的最小k个特征值对应的特征向量，这是一个典型的降维方法，此时的，对维度进行了规约，可以解决维度灾难问题。但同时降维后会损失少量信息使得优化后指示向量不能完整地表示各个样本的归属，因此一般需要在上再次进行传统的聚类，如k-means聚类。

#### 规范割集准则

原有的割集准则只能够度量同一簇内样本的相似度，而忽略了簇间样本的关系，为了克服此问题，Shi和Malik提出了规范割集准则（NCut）[6]。不同于RatioCut，Ncut在最小化cut(A\_1,A\_2,…,A\_k )的同时也考虑了最大化每一个划分中子图的度的和。

（2-17）

直接考虑任意的的情况。引入指示向量，对于任意的，定义：

（2-18）

那么对于有：

（2-19）

此时，（）。优化目标函数转换为：

（2-20）

令，可以得到，因此最优化问题可以转换为：

（2-21）

此时的最优解对应于的k个最小特征值对应的特征向量。规范割集准则对应于标准化拉普拉斯聚类。

### 拉普拉斯谱聚类算法流程

根据不同的划分准则，本文得到了三种不同类型的谱聚类，下面给出三种谱聚类算法。

首先本文给出非标准化拉普拉斯谱聚类的算法流程如表2-1所示:

表2-1 非标准化谱聚类算法流程

|  |
| --- |
| 算法输入：样本集，聚类簇数k，三种相似矩阵生成方式的一种  算法输出：簇划分 |
| 算法步骤：   * 构造相似矩阵； * 根据公式计算邻接矩阵和度矩阵； * 计算； * 计算前k小特征值对应的特征向量； * ，令是的第行； * 使用k-means对新特征聚类得到最终结果 |

有两种形式的标准化拉普拉斯谱聚类算法，本文先给出随机游走标准化拉普拉斯谱聚类算法的流程，如表2-2所示:

表2-2 随机游走标准化谱聚类算法流程

|  |
| --- |
| 算法输入：样本集，聚类簇数k，三种相似矩阵生成方式的一种  算法输出：簇划分 |
| 算法步骤：   * 非标准化谱聚类前三个步骤 * 通过解特征方程得到前k小特征值对应的特征向量（使用的性质）； * ，令是的第行； * 使用k-means对新特征聚类得到最终结果 |

注意到表2-2的算法中用到了性质2中的性质，其实我们也可以通过直接求解的特征分解得到算法流程。下面给出对称拉普拉斯谱聚类的算法流程，如表2-3所示:

表2-3 对称标准化谱聚类算法流程

|  |
| --- |
| 算法输入：样本集，聚类簇数k，三种相似矩阵生成方式的一种  算法输出：簇划分 |
| 算法步骤：   * 非标准化谱聚类前三个步骤 * 计算标准化拉普拉斯谱聚类 * 计算前k小个特征值所对应的特征向量 * 对特征向量进行行标准化后得到，其中 * ，令是的第行； * 使用k-means对新特征聚类得到最终结果 |

可以看出三个算法流程的大致步骤是相同的，只是在计算特征分解时使用到的拉普拉斯矩阵不同。另外，在对称拉普拉斯谱聚类当中，本文增加了对特征向量进行行标准化的步骤，这是因为，根据扰动理论[6]，非标准化谱聚类和具有的标准化谱聚类算法对于噪声点具有很好的鲁棒性，但是如果图形包含度数非常低的顶点时，可能会发生错误，因此需要进行行归一化来降低度数低顶点的影响。

## 多视角学习

现实中的数据往往是多视角的。不同视角的冗余信息是单视角数据和多视角数据之间的主要差异，每个视角都可以提供其他视角没有的信息正是因为这些差异的存在，才使得多视角数据可以提供更多的信息。为了利用这些信息需要遵循两个原则来对数据进行处理：一致性原则和互补性原则。一致性原则是为了利用数据之间的一致性关系提高算法性能。对于多视角数据，每个视角都会包含其他视角所不存在的信息，充分利用这些信息，可以更准确的预测数据，互补性原则就是为了充分利用这些信息。当前对于多视角数据的利用主要有三种方法：第一种是协同训练，协同训练经常用于半监督学习，但经过改进后可以用于无监督学习中，它旨在设计一种规则使得各个视角互相之间提供信息；第二种是多核学习，多核学习旨在根据每个视角的重要性为每个；第三种是子空间学习，子空间学习假设存在一个共享的子空间，不同视角从此空间转换而来，子空间学习的目的是恢复出这个共享的子空间。

### 协同训练

协同训练算法是最早的用于多视角学习的算法，通过交替优化以最大化不同视角之间的一致性，提升算法性能。协调训练算法需要满足三个基本的假设：充分性、兼容性和条件独立性[3]。充分性表明每个视角的数据只有少部分被噪声污染了，它仍然可以学习到一个好的学习器；兼容性表明不同视角的学习器对于同一个数据可以做出相同的结果的概率是很大的。

协同训练算法满足一致性原则和互补性原则。协调训练算法对于多视角数据的不同视角训练相互联系的学习器，使得这些学习器对于同一输入可以预测出相似的输出，并且协同训练算法在每次迭代之后都会极小化两个视角之间的不一致性，从而提高学习器的性能。具体的流程如图2-2所示：

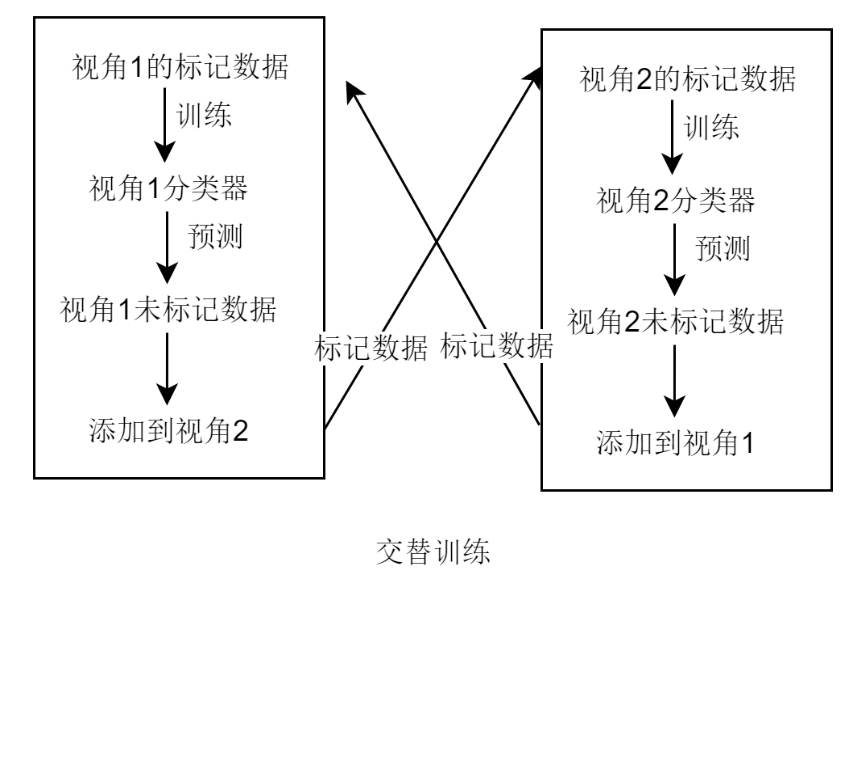


图2-2 协同训练算法流程示意图

本文以标准协同训练算法为例介绍协同训练算法。假设数据集有两个视角，可以将这两个视角分为有标签的数据集和无标签的数据集。具体算法流程如表2-4所示：

表2-4 标准协同训练算法

|  |
| --- |
| 算法输入：  算法输出：分类器 |
| 算法步骤：    ; |

表2-4 标准协同训练算法（续）

|  |
| --- |
| 算法输入：  算法输出：分类器 |
| 算法步骤：    ;    用训练视图上的学习器  用对未标记数据进行分类    把对的分类结果中前k个最置信的数据及其预测标签加入到  把这2k个数据从中删除 |

在实现过程中，为了减少每轮中计算置信度的次数，可以设置一个缓冲池。算法最终会输出两个分类器，然后用这两个分类器做集成可以得到最终的分类结果。算法可以使用相同的带置信度分类器，也可以采用不同的带置信度分类器。

### 多核学习

在多核映射背景之下，可以将高维空间的特征表示为多个低维特征空间的组合。不同的组合空间可以发挥出各个基本核的不同特特征映射能力，因此将多视角数据表示为低维特征空间，然后组合为高维特征空间可以很好的利用多视角数据。

通过不同的多核学习方法将不同视角的特征空间组合为不同的组合核空间可以利用多视角的不同方面的信息。但不同的多核学习方法都遵循统一的学习流程，多核学习的学习流程如图2-3所示：

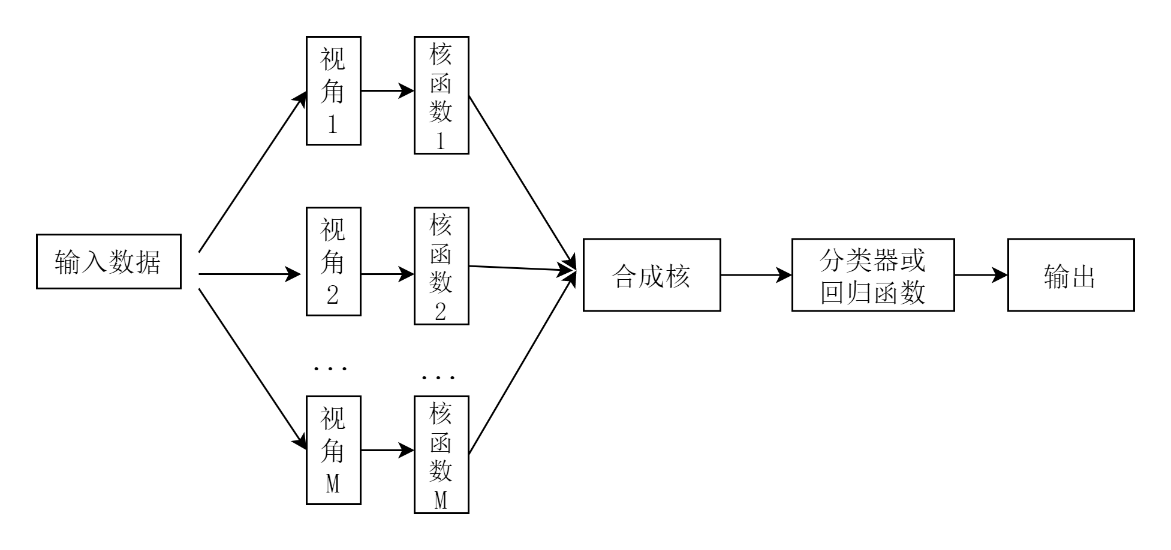


图2-3 多核学习流程

多核学习通过将不同核结合按照不同权重结合在一起来提高算法性能。其中最常用的方法是合成核方法。可以将几个基本核函数线性组合在一起，这是最基本的方法，但是效果不是很好。并且这种方法会导致某个核函数都分配相同权重的问题，针对这个问题，可以采用选通模型局部选择合适的核函数。这些和组成方法通过在训练数据上进行学习来获取一些权重用来标识每个核的重要性。另外，理论上基本核的个数可以是无限多个的，因此可以把多核学习从有限核扩展到无线核方法。

### 子空间学习

子空间学习采用空间投影技术恢复出不同视角共享的子空间，其具体算法流程如图2-4所示：

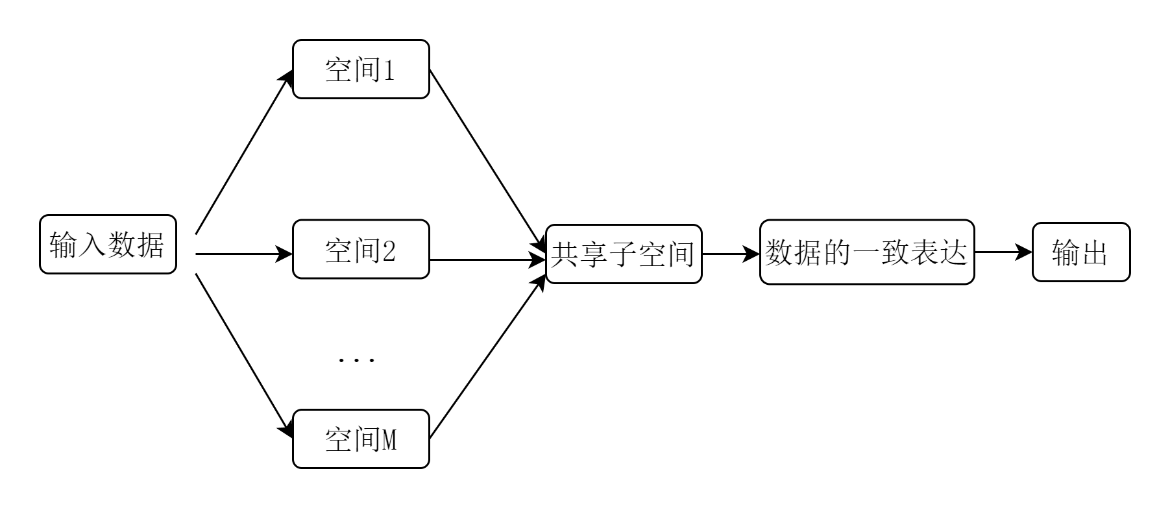


图2-4 子空间学习流程

子空间学习是一种降维方法，能有效解决维度灾难问题，并且映射之后的数据特征更有利于后期的学习，最常用的子空间学习方法是典型相关性分析（CCA）及其变形核典型相关性分析（KCCA）。

#### 典型相关性分析

假设多视角数据，，为了更好的得到和的投影向量，假设和投影后的基向量为和，使用皮尔逊相关系数度量和之间的关系，以解得最优的和使得和之间的关联程度最大，此时的目标函数为：

（2-22）

其中和都进行了数据居中处理，考虑到和的尺度问题可能导致多解的情况存在，可以采用固定分母，求解分子的方法进行求解，求解后的优化问题转换为：

（2-23）

这是有约束最优化问题，很容易求解。

#### 核典型相关性分析

当特征之间的关系为非线性关系时，CCA的效果不好，可以采用核方法解决这个问题，其基本思想是低维中不可分的数据在高维中可以变得线性可分，这就是KCCA。用，表示核矩阵，可以得到优化问题为：

（2-24）

可以通过类似CCA的方法进行求解。

## 本章小结

本章介绍了三种谱聚类算法的由来以及算法流程。三种谱聚类都采用了拉普拉斯矩阵的特征分解，不同的是对称拉普拉斯矩阵需要对特征向量进行标准化。谱聚类算法由于其对任意形状数据的良好性能受到广泛的关注。谱聚类算法只需要数据之间的相似矩阵，因此对于稀疏矩阵的聚类也很有效。同时谱聚类进行了降维，可以解决维度灾难问题。但谱聚类需要计算特征分解，对于海量数据运算速度很慢，同时它也依赖于相似矩阵，对于不同的相似矩阵，谱聚类的结果可能千差万别。另外，本章介绍了马尔科夫随机游走和谱聚类之间的关系。

同时本章介绍了多视角学习的三个基本方法，它们基于两个基本的原则——一致性原则和互补性原则。本章主要介绍了三个多视角学习方法的基本思想以及它们的代表性算法。协同训练算法在不同视角上训练学习器，然后迫使它们在视角之间保持一致性。多核学习在每个视角上计算单独的核并以基于核的方法组合起来。子空间学习通过降维的方法得到一个共享的子空间，在共享子空间上进行学习提高算法性能。

# GCN-RMSC算法设计

多视角聚类旨在从多个能够提供互补信息的多视角数据中实现对数据的划分，在近年来受到了国内外学者广泛的关注。在现实生活的聚类问题中，不同视角的数据可能会存在一定的噪声，但是，现有的聚类方法盲目地将来自多个视角的信息与一些可能存在的噪声结合在一起，这样做通常导致聚类算法性能的下降。本小节将介绍一种基于子空间学习的多视角谱聚类算法（RMSC），它将马尔科夫链和鲁棒多视角谱聚类结合在一起，采用了低秩稀疏分解的思想。它的主要思想是首先为每个视角构造转移概率矩阵，然后重构出这些视角的共享转移概率矩阵。

另外，谱聚类主要依赖于相似矩阵的构造，但是传统的相似矩阵构造方法只考虑距离的因素，没有考虑处于不同密度区域的数据点之间的相似度。可以直观的发现，数据点之间的相似性不仅与距离有关，而且跟数据是否位于同一密度区域也有关系。本文在鲁棒多视角谱聚类（RMSC）的基础上，考虑了图卷积神经网络（GCN）在近邻信息传递中的作用，提出了基于GCN特征重构的鲁棒多视角谱聚类（GCN-RMSC）。

## 基于低秩稀疏分解的鲁棒多视角谱聚类

RMSC基于低秩稀疏分解的思想，考虑了数据存在噪声的因素，得到了比较鲁棒的聚类结果。为了实现RMSC，首先需要考虑的是为每个视角构造相似矩阵和转移概率矩阵。在多视角聚类的背景下，给定m个视角下的一组数据点，在每个视角中，可以构造一个相似矩阵和相应的图。令是与相对应的转移概率矩阵。如图3-1所示，每个视角的转移概率矩阵含有足够的聚类信息，但是此转移概率矩阵被噪声污染，这些噪声是很少一部分的，因此可以将转移概率矩阵分解成两个部分: 共享转移概率矩阵和噪声矩阵即 。

一旦给定，就可以简单地用作为马尔可夫链谱聚类方法的输入转移矩阵，从而得到最终的聚类结果。

接下来的问题是如何对共享转移概率矩阵和误差矩阵进行建模。相似矩阵和转移概率矩阵是用来反映不同数据点之间的情况的，一般而言，相似矩阵和转移概率矩阵中用于描述位于同一簇的几个数据点的几行之间是有强相关性的，这就导致相似矩阵和转移概率矩阵的某一行或某几行可以由其他几行表示，也就是说相似矩阵和转移概率矩阵是低秩的，如图3-1所示，转移概率矩阵是一个块状矩阵。在构建不同视角的相似矩阵和转移概率矩阵时，数据集可能含有噪声，这种噪声破坏了相似矩阵和转移概率矩阵的低秩性，但是噪声在一般情况下是随机的，它对数据的破坏是很少一部分的，也就是说噪声矩阵是趋向于稀疏的，如图3-1所示，噪声矩阵只有很少一部分非零。

RMSC虽然已经考虑到了噪声对聚类结果的影响，但是它构造相似矩阵的过程中只考虑到了距离的因素，没有考虑到数据是否位于同一密度区域内这一因素，因此RMSC的聚类结果不是十分理想，为此本文设计了一种新颖的多视角谱聚类算法——基于GCN特征重构的鲁棒多视角谱聚类（GCN-RMSC）。

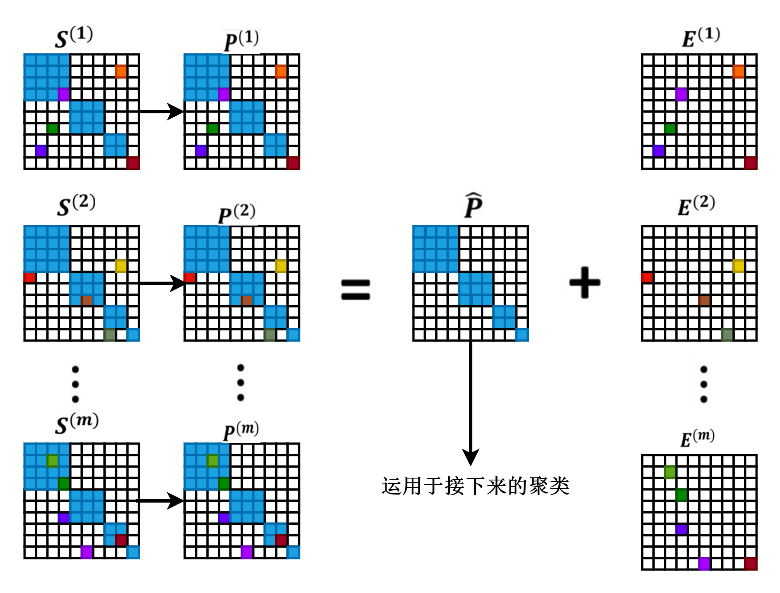


图3-1 真实转移概率矩阵构造图

## 基于GCN特征重构的鲁棒多视角谱聚类

### 引入

谱聚类的性能主要依赖相似矩阵的构造，传统的相似矩阵构造方法如全连接法只考虑了距离的因素，而没有考虑处于同一密度区域的因素。如图3-2所示，A和B与A和C的距离相等，按照传统的相似矩阵构造方法，A和B与A和C的相似度是一样的。但是在现实情况中，A和B与C位于不同的密度区域内，A和B的相似度应该大于A和C的相似度，因此仅仅通过距离的因素得到的相似矩阵不太合理。此时，考虑近邻的因素，如图3-2所示A和B有4个共享的近邻（红球），A和C有2个共享的近邻（蓝球），通过近邻可以对依靠距离得到的相似度进行更新。本文通过两个步骤采用近邻思想对相似矩阵进行改进，首先通过近邻传递的思想对原始数据改进得到每个数据点新的表示H，然后对H采用共享近邻的思想进行改进[15]。

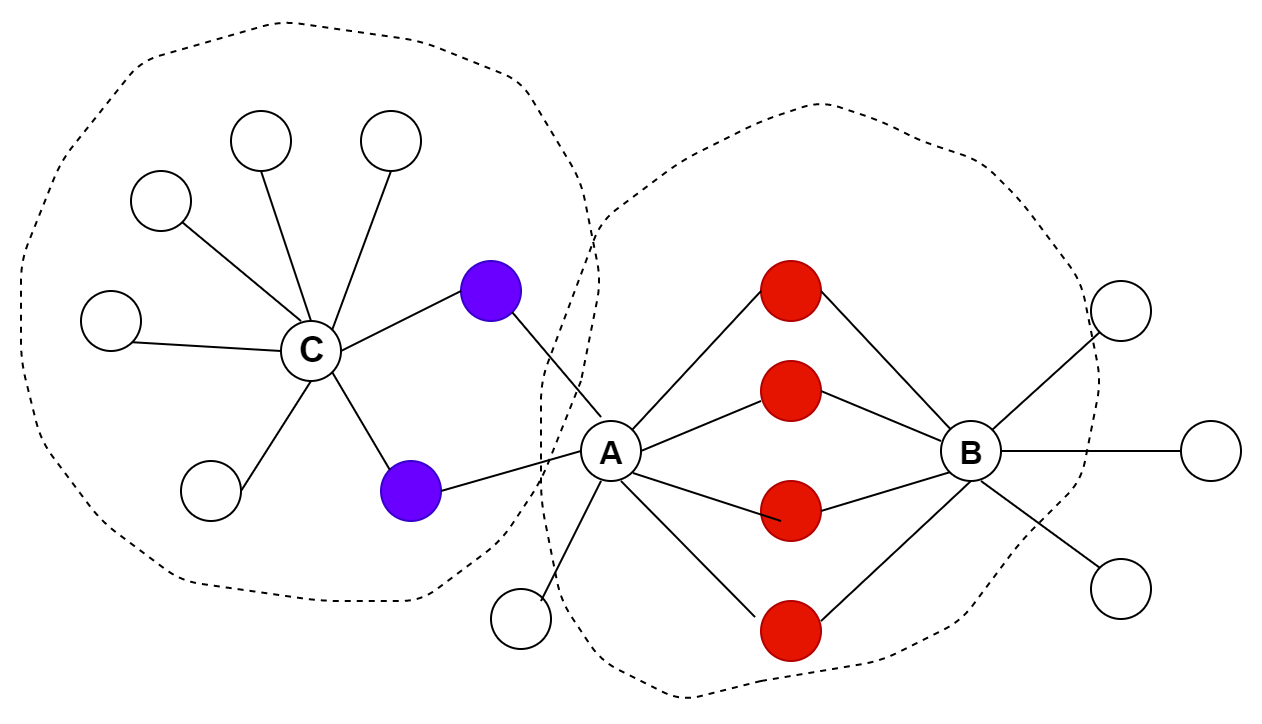


图3-2 近邻思想示意图

通过邻居结点的信息来更新结点自身的信息可以自然而然的令人想到图卷积神经网络（GCN）。GCN的核心思想是通过聚集结点本身的特征以及其邻居结点的特征来生成结点新的特征表示[22, 23]，因此本文采用了GCN来构造更符合实际情况的相似矩阵，以此来能够获得更好的聚类结果。本文所提出算法GCN-RMSC的完整步骤如图3-3所示，首先通过图卷积神经网络进行特征重构，得到数据集新的潜在特征表示，已经通过近邻信息的更新更有利于后续的聚类工作，它增强了位于同一密度区域的数据的近邻关系，而弱化了位于不同密度区域内的非近邻数据之间的关系。然后本文通过共享近邻思想利用为每个视角构造出相似矩阵，利用得到转移概率矩阵，利用共享近邻思想可以充分利用潜在特征表示中的近邻信息，为每个视角得到更符合实际情况的相似矩阵，利用得到的转移概率矩阵更有利于后续的聚类工作。最后本文将转移概率矩阵运用到RMSC中得到最终的聚类结果。

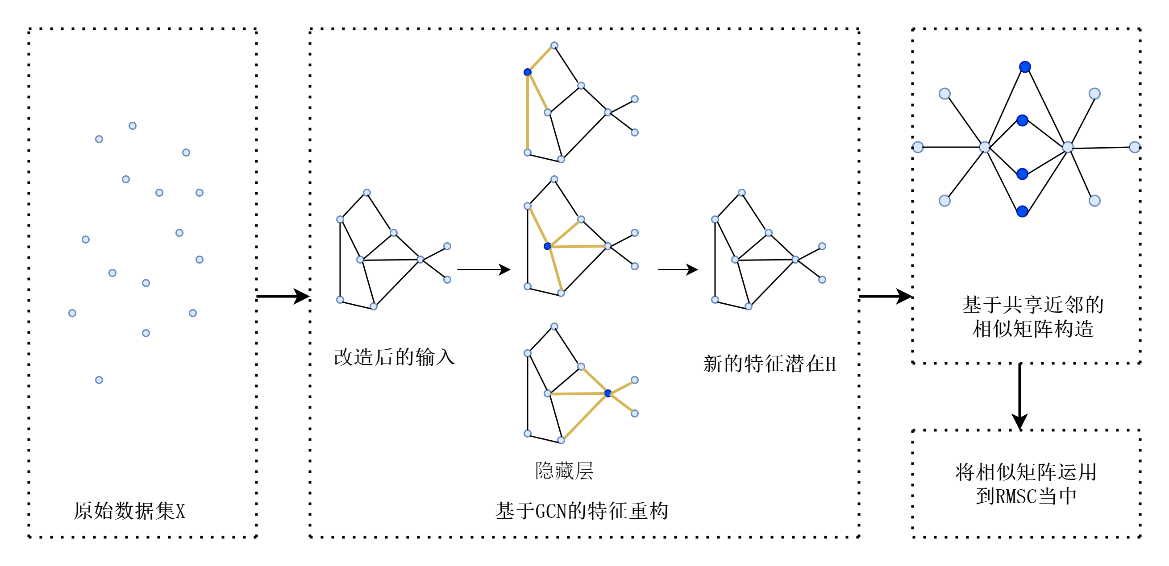


图3-3 GCN-RMSC完整步骤

### 基于GCN的特征重构

在传统的谱聚类相似矩阵的构造过程中，已经有了采用近邻思想构造相似矩阵，但是该方法只考虑了两个结点之间是否为近邻，而没有考虑共享的近邻以及近邻之间的传递，因此效率不高。考虑到图卷积网络（GCN）中每个结点可以根据其相邻结点的潜在表示更新自己的潜在表示，也就是说，在更新完成后，每个结点的潜在表示可以包含自己邻居结点的信息。因此本文采用GCN来完成数据结点的近邻信息传递，得到每个结点新的潜在表示。这个新的潜在表示比原始特征更能够反映数据之间的相似关系，也就是说，它更有利于后续的聚类工作。

GCN的核心思想是通过聚集结点本身的特征以及其邻居结点的特征来生成结点新的特征表示。给定一个空间，在这个空间当中有一张图，图中的每个结点代表一个数据点，其中相邻的结点比不相邻的结点之间的关系更密切，特别地，在该空间当中近邻信息可以流动传播，也就是结点可以传播近邻消息到它的近邻结点，这样可以使得相邻的结点具有更接近的特征关系，更有利于接下来的聚类工作。

GCN当中每个结点每个时刻的变化都是它的相邻结点把信息传播过来之后再进行聚合的结果。本文正式采用这种思想，在本文当中，本文将构造数据潜在表示的过程分为两个阶段：在第一个阶段，构造每个顶点的KNN图并计算顶点特征，捕捉数据之间的拓扑结构，此过程主要是用来构造相似矩阵以及初步的潜在表示，便于用到接下来的GCN工作当中；在第二个阶段，采用GCN传播顶点（数据）的近邻信息，此过程是一个迭代的过程，在此过程中，每个结点可以不断的利用当前时刻其邻居结点和其自身的近邻信息作为输入来更新它当前时刻的潜在表示，直到每个结点的潜在表示的变化幅度小于某个阈值，在整个图中的近邻信息的流动传播已经趋于平稳时暂停此过程，得到数据最后的潜在表示，最后将数据的潜在表示用于下一阶段的工作。具体的通过GCN进行潜在特征重构的流程如图3-4所示。



图3-4 GCN特征重构过程

图、顶点与边的介绍如上文2.1.1所述。定义为第个数据点的前个近邻，为第个数据点的个相互近邻，可以由以下公式得到：

（3-1）

#### 图的构造和顶点特征的计算

首先，本文采用余弦相似度构造相似矩阵。余弦相似度相比于采用径向基核得到的高斯相似度，更能捕捉到向量之间的关系，而高斯相似度只能捕捉到距离因素。余弦相似度的计算如下：

（3-2）

然后，需要根据近邻关系来计算每个顶点的特征，顶点特征计算此时的目标是从相似矩阵中提取出近邻信息进行编码，然后得到新的相似矩阵。为了避免歧义，这个新的相似矩阵用表示。在计算相似矩阵时，本文采用个近邻进行编码，具体而言，相似矩阵为：

（3-3）

此时的是一个非对称矩阵，直观上，相似矩阵应该为对称矩阵，因此本文定义新的。也就是说，的取值为：

（3-4）

此时的相比于获得了近邻的信息，这些近邻的信息可以为后面的聚类工作提供更多的帮助。另外，由于的对称性和稀疏性，它会比只用作为相似矩阵更好。

定义为此时数据集的新的潜在表示，其中是的新的特征表示，它可以从对称相似矩阵的第行提取，也就是说：

（3-5）

这种近邻编码后的特征在聚类当中比原始特征表现得更好，因为它可以反映处于同一密度区域的数据之间的关系。因此接下来的工作中不在采用原始特征，而是采用基于近邻相似度构造的顶点特征。

#### 近邻信息传递

此时，本文已经得到了包含近邻信息的特征表示矩阵，但是这个特征表示矩阵只包含了部分的近邻信息，它还不能够更好地反映出处于同一密度区域内数据之间不是近邻的相似关系，为了更好地反映这一相似关系，本文采用近邻信息传递的思想优化特征表示矩阵。考虑到近邻传递过程中的噪声因素，本文设置一个阈值，在近邻以内的数据点的近邻信息才会被传递。注意，并且。在这里，定义图中边的权重为：

（3-6）

近邻信息传递过程的本质是特征聚集的过程，近邻信息传递之后可以加强位于同一密度的数据点之间的联系，得到更有利于聚类的特征表示。本文采用GCN的思想迭代的更新每个数据点的潜在特征表示，在每次迭代过程中，数据点可以利用自己本身的特征以及自己邻居结点的特征来更新自己的特征，这种方法的关键公式描述如下：

（3-7）

其中表示顶点在GCN的第层的特征，是用来计算近邻传播权重的函数，希望利用函数通过边的权重捕捉到近邻之间的关系，增强位于同一密度区域内的近邻之间的相似度，削弱位于不同密度区域内数据点之间的相似度，可以采用，其中是一个固定值，一般取2。函数是一个聚合函数，一般有、和，本文采用，这样可以捕捉到更多的近邻信息。因此公式（3-7）变为以下形式：

（3-8）

其中，。GCN的层数并不是越多越好，层数过多会引入冗余信息，造成后续的聚类结果过拟合。一般情况下选择比较好。另外在每次近邻消息传播后都采用范数进行正则化。最后会得到每个数据点新的特征表示。具体的算法流程如表3-1所示：

表3-1 近邻信息传播构造潜在特征表示

|  |
| --- |
| 算法输入：数据集，超参数和，GCN层数  算法输出：数据集新的特征表示 |
| 算法步骤：   * 根据余弦相似度构造相似矩阵； * 根据公式（3-4）计算； * 由公式计算； * 令； * 根据公式（3-6）计算边权重； * 采用迭代方法传播近邻信息得到最终的特征表示； |

### 基于共享近邻的相似矩阵构造

在经过近邻信息传播后，得到了更有利于聚类的数据潜在特征表示，此潜在特征表示增强了位于同一密度区域内的近邻数据之间的关系，弱化了位于不同密度区域内的非近邻数据之间的关系。此时，本文采用了一种共享近邻的思想来利用得到的潜在特征表示，构造出相似矩阵用于后续的谱聚类。共享近邻算法改进了采用KNN构造的相似矩阵只考虑聚类因素而不考虑数据是否位于同一密度区域内的因素。具体的共享近邻思想如图3-2所示，图中A和B的7个近邻当中有4个是共享的（用红球表示），A和C的7个近邻当中有两个是共享的（用蓝球表示）。

为了更好的理解共享近邻，采用公式化方法，数据点和之间的共享近邻表示为和近邻集合的交集中元素的个数。采用共享近邻的思想可以得到新的构造相似矩阵的核函数：

（3-9）

其中表示到它的第7个近邻的欧式距离，这样做是为了解决高斯核函数中的不确定性，另外这样可以捕获到近邻的信息。

对于的解释如下：

1）是用来表示空间位置相近的两点之间的相似性。

2）当时，可以通过来比较两个数据点在不同密度情况下的相似性，加一是为了平滑，防止出现除0的情况发生。

基于共享近邻的相似矩阵的具体流程如表3-2所示：

表3-2 通过共享近邻思想构造相似矩阵

|  |
| --- |
| 算法输入：数据集新的*H*，近邻个数，阈值  算法输出：相似矩阵 |
| 使用KNN算法计算出的个近邻;  使用KNN算法计算出的个近邻; |

表3-2 通过共享近邻思想构造相似矩阵（续）

|  |
| --- |
| 算法输入：数据集新的*H*，近邻个数，阈值  算法输出：相似矩阵 |
| 计算两个近邻集合的交集，并计算 |

其中，参数过大会导致所有样本聚为一簇，过小会导致每个样本点独成一簇，因此需要选择合适的值。和按照前文所述分别取和到它们各自的第7个近邻的欧式距离。

### 求解

在3.2.3节当中本文得到了每个视角更符合实际情况的相似矩阵，可以利用这些相似矩阵得到每个视角的转移概率矩阵，此时这些转移概率矩阵包含了许多近邻的信息，但是它仍然被少量的噪声污染，在转移概率矩阵的低秩性和噪声矩阵的稀疏性约束下，本文得到公式：

（3-10）

其中表示的秩，范式表示中非零元素的个数，表示全为1的向量，是一个非负的权衡参数。注意约束强调是一个转移概率矩阵，也就是说它的每一行都是一个概率分布。

由于非凸秩和范式的存在，（3-10）对应的优化问题一般是NP难的。一种流行的解决方法是用核范数来代替，用范式来代替，得到下面的凸优化问题：

（3-11）

矩阵的核范数，计算方式简单，即将奇异值分解后的奇异值相加，它可以用来约束矩阵是低秩的并且是凸的。矩阵的范式，它是矩阵范式的天然凸代替。

采用增广拉格朗日乘子法对优化问题（3-11）求解然后采用交替方向最小化算法进行优化。

对于优化问题（3-11），为了使得变量和可分离，引入新的辅助变量，将优化问题（3-11）转换为一下等价形式：

（3-12）

优化问题（3-12）的相应增广拉格朗日函数是:

（3-13）

其中，是拉格朗日乘子，，表示一个自适应惩罚系数。

求解优化问题（3-13）有三个变量,和,同时考虑这三个变量的优化问题是非凸的，但是它们之间是非相关的，是变量可分离的，对于每一个变量，优化问题（3-13）是凸优化问题，因此可以采用交替方向最小化进行求解。

#### 求解

其他变量固定时，求解的子问题是：

（3-14）

这个公式可以由奇异值阈值算法来解决。更具体的说，令为的SVD形式，公式（3-14）的求解如下所示：

（3-15）

其中是收缩算子。

#### 求解

求解的子问题可以简化为：

（3-16）

这个问题有闭式解。

#### 求解

当其他参数固定时，本文通过求解一下公式来更新：

（3-17）

对于每一个表达式，本文定义：

（3-18）

然后通过简单的线性代数可以将（3-18）中的问题改写为：

（3-19）

其中是矩阵的第i行，也就是说，（3-19）中的问题可以分解为n个独立的子问题：

（3-20）

每一个子问题可以通过求解概率单纯性约束的投影算法有效地求解[20]。

#### 更新乘子：

直观上来说，乘子的更新与等式约束的违规程度成正相关，所以更新乘子可以通过：

（3-21）

整个构造转移概率矩阵的算法如表3-3所示：

表3-3 转移矩阵构造算法

|  |
| --- |
| 算法输入：  算法输出：， |
| 初始化：  *:*  根据公式（3-18）计算    根据[20]中的投影算法更新    根据公式（3-16）更新  根据公式（3-15）更新  令    根据公式（3-21）更新  令 |

## 本章小结

本章简要地介绍了多视角谱聚类的背景，多视角谱聚类由于各个视角之间互相提供互补信息提高了聚类性能，因此受到了广泛的关注。本文发现基于低秩稀疏分解思想的多视角聚类由于解决了噪声污染的问题，往往可以取得不错的效果。因此本章重点介绍了基于低秩稀疏分解的鲁棒多视角谱聚类（RMSC），它将马尔可夫随机游走运用到谱聚类当中，其本质思想类似于低秩稀疏分解。RMSC可以从噪声矩阵中恢复出每个视角共享的转移概率矩阵。

本章对于RMSC当中相似矩阵的构造方法只考虑距离因素而没有考虑位于同一密度区域的因素采用GCN和共享近邻的思想进行改进。首先本文通过GCN重构出数据集新的潜在特征表示，该潜在特征表示比原始数据含有更多的近邻信息，使得位于同一密度区域的数据更加紧凑，同时削弱了位于不同密度区域数据的相似度。然后本文通过共享近邻思想利用潜在特征表示为每个视角构造出相似矩阵，利用该相似矩阵得到转移概率矩阵。利用共享近邻思想可以充分利用潜在特征表示中的近邻信息，为每个视角得到更符合实际情况的相似矩阵。最后将由相似矩阵得到的转移概率矩阵运用到RMSC的后续部分。对于本文提出的算法，本文采用了交替方向最小化的增广拉格朗日乘子法进行求解。

# 实验

## 实验设置

### 数据集

本文在六个现实数据集上进行比对。

3Source数据集（3Source）：它由169条新闻组成，由三家新闻机构报道，分别是BBC、Reuters、The Guardian。每条新闻都用6个主题标签中的一个手工标注。

Handwritten digit 2 source数据集（HW）：手写数字数据集， 来自于MNIST手写数字和USPS手写数字，该数据集由10000个样本组成，本文选取其中2000个样本

BBC数据集（BBC）：它是从BBC新闻网站收集的。BBC数据集由685份文件组成。每一份文件被分成四个部分，并用五个主题标签中的一个进行手工注释，也就是说它有四个视角，共有5个类别。

UCI\_digits数据集（UCI\_digits）：该数据集一共有2000个样本， 有十个类别即1-10。它由手写数字图片数据集的不同特征来创建不同视角，本文选取其中三个视角，分别是轮廓相关性特征、强度均值特征和karhnnen-loeve系数特征。部分图像示例如图4-1所示：

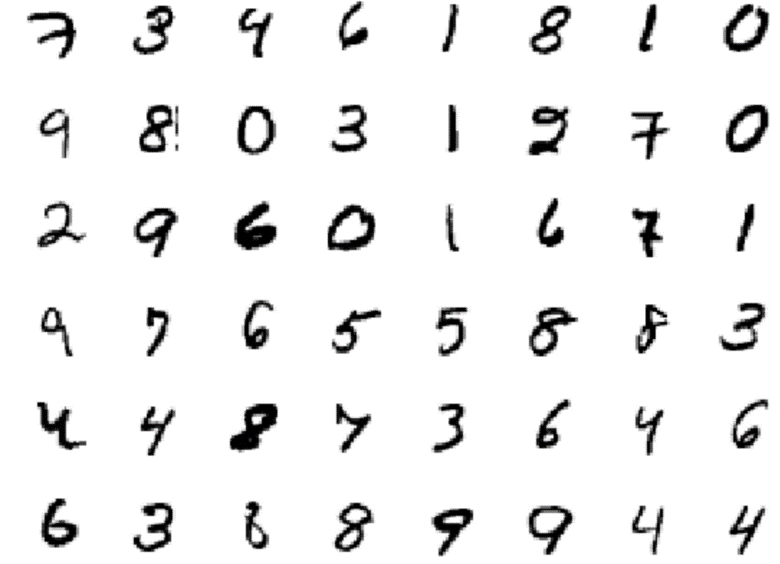


图4-1 UCI手写数字集部分示例图

BBCSport数据集（BBCSport）：该数据集包含从2004-2005两年时间BBCSport新闻网页上对应于5个主题的737个体育新闻文档，共有5个类别，分别是athletic、cricket、football、rugby和tennis。本文选取了544个样本。

Newsgroups数据集（NGs）:NGs是由20个新闻组数据集的子集，它由500个新闻组文档组成，每个原始文档都用三种不同的方法进行预处理（提供三个视角），并使用五个主题标签中的一个进行注释（五个类别）。

实验使用的具体数据集如表4-1所示:

表4-1 多视角数据集

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 实例数 | 视角个数 | 簇个数 | 视角1 | 视角2 | 视角3 | 视角4 |
| 3Source  HW  BBC  UCI\_digit  BBCSport  NGs | 169  2000  685  2000  544  500 | 3  2  4  3  2  3 | 6  10  10  5  5  5 | 3560  784  4659  240  3183  2000 | 3631  256  4633  76  3203  2000 | 3068  \  4665  6  \  2000 | \  \  4684  \  \  \ |

### 聚类指标

本文通过下面的4个聚类指标来评价聚类的效果。其中，除了调整兰德系数的取值范围为[-1, 1]外，其他三个聚类指标的取值范围都在[0,1]。另外，四个聚类指标的结果都是越大越好。

NMI（归一化互信息）：互信息可以用于度量聚类结果与真实值的差异。NMI将互信息缩放到[0, 1]，可以更好地比较聚类算法的好坏。

ACC（准确率）：它主要用于分类任务，在聚类数据有标签的情况下，可以进行改进用于聚类任务，表示为聚类中分簇正确的数据个数占总样本个数的比例。

F-score：查准率用来判断判断为真的有多少是对的，查全率用来判断有多少正类被查出来。F-score用来平衡这两个指标，是它们两个的加权调和平均。

Adj-RI（调整兰德系数）：对于随机划分的结果，兰德系数不能够保证结果近似为0，wield实现此目标，调整兰德系数被提出，它的取值范围为[-1, 1]，比兰德系数有更高的区分度。

### 对比算法

本文主要采用五个基准算法与本文提出的GCN-RMSC进行比较，分别是：

最佳单视角（Best Single View:**BSV**）：使用在单个数据视角中实现最佳谱聚类性能的那个视角；

特征连接（Feature Connection:**FC**）：连接每个视角的特征，然后直接对该连接的特征表示执行谱聚类；

核加权（Kernel Addition:**KA**）：为每个视角构造核矩阵，然后采用不同的权重得到单一的核矩阵，将此核矩阵用于聚类；

协同正则谱聚类(**Co-Reg**)：谱聚类的协同正则化方法。

基于低秩稀疏分解的鲁棒多视角谱聚类（**RMSC**）：本文主要参考的算法。这些算法的前期工作各不相同，但最后都需要运行k-means算法得到最终的聚类结果。

## 实验结果分析

### 实验结果

为了验证谱聚类算法的有效性，本文在双月型数据集和双环数据集这两个典型的非凸型数据集上进行实验，如图4-2和图4-3所示：

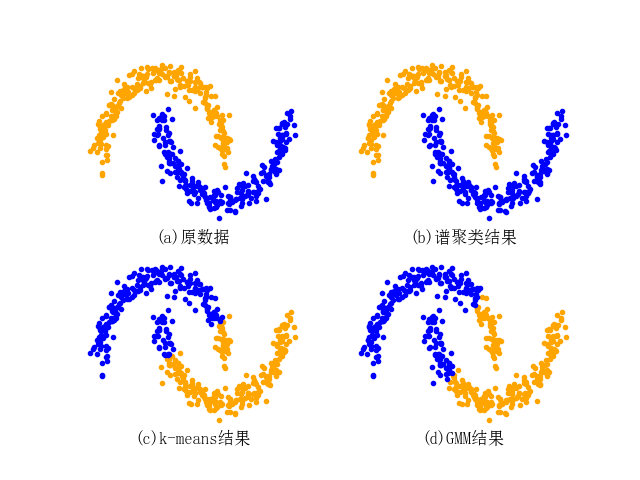


图4-2 双月数据集聚类结果比较

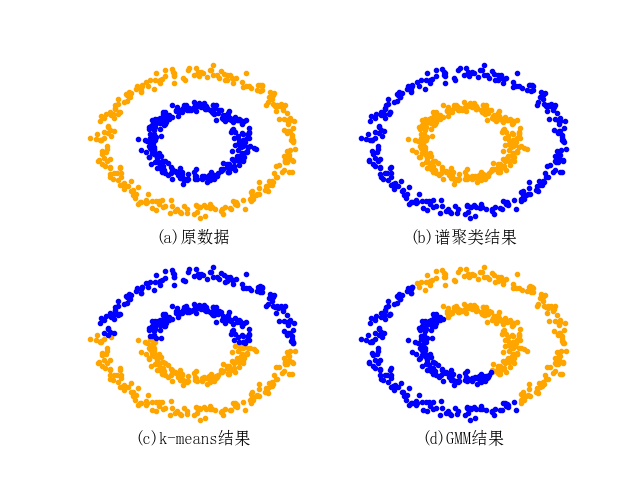


图4-3 双环数据集聚类结果比较

在图4-2和4-3中，分别以蓝色和橙色表示两类数据。（a）表示原始数据集的类型。（c）和（d）表示传统聚类算法的结果，可以看出k-means和GMM这两个传统的聚类算法陷入了局部最优解，聚类效果不好。图（b）表示谱聚类的结果，可以看出谱聚类得到了全局最优解，聚类效果不错。

本文所提出的GCN-RMSC算法共有6个参数，在采用GCN重构潜在特征过程中需要设置近邻个数、近邻个数和GCN层数，在采用共享近邻计算相似矩阵时需要设置近邻个数、阈值和计算高斯核函数时的标准差。为了简化模型的复杂度，避免大量的调参工作，本文采用了折中的思想。在近邻信息传播阶段当中，如果和的值设置的太小就无法捕捉到近邻信息，如果和的值设置的过大就会引入噪声，因此本文将近邻个数和分别设置为40和15；如果GCN层数过小，传播的信息会不足，如果GCN层数过大，会造成过拟合，因此本文所采用的的GCN层数为3。在采用共享近邻构造相似矩阵阶段，本文将近邻个数当做可变超参数进行调参；一种直观上的感觉是阈值应该小于近邻个数，否则会引入噪声，在本文当中阈值被固定为30；标准差被设置为到数据点第7个近邻的欧几里得距离，这样可以捕获到更多的近邻信息，同时能够避免参数的调整。经过折中，本文固定了5个参数，只需要调整利用共享近邻计算相似矩阵阶段的近邻个数，的调整范围为50到150。具体的结果如表4-2所示：

表4-2 六个数据集的聚类结果比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 方法 | NMI | ACC | F-score | Adj-RI |
| 3Source | BSV  FC  KA  Co-Reg  RMSC  Ours | 0.489(0.048)  0.440(0.052)  0.426(0.033)  0.497(0.054)  0.603(0.012)  **0.676(0.019)** | 0.531(0.043)  0.517(0.039)  0.517(0.032)  0.541(0.062)  0.637(0.017)  **0.797(0.018)** | 0.475(0.042)  0.435(0.049)  0.434(0.030)  0.481(0.054)  0.603(0.029)  **0.767(0.024)** | 0.344(0.054)  0.291(0.061)  0.293(0.040)  0.351(0.070)  0.509(0.019)  **0.703(0.014)** |
| HW | BSV  FC  KA  Co-Reg  RMSC  Ours | 0.468(0.009)  0.635(0.016)  0.629(0.013)  0.682(0.016)  0.743(0.012)  **0.834(0.025)** | 0.540(0.020)  0.704(0.038)  0.675(0.032)  0.754(0.039)  0.771(0.025)  **0.895(0.046)** | 0.410(0.013)  0.590(0.022)  0.570(0.019)  0.643(0.026)  0.704(0.020)  **0.830(0.041)** | 0.344(0.014)  0.543(0.025)  0.521(0.022)  0.602(0.029)  0.671(0.022)  **0.810(0.047)** |
| BBC | BSV  FC  KA  Co-Reg  RMSC  Ours | 0.488(0.035)  0.438(0.033)  0.433(0.028)  0.500(0.044)  0.538(0.043)  **0.741(0.031)** | 0.504(0.047)  0.508(0.031)  0.523(0.020)  0.552(0.044)  0.604(0.021)  **0.888(0.021)** | 0.462(0.036)  0.433(0.033)  0.442(0.025)  0.483(0.043)  0.596(0.037)  **0.825(0.036)** | 0.326(0.050)  0.290(0.043)  0.301(0.033)  0.355(0.055)  0.484(0.025)  **0.773(0.035)** |
| UCI\_digit | BSV  FC  KA  Co-Reg  RMSC  Ours | 0.659(0.026)  0.557(0.016)  0.784(0.012)  0.802(0.018)  0.834(0.039)  **0.892(0.028)** | 0.728(0.048)  0.539(0.038)  0.826(0.069)  0.833(0.042)  0.868(0.078)  **0.943(0.037)** | 0.618(0.038)  0.451(0.017)  0.748(0.034)  0.796(0.031)  0.816(0.067)  **0.903(0.034)** | 0.575(0.042)  0.388(0.020)  0.720(0.039)  0.773(0.035)  0.795(0.075)  **0.892(0.045)** |
| BBCSport | BSV  FC  KA  Co-Reg | 0.721(0.002)  0.680(0.050)  0.620(0.022)  0.720(0.004) | 0.854(0.001)  0.726(0.033)  0.741(0.031)  0.843(0.028) | 0.769(0.002)  0.666(0.024)  0.668(0.019)  0.770(0.007) | 0.699(0.002)  0.559(0.028)  0.565(0.022)  0.700(0.007) |

表4-2 六个数据集的聚类结果比较（续）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 方法 | NMI | ACC | F-score | Adj-RI |
| BBCSport | RMSC  Ours | 0.817(0.023)  **0.905(0.022)** | 0.898(0.934)  **0.971(0.026)** | 0.872(0.043)  **0.940(0.011)** | 0.833(0.054)  **0.922(0.035)** |
| NGs | BSV  FC  KA  Co-Reg  RMSC  Ours | 0.390(0.031)  0.293(0.022)  0.228(0.036)  0.468(0.016)  0.325(0.009)  **0.880(0.027)** | 0.595(0.050)  0.544(0.054)  0.457(0.043)  0.703(0.015)  0.569(0.013)  **0.962(0.036)** | 0.452(0.030)  0.403(0.027)  0.354(0.029)  0.537(0.017)  0.442(0.011)  **0.925(0.039)** | 0.312(0.043)  0.254(0.033)  0.191(0.041)  0.421(0.021)  0.302(0.014)  **0.907(0.041)** |

对于每个结果，本文进行20次训练，并按照均值（方差）的形式记录下来。如表4-2所示，在3Source数据集上，本文采用的值为60，可以看出聚类指标F-score、NMI、Adj-RI和ACC相较于最好的基本算法分别提升了16.4%、7.3%、19.4%和16.0%。在HW数据集上，本文采用的值为64，可以看出聚类指标F-score、NMI、Adj-RI和ACC相较于最好的基本算法分别提升了12.6%、9.1%、13.9%和12.4%。在BBC数据集上，本文采用的值为110，可以看出聚类指标F-score、NMI、Adj-RI和ACC相较于最好的基本算法分别提升了22.9%、20.3%、28.9%和28.4%。在UCI\_digit数据集上，本文采用的值为140，可以看出聚类指标F-score、NMI、Adj-RI和ACC相较于最好的基本算法分别提升了8.7%、5.8%、9.7%和7.5%。在BBCSport数据集上，本文采用的值为76，可以看出聚类指标F-score、NMI、Adj-RI和ACC相较于最好的基本算法分别提升了6.8%、8.8%、8.9%和7.3%。在NGs数据集上，本文采用的值为90，可以看出聚类指标F-score、NMI、Adj-RI和ACC相较于最好的基本算法分别提升了38.8%、41.2%、48,。6%和25.9%。根据比较结果可以看出本文所提出的方法在六个数据集上的四个指标都有明显的提升，尤其是在BBC数据集和NGs数据集上各项指标得到了极大地提升，证明了本文所提出算法的有效性。

为了更好地比较NN-RMSC与基准算法，本文采用柱状图进行可视化，具体结果如图4-4所示。由图4-4可知，本文所提出的算法的有效性。需要注意的是，所有的算法都是针对具体的数据集的，不可能存在一个算法对所有的数据集都有效，但是，由图可知，在核加权和特征连接可能会造成过拟合，同时引入噪声因素，聚类效果甚至低于最好视角，协同正则解决了过拟合情况，得到了较好的结果，RMSC考虑了噪声的因素，得到了更好的结果，本文所提出的GCN-RMSC在构造相似矩阵时，在RMSC的基础上考虑了近邻的因素，因此得到了更好的结果。

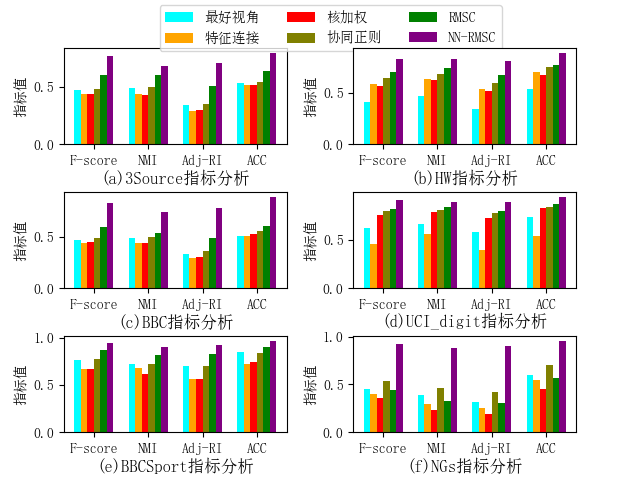


图4-4 六个数据集的聚类结果对比

### 参数敏感性分析

本文所提出的GCN-RMSC算法共有6个参数，在采用GCN重构潜在特征过程中需要设置近邻个数、近邻个数和GCN层数，在采用共享近邻计算相似矩阵时需要设置近邻个数、阈值和计算高斯核函数时的标准差。在实验过程中，本文采用折中的思想，固定参数近邻个数设置为40、近邻个数设置为15、GCN层数设置为3、阈值设置为30以及标准差设置为到数据点第7个近邻的欧几里得距离，本文只需要对超参数近邻个数进行调参。以BBCSport数据集和3Source数据集为例进行参数敏感性分析，聚类指标采用了F-score和准确率（ACC），在BBCSport数据集中，近邻个数的设置范围为60-90；在3Source数据集中，近邻个数的设置范围为40-70。参数敏感性的实验结果如图4-5所示。由图4-5可知，BBCSport数据集和3Source数据集的F-score和准确率（ACC）的整体趋势都是先上升后下降的。根据实际情况，当近邻个数过小时，捕捉不到近邻信息，聚类效果不好，当近邻个数过大时，引入太多的噪声，聚类效果也不好，因此本文的聚类结果符合实际情况。另外，刚开始聚类评价指标的值随着近邻个数的值迅速增长，然后会有一段的近似平稳期，因此本文的算法对参数是不敏感的，在最优值附近的结果比较平缓。

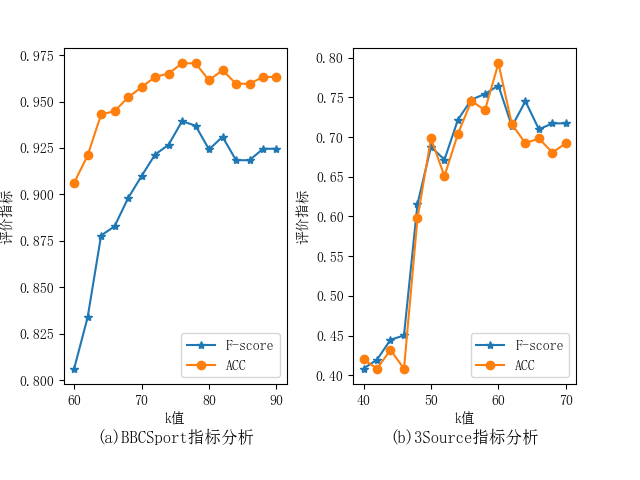


图4-5 参数敏感性分析

除了本文提出算法含有的六个超参数外，RMSC也有一个需要调整的超参数，现在对该超参数进行参数敏感性分析。首先需要固定GCN-RMSC当中的六个超参数为它们各自的最好值，接着本文以BBCSport数据集和3Source数据集为例对超参数进行参数敏感性分析，其中聚类指标采用了F-score和准确率（ACC），的取值为[0.0001, 0.001, 0.01, 0.1, 1, 10, 100]，最后将结果进行可视化。超参数的参数敏感性分析的具体结果如图4-6所示：

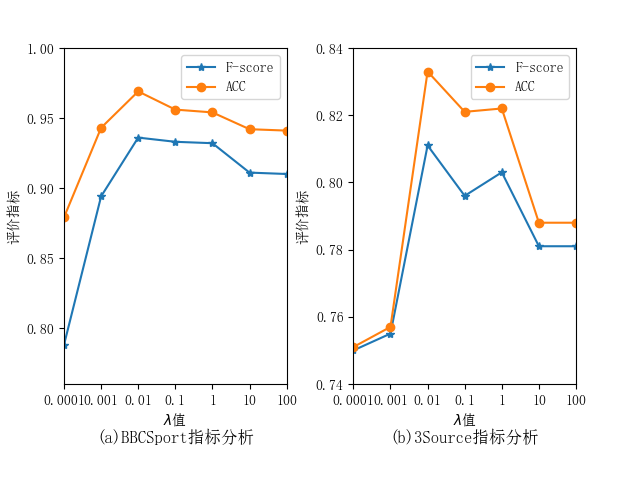


图4-6 参数敏感性分析

由图4-6可知，本文所提出的算法GCN-RMSC对于超参数的变化，其聚类指标整体上也是先增大后减小，并且在超参数的最优值附近，GCN-RMSC对超参数是不敏感的。

### 收敛性分析

最后本文以BBCSport数据集和3Source数据集为例进行GCN-RMSC算法的收敛性分析，其具体结果如图4-6所示：

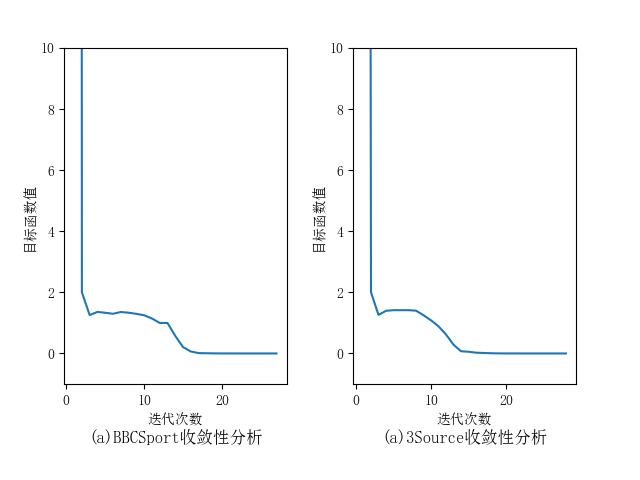


图4-7 收敛性分析

由图4-7可知，在迭代1次后算法就开始快速收敛，在之后目标函数值变化比较平缓，最后趋于稳定，因此，本文的算法是可收敛的。

## 本章小结

本章介绍了GCN-RMSC算法的实验部分，将本文所提出的算法GCN-RMSC与五个基准算法进行比较，实验结果表明，本文所提出的算法优于原有的RMSC算法。另外本章在BBCSport数据集和3Source数据集上对GCN-RMSC算法进行了参数敏感性分析和收敛性分析，实验结果表明，GCN-RMSC算法是对超参数近邻个数和超参数是不敏感的，而且在迭代过程中，本文算法的损失函数会急剧下降，最后趋于稳定。

# 结论与展望

## 结论

本文在学习了多视角谱聚类的基础上，对谱聚类和多视角学习进行了简单地总结。本文重点介绍了基于低秩稀疏分解的鲁棒多视角谱聚类（RMSC），它是一种关于马尔科夫链的多视角谱聚类方法，其核心思想类似于低秩稀疏分解。RMSC考虑到了噪声对重构出共享转移概率矩阵的影响，其聚类结果优于原先所提出的多视角聚类算法。但是RMSC的相似矩阵构造过程中只考虑到了距离的因素，没有考虑到位于同一密度区域的因素，这与实际情况不相符。基于此，本文提出了一种新颖的基于近邻思想的鲁棒多视角谱聚类（GCN-RMSC），其主要步骤为：首先通过图卷积神经网络进行特征重构，得到数据集新的潜在特征表示，该潜在特征表示已经通过近邻信息的更新更有利于后续的聚类工作，它增强了位于同一密度区域的数据的近邻关系，而弱化了位于不同密度区域内的非近邻数据之间的关系。然后本文通过共享近邻思想利用潜在特征表示为每个视角构造出相似矩阵，再通过相似矩阵得到转移概率矩阵。利用共享近邻思想可以充分利用潜在特征表示中的近邻信息，得到更符合实际情况的相似矩阵和转移概率矩阵。最后，考虑到噪声的因素，可以采用低秩稀疏分解的思想，从噪声矩阵的影响中恢复出各个视角共享的转移概率矩阵，该共享转移概率矩阵已经十分符合实际情况，因此可以将此共享转移概率矩阵用于最后的谱聚类当中。

本文所提出的GCN-RMSC算法共有六个参数，分别是在采用GCN重构潜在特征过程中需要设置的近邻个数、近邻个数和GCN层数，在采用共享近邻计算相似矩阵时需要设置的近邻个数、阈值和计算高斯核函数时的标准差。为了简化模型，本文利用折中思想，固定了五个参数，只需要调整超参数近邻个数。在实验当中，本文采用了五个baseline算法，选取了六个数据集验证，实验结果表明，本文所提出的算法优于原先的算法RMSC。最后本文进行了参数敏感性分析和收敛性分析，实验结果显示，在最优值附近，算法对于超参数近邻个数和参数是不敏感的，算法在一次迭代后就会急剧下降，最后趋于稳定，本文的算法是收敛的。

## 展望

RMSC解决了多视角聚类存在噪声的问题，本文在RMSC的基础上解决了RMSC的相似矩阵构造不合理的问题，将近邻信息传播和共享近邻的思想引入到相似矩阵的构造过程中，虽然取得了不错的结果，但是多视角谱聚类是当前比较火热的方向，仍然有一些尚待解决的问题：

1）大规模样本集和数据高维度问题。由于谱聚类需要计算矩阵的特征值，因此数据集过大和数据维度过高会导致过高的时间复杂度，目前的多视角谱聚类只能够处理小规模数据集，如何将多视角谱聚类扩展到大规模问题是未来的热点。

2）视角不完整和丢失数据问题。目前对于此问题采用的是丢弃整个视角数据或采用周围数据进行填充，但是这两个都不合理，如何处理丢失的数据也是未来的研究热点。

3）深度学习问题。深度学习在各个领域都表现出了较好的性能，但是很少有将聚类算法和深度学习结合起来的算法，更不用说多视角聚类了。未来的研究热点是将聚类和深度学习结合在一起。

4）多解问题。对于不同的划分准则，聚类算法会得到不同的聚类结果，如何选择出最符合实际需要的聚类结果或者返回所有可能的聚类结果也是未来的研究热点。

# 参考文献

[1] 陶红. 多视角数据分析算法研究 [D] . 长沙：国防科学技术大学, 2014.

[2] 孔阿栋. 多视角子空间学习方法及其应用研究 [D] . 大连：大连理工大学, 2018.

[3] Xu C, Tao D, Xu C. A Survey on Multi-view Learning [J/OL]. arXiv, 2013, 1304.5634. https://ui.adsabs.harvard.edu/abs/2013arXiv1304.5634X.

[4] 邢洁清, 符传谊. 谱聚类算法及其研究进展 [J]. 电脑知识与技术, 2016, 12(19): 159-61.

[5] 李玲俐. 谱聚类算法及其应用综述 [J]. 软件导刊, 2016, 15(07): 54-56.

[6] Von LU. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4): 395-416.

[7] 唐静静, 田英杰. 多视角学习综述 [J]. 数学建模及其应用, 2017, 6(03): 1-15+25.

[8] Chao GQ, Sun SL, Bi JB. A Survey on Multi-View Clustering [J/OL] . arXiv, 2017, 1712.06246. https://ui.adsabs.harvard.edu/abs/2017arXiv171206246C.

[9] 贺艳芳, 邵亚丽, 向志华. 基于谱聚类的多视角聚类算法 [J]. 河南教育学院学报(自然科学版), 2018, 27(01): 15-18.

[10] 黄宗超, 王思为, 祝恩等. 基于子空间融合的多视图聚类算法 [J]. 郑州大学学报(理学版), 2021, 53(01): 68-73.

[11] Hao WY, Pang SM, Zhu JH, et al. Self-Weighting and Hypergraph Regularization for Multi-view Spectral Clustering [J]. IEEE Signal Processing Letters, 2020, 27: 1325-1329.

[12] Kumar A, Rai P, Daume H. Co-regularized multi-view spectral clustering [M]. Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011: 1413–1421.

[13] Kumar A, Daume H. A co-training approach for multi-view spectral clustering [M]. Proceedings of the 28th International Conference on International Conference on Machine Learning. 2011: 393–400.

[14] Wen J, Sun H, Fei L, et al. Consensus guided incomplete multi-view spectral clustering [J]. Neural Netw, 2021, 133: 207-219.

[15] 宋艳, 殷俊. 基于共享近邻的多视角谱聚类算法 [J]. 计算机应用, 2020, 40(11): 3211-3216.

[16] 杨金鸿, 邓廷权. 基于距离度量学习的半监督多视角谱聚类算法 [J]. 四川大学学报(工程科学版), 2016, 48(01): 146-151.

[17] 程士卿, 郝问裕, 李晨等. 低秩张量分解的多视角谱聚类算法 [J]. 西安交通大学学报, 2020, 54(03): 119-125+133.

[18] 胡世哲, 娄铮铮, 王若彬等. 一种双重加权的多视角聚类方法 [J]. 计算机学报, 2020, 43(09): 1708-1720.

[19] Zhang CQ, Hu QH, Fu HZ, et al. Latent Multi-view Subspace Clustering [J]. Proc Cvpr Ieee, 2017, 4333-4341.

[20] Xia RK, Pan Y, Du L, et al. Robust Multi-View Spectral Clustering via Low-Rank and Sparse Decomposition [J]. Proceedings of the Twenty-Eighth Aaai Conference on Artificial Intelligence, 2014, 2149-2155.

[21] 于心源. 基于马尔科夫判别谱聚类的极化SAR图像分类方法 [D] . 西安：西安电子科技大学, 2019.

[22] 徐冰冰, 岑科廷, 黄俊杰等. 图卷积神经网络综述 [J]. 计算机学报, 2020, 43(05): 755-780.

[23] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs[C]//International conference on machine learning. PMLR, 2016: 2014-2023.