班 级 1413011

学 号 14130110075

西安电子科技大学

# 本科毕业设计论文


题　　目　基于组学数据集成的聚类方法在癌症分类中的实现

学　　院　软件学院

专　　业　软件工程

学生姓名　汪自力

导师姓名　徐悦甦

# 摘 要

随着人类基因组测序项目的完成，人类对于疾病的基因表达有了更深的了解。癌症是一类致死率极高的疾病，因此，针对癌症基因进行探索并且提出准确的治疗和诊断刻不容缓。本文将主要研究聚类算法在癌症的多元组学数据集的应用。

本文首先介绍了 LRACluster、iCluster、SNF 等聚类算法。对于多元组学数据集来说，主要的困难是将不同分布类型的数据处理到同一维度的数据空间。LRACluster 是一种针对多元组学数据集的高效的降维分类、聚类算法。本文选取由 TCGA 上的关于 BRCA、Pan-Cancer 等数据集进行实验。首先，对于 BRCA 来说，实验结果表明其存在两种癌症亚型，这与利用组织细胞进行分类的结果一样。其中，本文提出利用“解释方差”来衡量降维后的低维矩阵与原始矩阵的相关性，当矩阵维度时，曲线出现转折点，因此选择该维度。另外，本文利用 “轮廓系数”来衡量非监督聚类的效果，当聚类个数时，轮廓系数最高故取该值。其次，为了探索该算法在多种癌症类型数据集的分类效果，将 LRACluster 应用在 Pan-Cancer 上，可知当矩阵维度，聚类个数时，LRACluster 效果最好。由结果可知：由组织源细胞进行分类的癌症类型可能被归结为同一个类型。

最后，本文通过将 LRACluster 分别应用在多个癌症类型上。可知对于不同的癌症类型来说其“轮廓系数”不同，这是由于降维后的分子特征的数据矩阵，其与原始数据矩阵相关性较低。因此，在今后的研究需要对分子特征间的联系进行探索，从而提高癌症亚型的效果。

**关键词：** **多元组学数据集** LRACluster **聚类** **概率模型** **梯度下降**

# ABSTRACT

With the completion of the human genome sequencing project, a further understanding of the disease gene expression has been achieved. Since cancer is a type of disease with a very high mortality rate, it is imperative to explore cancer genes and propose accurate treatments and diagnoses. This paper will focus on the application of clustering algorithms in the cancer multivariate data set.

This paper first introduce algorithms of cluster such as LRACluster, iCluster, and SNF. For the multivariate data set, the main difficulty is to process data of different distribution types into the data space of the same dimension. LRACluster is an efficient algorithm for reduction dimension,classification and cluster aiming at multivariate data sets. This paper selects experiments on BRCA, Pan-Cancer, and other data sets from the TCGA. First, for BRCA, the results of the experiment indicate that there are two sub-types of cancer, which is the same as using tissue cells for classification. Besides, this paper proposes the use of "interpretative variance" to measure the correlation between the reduced dimension matrix and the original matrix, and thus the curve appears a turning point when the matrix dimension, so this dimension is chose. In addition, this paper adopts “contour factor” to measure the effects of unsupervised cluster. When the number of clusters is less than 10, the contour factor is the highest. Secondly, in order to explore the classified effects of the algorithm in various cancer type data sets, LRACluster is applied to Pan-Cancer. It can be seen that LRACluster works best when the dimensions of the matrix and the number of clusters are used. From the results, it can be known that the type of cancer which is classified by the tissue-derived cells may be attributed to the same type.

Finally, this paper applies LRACluster respectively to multiple cancer types. It can be seen that for different types of cancers, their "contour coefficients" are different, which is due to the dimensionally reduced data matrix of the molecular features, and the relevance to the original data matrix is relatively less. Therefore, future research needs to explore the links between molecular features to improve the effect of cancer sub-types.

**Keywords: Multivariate data set LRACluster Clustering Probability Model Gradient Descent**

.

# 目 录

**第一章 绪论 1**

**1.1 研究背景及意义 1**

1.1.1 课题背景 1

1.1.2 研究意义 2

**1.2 国内外研究现状 3**

**1.3 本文主要研究内容及工作安排 4**

1.3.1 主要研究内容 4

1.3.2 本文的工作安排 4

**第二章 相关基础知识 7**

**2.1 聚类算法 7**

2.1.1 聚类算法概述 7

2.1.2 聚类算法简介 7

**2.2 K-Means 聚类 8**

2.2.1 K-Means 算法概述 8

2.2.2 伪代码及算法思想 8

**2.3 ConsensusClustering 算法 8**

**2.4 多元组学数据的介绍 9**

**2.5 癌症分类方法的介绍 10**

2.5.1 癌症分类方法概述 10

2.5.2 直接聚类 10

2.5.3 集群聚类 11

2.5.4 规制一体化聚类 12

**第三章 基于数据集成的低维矩阵近似聚类算法 13**

**3.1 低维矩阵近似（LRA）的介绍 13**

**3.2 低维矩阵近似聚类(LRACluster)算法的介绍 13**

3.2.1LRACluster 概况 13

3.2.2 概率模型及 LRACluster 介绍 14

**3.3 构建关于 LRA 的凸函数 15**

**3.4 求解低维矩阵近似的凸函数 16**

3.4.1 相关知识基础.....16
3.4.2 迭代求解的过程.....16
3.4.3 初始化.....17
3.4.4 确定降维后矩阵的维度$r$.....17
3.4.5 矩阵降维.....17
3.4.6 非监督聚类.....18
**第四章 LRACluster 在多元组学数据集上的应用.....19**
**4.1 数据集介绍.....19**
**4.2 对于 LRACluster 性能的探究.....19**
4.2.1 LRACluster 与 iCluster+的对比.....19
4.2.2LRACluster 收敛性的探究.....21
**4.3 LRACluster 在 BRCA（乳腺癌）的应用.....22**
4.3.1 LRACluster 的参数的确定.....22
4.3.2 K-Means 聚类的应用.....23
4.3.3 ConsensusClustering 的应用.....24
**4.4 对于 LRACluster 在 Pan-Cancer（泛癌类型）的应用.....24**
**4.5 将 LRACluster 分别应用于多种癌症类型.....27**
4.5.1 概述.....27
4.5.2 LRACluster 在 COAD（肺癌）上的应用.....27
**第五章 总结与展望.....29**
**5.1 对于 LRACluster 的讨论.....29**
**5.2 总结.....29**
**5.3 展望.....29**
**致 谢.....31**
**参考文献.....33**

# 第一章 绪论

目前，机器学习以及数据挖掘等技术迅速发展，对于生物信息学来说，我们所能掌握的关于各类疾病、基因的数据越来越多，其中对于癌症分类和癌症亚型分类已经成为生物信息学的重要方向。本章将介绍关于目前对于该课题的研究现状以及一些相关背景。

## 1.1 研究背景及意义

### 1.1.1 课题背景

众所周知，从后基因组时代开始，随着对于人类基因组测序项目的完成，以及各种模式下生物基因组测序的完成，生物科学的发展正式进入到了后基因组时代，并且对于基因研究的重心由原来的基因组的结构向基因的功能发展。目前，已经存在诸多的关于生物信息的数据库，通过处理这些生物信息数据，我们可以探索关于生物信息的特征，以此对人类的健康发展有一定的意义。另一方面，随着目前社会的发展，癌症距离人们越来愈近。

癌症[1]是已经成为一个致命疾病，每年要夺走数百万人的生命。由于癌症具有高度遗传异质性，因此，对于探索一种有效的抗癌治疗变得十分困难。根据世卫组织 2011 年的数据统计：癌症现在比所有由冠心病或者是中风造成死亡案例更多。根据目前持续的全球人口和流行病学的转变表明，未来几十年，特别是在中低收入国家（中低收入国家），每年预计将产生 20 多万例新的癌症病例。因此，利用生物信息探索潜在的、可能的癌症的亚型，已经成为一个重要的研究方向，通过探究癌症的亚型，可以提供精准的治疗以及诊断，并且，近些年来，对于特定癌症的类型判别已经出现了一定的方法，以及对于单一癌症类型的亚型判别也有了一定的进步。

快速发展的技术正在逐渐收集多种多样的基因组数据集以解决临床和生物问题。例如，癌症基因组图谱（TCGA）的大规模努力已经为来自数千名患者的超过 20 种癌症积累了基因组，转录组和表观基因组信息。可获得如此丰富的数据使得整合方法对于捕获生物过程和表型的异质性至关重要，从而导致例如鉴定乳腺癌中的同质亚型。数据集成方法需要克服至少三个计算挑战：1.与大量测量相比，少量样本。2.每个数据集的规

模，收集偏差和噪音的差异，3.不同类型数据提供的信息的互补性。目前的整合方法还没有一起解决所有这些挑战。

在目前的临床实践和科学研究中，癌症通常基于其组织/细胞类型来源和病原体进行分类。但是，由于癌症通常涉及复杂的分子改变。因此，基于分子特征的癌症子类型和重新分类对于精确肿瘤学变得越来越重要。基本上基于分子的分类主要有两类:监督分类和无监督聚类。监督聚类一般基于标记数据集探索患者的基因型-表型相互作用以及患病风险。而非监督聚类在识别癌症的亚型方面起到了重要作用。非监督聚类算法在对于癌症亚型聚类在临床应用方面应用的十分广泛。

### 1.1.2 研究意义

精确肿瘤学的一个目标是根据分子特征，而不是其组织来源对癌症进行重新分类。大规模多组学数据的整合聚类是分子癌症分类的重要途径。数据的异质性和组间差异的复杂性是整合聚类分析的两大挑战。

许多癌症基因组项目的一个主要目标是通过对癌症基因组进行全面的基因组分析，通过这来癌症中的关键基因变异并发现治疗靶点。TCGA 研究通过全基因组和全外显子组测序，DNA 拷贝数分析，启动子甲基化谱分析和大量肿瘤中的 mRNA 表达谱揭示了几种癌症类型的遗传情况。CCLE 项目与肿瘤项目相辅相成，编制了近 1000 个人类癌细胞系的遗传学和分子学数据汇编。这些大规模的综合性基因组工作已经着眼于全面而非个别的基因改变，类似于逆向工程过程，其中成千上万的个体癌症基因组已经被探索，以此揭示常见的生物信息。但是由于癌症基因组显示出相当大的异质性，不同个体之间的不同基因发生异常，对鉴定癌症类别具有功能重要性和治疗意义的基因提出了巨大挑战。因此，需要一个相应的工程去实现，整合信息以从大量数据中提取生物学原理，为推进诊断，预后和治疗策略提供有用的见解。

另外，癌症多元组学研究的一个主要目标是使用分子水平的特征来发现可能的癌症亚型，这可以用于更准确的诊断和治疗。TCGA，ICGC 和 CCLE 等几个国际合作项目收集了大量癌症多元组学数据。然而，我们在分析这种大规模癌症多组学数据时仍面临一些挑战：首先，我们需要同时处理不同平台的不同数据类型，如基于计数的测序数据，微阵列连续数据和二进制数据遗传变异。其次，由于数据维度很高和大数据量产生。因此，需要有效且强大的计算算法。

## 1.2 国内外研究现状

首先，参与相同生物过程的分子通常高度相关。人们普遍认为高维癌症基因组数据可以归结为与少数几个主要生物过程相关的低维子空间[2]，如可持续增殖，细胞凋亡抵抗，激活入侵和免疫避免。对此，前人已经做了许多努力来做这种综合分析。

肿瘤的分子复杂性表现在基因组，表观基因组，转录组学和蛋白质组等不同类型的组学数据。在这些多层次的基因组分析应该存在对于肿瘤病因学的综合表征。但是，真正的综合数据分析缺乏有效的统计和生物信息学工具。集成式聚类的标准方法是单独集群，然后再进行手动集成。另外，一种具有统计功能并且有效的方法是将所有数据类型同时并入并生成单个集成的集群然后再进行聚类。为此，前人开发了一个用于集群聚类的联合隐藏变量模型。该方法被称为 iCluster[3]。iCluster 整合了不同数据类型之间关联的建模以及单个框架中数据类型内的方差-协方差结构，同时降低了数据集的维度。其中，似然估计的推论是通过期望最大化算法完成的。

虽然 iCluster 该方法最近被用于一项具有里程碑意义的研究，以预测具有不同临床结果的新型乳腺癌亚型，并且发现拷贝数和基因表达谱的联合聚类解决了仅表达亚组的相当大的异质性。然而，却存在两个问题还没有被解决：首先，现有方法的设计不包括离散（例如体细胞突变）和连续变量，从而限制了利用大规模综合基因组数据集全部潜力的能力。事实上，以前的大多数方法都只着眼于整合拷贝数和基因表达。第二个问题在于：系统无法区分可靠和不变的亚型特征与不可靠特征的癌基因。为了在多种数据类型中找到共享的低维子空间，提出了基于概率主成分分析的潜在模型 iCluster +[4]，该模型使用广义线性模型将连续，离散化和计数变量变换为一组潜在驱动因素的稀疏线性回归。至此，癌症亚型可以在由隐藏驱动因素组成的低维子空间中进行探索。

大多数当前的多源聚类方法要独立地为每个数据源确定一个单独的聚类，要为所有数据源确定一个“联合”聚类。需要更灵活的方法来同时处理数据源的依赖性和异质性。因此，Lock 等人提出了另一种贝叶斯模型（Bayesian consensus clustering，BCC）[5]来同时找到潜在的低维子空间并将样本分配到不同的聚类中，允许为每个数据源分别聚类对象。该方法主要关注集群聚类。聚类是一种广泛使用的探索工具，用于识别类似的对象组（例如，临床相关的疾病亚型）。 已经提出了数百种执行聚类的通用算法。 但是，我们的工作受到需要一种集成式集群方法的启发，该集成方法具有计算可扩展性，并且

对每个数据源的独特功能具有强大的可靠性。这些分离的聚类松散地遵循整体共识聚类，因此它们不是独立的。该方法描述了计算可扩展的贝叶斯框架，用于同时估计共识聚类和源特定聚类。这种灵活的方法比所有数据源的联合聚类更加稳健，并且比独立聚合每个数据源更强大。我们使用来自 The Cancer Genome Atlas 的公开可用数据提出了应用于乳腺癌肿瘤样品的亚型鉴定。

## 1.3 本文主要研究内容及工作安排

### 1.3.1 主要研究内容

对于大规模的组学数据进行集成分析是基于分子进行癌症分类的一种重要方法，由于组学数据中，对于不同的癌症类型存在着不同类型的数据，这些数据服从高斯分布、泊松分布、伯努利分布等概率分布，因此，对于这些多元组学数据，利用这些基于分子特征的数据去探索癌症类型的亚型，主要的问题是，不同类型的数据无法整合成为一个共同的数据矩阵，因此，对于数据的处理是本文的一个研究内容。另外，目前，尚未存在一个高效的、可以处理大数据量的对数据矩阵降维的算法，因此本文基于概率模型，通过对于不同类型的数据利用概率密度函数建立概率模型， 通过概率模型将数据处理到一个共同的具有隐藏参数的矩阵中，利用该矩阵以及最大似然函数构造 LRA（低维矩阵近似），由于构造的函数一般都是凸函数[6]，因此可以通过快速算法进行求解，将通过梯度下降进行求解。最后，可以通过无监督聚类算法进行聚类。其中，无监督算法存在许多聚类算法，我们可以选择 k-means 聚类算法，以及其他聚类算法，对于不同的聚类个数探究不同的效果。

### 1.3.2 本文的工作安排

第一章：本文将对基于的多元组学数据进行聚类分类的研究意义以及研究背景进行阐述。除此之外，本文将从国内外的研究现状来对该课题进行阐述，最后叙述关于本文的主要研究内容，对该课题进行了概述并且提出主要的工作安排。

第二章：本文将对聚类算法进行简介，聚类算法一般包含监督聚类、非监督聚类算法。对于非监督聚类算法着重介绍 K-Means 聚类算法的内容以及优缺点等。其次，本文将对应该课题中的多元组学数据进行介绍，将目前应用于的多元组学数据集的聚类算

法进行概况介绍，目前大多分为以下聚类方法：直接一体化聚类、集群聚类、调节整合聚类。

第三章：本文将对 LRACluster 算法进行详细阐述。首先，对低维矩阵近似的构造进行解释，低维矩阵近似构造的关键是降维后的矩阵与原始矩阵需要具有较强的相关性。对于 LRACluster 来说，主要应用概率模型将不同概率分布的数据映射到相同的数据矩阵中，之后通过对于所有概率进行求和，然后将其作为似然函数，通过利用奇异值分解、梯度下降等方法对这个具有凸函数性质的似然函数进行优化可得降维矩阵。本文将应用非监督聚类算法进行聚类验证，例如 K-Means 聚类算法。最后，对于降维后的矩阵维度以及聚类个数进行探究。

第四章：本文将 LRACluster 应用在 BRCA 以及 Pan-Cancer 等分子特征的数据集上，对 LRACluster 的性能进行分析、对 LRACluster 在单个癌症类型亚型的分类能力进行探究、对 LRACluster 对于多种癌症类型的分类进行探究。

第五章：对 LRACluster 进行总结、讨论，并且通过以上实验对 LRACluster 的不足进行分析，提出相应的改进措施，作出期望。

# 第二章 相关基础知识

## 2.1 聚类算法

### 2.1.1 聚类算法概述

聚类是数据挖掘、机器学习的一个重要研究方向。目前，存在许多不同的聚类算法，不同的聚类算法在不同的领域有许多不同应用。在“无监督学习”中，训练样本的标记信息是不可知的，目标是通过对无标记训练样本的学习来探索数据的内在规律以及性质，这对于数据分析、数据挖掘提供了帮助。在“无监督学习”中，目前研究最多、应用最为广泛的算法便是——“聚类”算法。

### 2.1.2 聚类算法简介

聚类分析是一种基于大数据的分析方法，该技术目前主要被用来分析数据内部之间的关系。聚类算法主要是通过将全部的数据按照它们之间的关联程度分成不同的相似组。处于同一类中的数据项相似性很强，即它们之间的关系很密切，而处于不同聚类中的数据彼此差异很大。目前，聚类分析又被称作无监督学习算法，相比于监督学习来说：监督学习是将数据分成两类：训练集和测试集，然后建立一些模型对训练集数据进行训练，得到合理的参数，从而对测试集数据进行分类。而无监督学习的主要思想则为：通过直接对数据集的内部关系进行探索，直接对数据进行分类。由于该方法不需要训练集数据，因此，目前对于聚类算法的应用越来越广泛。

聚类分析和分类分析它们的作用非常的相似，都是对已知的数据进行分类的一种方法。但是有所不同的是，分类分析是根据已有的数据得出一种模型，它是一种有指导性的学习；然而聚类分析则是针对这样一批原始样本数据，我们不知道这些原始数据的内部分类情况，也不清楚应该分成几类，只是希望通过建立某种模型，来把观测到的数据分类的更合理，使得同一类中的观测数据比较接近，更加的符合事物的特性，不同类的数据之间相差较大，这就是无指导的学习。所以，聚类分析最重要的是对观测数据之间的相似程度和接近程度的理解和描述，这个可以通过定义不同的距离公式和相似性度量产生不同的聚类结果。

## 2.2 K-Means 聚类

### 2.2.1 K-Means 算法概述

基本 K-Means 算法的思想很简单，事先确定常数 K，常数 K 意味着最终的聚类类别数，首先随机选定初始点为质心，并通过计算每一个样本与质心之间的相似度(这里为欧式距离)，将样本点归到最相似的类中，接着，重新计算每个类的质心(即为类中心)，重复这样的过程，知道质心不再改变，最终就确定了每个样本所属的类别以及每个类的质心。由于每次都要计算所有的样本与每一个质心之间的相似度，故在大规模的数据集上，K-Means 算法的收敛速度比较慢。

### 2.2.2 伪代码及算法思想

算法主要思想：

1.初始化常数 K，随机选取初始点为质心；

2.重复计算一下过程，直到质心不再改变：

2.1 计算样本与每个质心之间的相似度，将样本归类到最相似的类中，

2.2 重新计算质心；

3. 输出最终的质心以及每个类。

K-Means 聚类算法伪代码如下页所示。

## 2.3 ConsensusClustering 算法

ConsensusClustering[7]是一种进行分类的非监督聚类算法。该算法是一种适用于探索基因表达的类别和聚类验证的新方法。它可以广泛应用于指导和协助使用任何广泛的可用聚类算法。我们称这个新方法为：consensus clustering(一致性聚类)。通过结合重采样技术，可以多次运行聚类算法，并且得到每次聚类情况的一致性水平，以此来评估聚类的结果的优劣。除此之外，为了衡量该算法在不同的情况下的表现，它可以采用不同的聚类的算法来运行，如：k-means 聚类算法、贝叶斯聚类、SOM 等方法。通过采用模拟数据和真实基因表达数据，可以评估该方法在对于基因表达数据进行分类的效果。

**K-Means 聚类算法伪代码**

输入：样本集 $D=\{x_1,x_2,...,x_m\}$;，聚类簇数 $k$.

1. 从 $D$ 中随机选择 $k$ 个样本作为初始均值向量 $\{\mu_1,\mu_2,...,\mu_k\}$
2. repeat
3. 令 $C_i=\varnothing(1\le i\le k)$
4. for $j=1,2,...,m$ do
5. 计算样本 $x_j$ 与各均值向量 $\mu_i(1\le i\le k)$ 的距离：$d_{ji}=\left\|x_j-\mu_i\right\|_2$
6. 根据距离最近的均值向量确定 $x_j$ 的簇标记 $\lambda_j=\arg\min_{i\in\{1,2,...,k\}}d_{ji}$
7. 将样本 $x_j$ 划入相应的簇 $C_{\lambda_j}=C_{\lambda_j}\bigcup\{x_j\}$;
8. end for
9. for $i=1,2,..k$ do
10. 计算新均值向量 $\mu_i^{'}=\frac{1}{|C_i|}\sum_{\infty\in C_i}x_i$
11. if $\mu_i^{'}\neq\mu_i$ then
12. 将当前均值向量 $\mu_i$ 更新为 $\mu_i^{'}$
13. else
14. 保持当前均值向量不变

## 2.4 多元组学数据的介绍

组学主要包括基因组学，蛋白组学，代谢组学，转录组学，脂类组学，免疫组学，糖组学和 RNA 组学等。各组学就是研究他们各自以及它们之间的关系，例如基因组学这门学科就是研究这些基因以及这些基因间的关系。组学大数据就是这些组学在生物医学等领域中的研究应用所收集到的庞大数据。本文利用多个组学数据中分子特征在多个样本中的基因表达进行分类。

## 2.5 癌症分类方法的介绍

### 2.5.1 癌症分类方法概述

目前，癌症通常基于其组织/细胞类型来源和病原体进行分类。但是，癌症通常涉及复杂的分子改变。因此，利用分子特征对癌症的亚型进行分类越来越重要。一般情况下，基于分子特征的分类有两类：监督分类和非监督分类。监督聚类一般基于标记数据集探索患者的基因型-表型相互作用以及患病风险。而非监督聚类在识别癌症的亚型方面起到了重要作用。

基于大规模多组学数据的非监督聚类是建立以分子为基础的癌症分类的重要途径。但是，目前实现这些方法存在一定的困难：第一个是“数据异质性”：不同的组学数据具有不同的数据分布和变异模式。例如，测序数据通常以基于计数的分布为模型，微阵列数据通常使用高斯分布，变异数据遵循二项分布。另一个困难是：“组间变异的复杂性”：不同的组学数据具有层间规律的共变（例如，基因表达受拷贝数变异和启动子 DNA 甲基化的调控），但每一层也具有它自己的特定变化模式。此外，由于分子特征的数量通常远大于样本大小，因此，这些方法应该考虑高维度的因素。针对上述困难的不同处理方法，一般存在三类聚类方法：直接一体化聚类、集群聚类、调节整合聚类，并且非监督聚类方法在临床阶段有着广泛的应用，利用其进行分类的结果可能与利用病理特征定义的癌症亚型结果不同。

### 2.5.2 直接聚类

整合多组学数据的一个直观的想法是将它们全部放入一个堆积矩阵中，并将该堆积矩阵作为后续聚类分析的输入，为了解决数据异构性问题，需要使用统一的框架来处理所有类型的数据。

一种简单的策略：通过比较肿瘤和相邻正常组织的数据，将所有类型的数据转化为相对分数。首先，他们使用 Cox 回归来获得相关特征，包括拷贝数变异，DNA 甲基化和基因表达。然后，将所有选择的特征归一化，最终，对该矩阵利用 k-means 聚类进行聚类，其中，集群个数采用贝叶斯信息准则(BIC)来选择

另一种策略：将不同类型的数据降维到一个共同的低维矩阵。该方法主要假设是：一些主要的生物因素决定了一系列高维但低秩的参数，并且这些参数决定了组学数据的

产生与分布。例如，iCluster +便是采用这种策略，其通过使用广义线性回归模型来处理不同类型的数据。利用不同的概率函数来构建观测数据和低维潜在变量之间的回归。该方法还使用回归模型中的范数惩罚来处理观测数据与潜在变量之间的稀疏关系。另外，低秩矩阵与非负矩阵分解（NMF）直接相关。目前一般联合 NMF 方法来寻找跨多组学数据集的共享特征矩阵，这是原始高维数据的低维表示。

在将多元组数据转化为低维子空间之后，可以使用经典的 k-means 聚类方法来获得最终的聚类。然而，由于这些潜在因子模型的目标函数通常是非凸的。他们只能使用基于抽样的算法来寻找次优解，这通常是缓慢和不稳定的。随着数据量的快速增长，这些问题将变得更加严重。因此，本文采用一种基于低秩近似的集成聚类方法 LRAcluster[8]，它所构造的目标函数一般是凸函数，因此，该算法可以通过简单的梯度上升算法快速而稳定地找到全局最优。

### 2.5.3 集群聚类

与通过直接堆叠不同类型的多组数据的方法不同，集群聚类(COC)是对单个组学数据集执行聚类分析的另一种方法，该方法通过将初级聚类结果再次进行聚类，之后可以得到最终的集群类型。初级聚类结果可以被认为是特定数据集相似性的中间表示。最终的聚类分配将基于这些相似性做出相应的调整，最终可以得到相应的聚类结果，这便是集群聚类的大致想法。

一个最简单的策略便是：将每个组学数据集的初级聚类的结果利用虚拟编码编制成二进制的变量。例如，如果我们对基因表达数据执行 k-means 聚类并获得三个聚类，属于第一个聚类的样本将被[1,0,0]编码、第二个聚类将为表达成[0,1,0]、第三个聚类结果表示为[0,0,1]。对于每一个数据集重复此过程，我们将每个样本编码到长度为 $\sum_{i=1}^{t} k_i$ 的新的二进制向量，其中 $t$ 是数据集的数量，$k_i$ 是数据集 $i$ 的集群数。这个向量形成了对于原始组学数据初次聚类后的相似性简明表示。基于这个根据初次形成的相似性矩阵，后续聚类可以通过分析实现整体集群分配。

另一个策略是：将每个数据集中的样本相似性直接组合到集成的相似性网络。然后，可以使用基于网络的聚类方法，如社区结构分析，光谱聚类和马尔可夫聚类算法(MCL)来识别最终的聚类。一种称为 SNF(相似性网络融合)的方法首先根据每对患者的“距离”建立了每个数据集的患者相似度网络，最初的步骤是采用相似性度量来为每个可用的数

据类型构建逐个样本的相似性矩阵。矩阵等价于节点为样本(例如患者)的相似性网络，其中加权边表示样本相似度。矩阵和网络都是比较直观的表示:相似矩阵有助于识别全局模式(集群)，而网络表现了节点与节点之间的关系。相似网络融合使用了基于消息传递理论的非线性方法，该方法其迭代地更新每个网络，在进行每次迭代之后更接近于其他网络。经过几次迭代，SNF 将会收敛于单个网络。本文所进行的处理的优点是融合过程中那些弱相似边(低重量边)消失，这有助于降低数据中的噪声，并且在其他网络中增加了一个或多个其他网络中存在强相似的(高权重)边。除此之外，由所有网络支持的低权重边都会被保留，这取决于它们的邻域在网络之间的紧密连接。这种非线性允许 SNF 充分利用网络的局部结构，整合共同信息和互补信息。该方法仅分析 mRNA，miRNA 表达和 DNA 甲基化数据，这些均为实数数据，可以通过欧几里德距离进行测量。除此之外，应使用其他测量距离方法分析其他种类的数据。

COC 的主要考虑是：中间结构或由初次聚类的结果的相似性是否可以保持原始数据集中的聚类信息。而这种转换将会使得在不同组学数据集之间检测共享结构变得更加困难。另一个问题是：如果不同组学数据集的聚类结果之间存在很大差异，则应检查以下共识聚类是否有意义。虽然这个问题几乎存在于各种综合方法中，但对 COC（集群聚类）来说更为严重。应采用一些措施来评估不同聚类结果的一致性，如 Jaccard 相似性，调整后的兰德指数和其他基于信息论的措施。

### 2.5.4 规制一体化聚类

除了在癌症发生和进展中发挥重要作用的癌症驱动分子改变之外，还有许多其他改变同时共存。为了减少这些改变所引起的“噪音”，基于分子的癌症分类应着重于 驱动改变。规制一体化聚类的基本思想是通过考虑不同分子层之间的调节结构来使用驱动变量，这有助于区分驱动变化。

候选癌症驱动因素通常被定义为重要的遗传改变，例如体细胞突变和拷贝数变 异。复发性表观遗传改变，特别是 DNA 甲基化和大规模功能筛选也可用于鉴定新的候选驱动因素。然后，将这些改变或扰动与基因表达变异，遗传改变的功能指征进行综合分析。与直接集成聚类和集群聚类相比，通过规制一体化聚类方法识别的群体的分子机制更为清晰。这种方法更像是用于集成分析的框架或管道，不同于其他的聚类算法。

# 第三章 基于数据集成的低维矩阵近似聚类算法

## 3.1 低维矩阵近似（LRA）的介绍

LRA(low-rank approximation)是指低秩近似[9]，一般对于组学数据来说，数据的维度都远高于样本的分子特征个数。因此，我们需要将原始矩阵进行降维处理，并且要求降维后的矩阵与原始矩阵应具有较强的相关性。目前，LRA 是目前一种理想的降维方法，在大多数的情况下，LRA 构建的目标函数是凸性的，因此，可以采用快速算法来求解。在一些研究中可以发现，该方法对于探索单个癌症类型的亚型癌症的效果突出。

本文应用了一个基于概率模型的低维矩阵近似进行降维处理，这个方法可以处理来自多元组学数据的不同的数据类型，并且具有很高的计算效率和稳定性。对于该方法：它假设一些主要生物因素决定了一组高维但是低秩的矩阵组学参数集，并且数据集与该参数集存在较强的相关性。根据实验结果分析可知：本文所应用的聚类方法 LRAcluster 的运行速度比 iCluster +更快以及更加稳定。因此，这使得可以在小型服务器甚至个人计算机上分析大规模癌症多元组学数据。

## 3.2 低维矩阵近似聚类(LRACluster)算法的介绍

### 3.2.1 LRACluster 概况

LRACluster 聚类方法是一种非监督聚类的方法，LRACluster 主要用于探索大规模的高维度多元组学数据中的低维癌症亚型。

在 LRAcluster 模型中，多个样本在分子特征（如体细胞突变，拷贝数变异，DNA 甲基化和基因表达）上的基因表达数据被表示为多个数据矩阵。本文利用概率模型来构造数据矩阵，在该概率模型情况下，每个分子特征的数据在一定的概率分布下，数据被转化成为一定的概率，并且原始数据集被转换为与其大小相对应的隐藏参数矩阵。 因此，需要利用惩罚函数对该参数矩阵的低秩假设进行约束。通过该概率模型，可以将原始矩阵降维成为低维矩阵，通过该低维矩阵进行癌症分类。

本文将 LRAcluster 应用于由样本在多元组学数据集，该数据集包含 11 种癌症类型，

11 种癌症类型的样本在四种不同分子特征（DNA 甲基化、拷贝数变异、mRNA、体细胞突变）上的基因表达。对于该数据集的癌症亚型探索，这是以前的方法难以处理的。经过 LRACluster 进行聚类后，其中：泛癌分析结果表明：大多数不同的癌症类型（或不同的组织成分）通常可以归入独立的集群中，但是存在一些个例不满足该种情况。 而单一的癌症类型分析结果显示：存在一部门癌症类型可以通过组学数据进行探索癌症亚型。

### 3.2.2 概率模型及 LRACluster 介绍

为了介绍关于本文所应用的概率模型，因此需要以下符号：

1. $X^{(k)}$：代表第 $k$ 种组学数据类型

2. $x_{ij}$ ：第 $i$ 行代表第 $i$ 个分子特征，$j$ 代表第 $j$ 个样本，$x_{ij}$ 代表第 $j$ 个样本在第 $i$ 个分子特征上的基因表达

3. $\Theta^{(k)}$： $\Theta^{(k)}$ 代表与 $X^{(k)}$ 大小相符合的矩阵

该概率模型对于每种数据类型都具有概率密度函数，如下所示：

1.对于正态类型的数据（real-type data），符合高斯分布其关系为：

$$\Pr(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}) \propto \exp(-\frac{1}{2}(X_{ij}^{(k)} - \Theta_{ij}^{(k)})^2) \qquad \text{式（3-1）}$$

一般拷贝数变异以及 DNA 甲基化的数据符合这种概率分布。

2.对于二维类型的数据（binary data），符合伯努利分布其关系为：

$$\Pr(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}) = \frac{e^{\Theta_{ij}^{(k)}}}{1+e^{\Theta_{ij}^{(k)}}} I(X_{ij}^{(k)} = 1) + \frac{1}{1+e^{\Theta_{ij}^{(k)}}} I(X_{ij}^{(k)} = 0) \qquad \text{式（3-2）}$$

一般体细胞突变的数据符合这种概率分布。

3.对于计数数据（count data），符合泊松分布其关系为：

$$\Pr(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}) \propto (\lambda_{ij}^{(k)})^{X_{ij}^{(k)}} e^{(-\lambda_{ij}^{(k)})}, \lambda_{ij}^{(k)} = e^{\Theta_{ij}^{(k)}} \qquad \text{式（3-3）}$$

一般 RNAseq 数据符合这种概率分布。

除此之外，分类数据可以使用虚拟代码进行转换，因此结果可以视为二维变量。通过概率模型，可将原始矩阵转换成符合条件的输入矩阵。如图 3.1 所示：

| 体细胞突变 | | | |
|---|---|---|---|
| | 样本 1 | 样本 2 | 样本 3 |
| 基因 A | 1 | 0 | 0 |
| 基因 B | 0 | 1 | 1 |
| 基因 C | 0 | 1 | 0 |
| **RNA 序列** | | | |
| 基因 D | 样本 1 | 样本 2 | 样本 3 |
| 基因 E | 103 | 96 | 132 |
| 基因 F | 27 | 24 | 35 |
| **拷贝数变异** | | | |
| 基因 G | 样本 1 | 样本 2 | 样本 3 |
| 基因 H | -0.87 | 1.02 | 0 |
| 基因 I | -0.87 | 0 | -0.34 |

Probabilistic model

$$X_{ij} \sim \Pr(X_{ij} \mid \Theta_{ij})$$

| 体细胞突变 | | | |
|---|---|---|---|
| | 样本 1 | 样本 2 | 样本 3 |
| 基因 A | Θ | Θ | Θ |
| 基因 B | Θ | Θ | Θ |
| 基因 C | Θ | Θ | Θ |
| **RNA 序列** | | | |
| 基因 D | 样本 1 | 样本 2 | 样本 3 |
| 基因 E | Θ | Θ | Θ |
| 基因 F | Θ | Θ | Θ |
| **拷贝数变异** | | | |
| 基因 G | 样本 1 | 样本 2 | 样本 3 |
| 基因 H | Θ | Θ | Θ |
| 基因 I | Θ | Θ | Θ |
| 基因 J | Θ | Θ | Θ |

图 3.1 将原始矩阵转换成符合条件的输入矩阵

如图 3.1 所示：左图中：分子特征（如体细胞突变，拷贝数变异，DNA 甲基化和基因表达）被表示为多个观测数据矩阵。根据概率假设：每个样本的每个观察分子特征是一个随机变量，并且以隐含参数为条件。因此，每个观察到的数据矩阵都以大小匹配的参数矩阵为条件，不同类型的数据遵循不同的概率分布。最终，可将原始矩阵转换成符合条件的输入矩阵。

## 3.3 构建关于 LRA 的凸函数

通过以上关于不同组学数据类型的概率分布(密度)函数，可以将它的似然函数得到：对于多元组学数据分析来说，存在不同类型的组学数据集 $X^{(k)}(k=1,2,...,K)$，因此，可以将由所有参数矩阵 $\Theta$ 组成与不同类型的数据集的集合 $\Theta^{(k)}$ 作为数据矩阵，将所有的数据类型的数据矩阵的似然函数求和可得：

$$L(\Theta)=\sum_k L(\Theta^{(k)},;X^{(k)}) \qquad \text{式（3-5）}$$

其中，通过概率模型假设数据集 $X^{(k)}$ 是独立分布在超高维参数矩阵 $\Theta$ 上的。并且，

根据先前的假设可知，$\Theta$具有低维结构的矩阵，因此，最终可将该问题归结为一个优化问题：

$$\arg\min_{\Theta} L(\Theta)+\mu|\Theta|^{*} \qquad \text{式（3-6）}$$

其中，$\mu$代表惩罚参数，$|\bullet|^{*}$代表矩阵的核范数[10]。

## 3.4 求解低维矩阵近似的凸函数

### 3.4.1 相关知识基础

对于优化问题，一般采用奇异值阈值(SVT)方法[11]解决，由于本文所构造的函数是凸函数，因此可以应用该方法进行求解。

### 3.4.2 迭代求解的过程

对于该问题迭代求解的方案可以简单地概述如下：

(1).初始化$\Theta^{0}$，并且一直迭代(2)，(3)直到收敛

(2). $\Theta^{2n+1}=\Theta^{2n}-\delta_n\nabla f$ 式（3-7）

(3). $\Theta^{2n+2}=D_{\mu}(\Theta^{2n+1})$ 式（3-8）

其中，$\nabla f$代表似然函数$L(\Theta)=\sum_k L(\Theta^{(k)},;X^{(k)})$的梯度，$\delta_n$代表梯度下降的固定步长，$D_{\mu}$代表代表奇异值阈值操作。

由于对于$\Theta$的奇异值分解(SVD)可得：

$$\Theta=U\Sigma V^{T} \qquad \text{式（3-9）}$$

推导可得：

$$D_{\mu}(\Theta)=UD_{\mu}(\Sigma)V^{T} \qquad \text{式（3-10）}$$

其中，$D_{\mu}(\Sigma)$是与$\Sigma$维度相同的对角矩阵，并且该矩阵的对角元素值等于$\Sigma$的奇异值的收缩值。对于每一个$\Sigma$的奇异值$\lambda$来说：

如果$\lambda>\mu$时，那么收缩值为$\lambda-\mu$。

如果$\lambda \le \mu$时，那么收缩值为0。

### 3.4.3 初始化

由于LRACluster 的目标函数是凸函数，因此存在一些初始值使其收敛至全局最小来说。LRACluster 将矩阵$\Theta$初始化为一个零矩阵。但是，由于最初的计算框架需要一个由用户定义的惩罚参数$\mu$，但是这个$\mu$很难确定。相比于$\mu$来说，由用户定义的降维后的矩阵维度$r$可以确定。因此，在每次的迭代过程中，$\mu$定义为维度$r+1$矩阵中的最大的奇异值。这个方法可以保证矩阵$\Theta$存在维度$r$的降维矩阵，并且对于收缩值的影响最小。最后，对于$\delta_n$来说，如果是“矩阵完成问题”，那么当$\delta_n$为0.5或者2时，那么该算法一定会收敛。因此，本文采用$\delta_n$（步长）为0.5，该值可以保证在对于LRACluster在研究应用的过程中，该算法可以一定会收敛。

### 3.4.4 确定降维后矩阵的维度$r$

对于LRACluster来说，其中，唯一由用户定义的参数即为确定降维后矩阵的维度$r$。似然函数$L(\theta;X)$对应于需要优化的方案，采用$L_r^*$来指导对于参数$r$的选择。对于相同的数据来说，越大的数据维度$r$导致降低模型自由度的惩罚效果和数据拟合效果(降维后的数据矩阵与初始矩阵的相关性)。因此，$L_r^*$代表降维后数据矩阵的拟合效果，并且，$L_{r=0}^*$是最小值，$L_{r=+\infty}^*$是最大值。但是，由于对于LRACluster来说，它的数据维度一般很大，因此$L_r^*$的值一般会很大，因此效果较差，所以本文将$L_r^*$的值进行归一化，利用$\frac{L_{r=+\infty}^* - L_r^*}{L_{r=+\infty}^* - L_{r=0}^*}$来衡量降维后矩阵的维度$r$的效果，$\frac{L_{r=+\infty}^* - L_r^*}{L_{r=+\infty}^* - L_{r=0}^*}$的值变化范围为$[0,1]$。

### 3.4.5 矩阵降维

对于原始矩阵来说，经过矩阵降维后会直接得到低维近似矩阵$\Theta$，由于$\Theta$的维度不超过$r$。因此，对于$\Theta$进行奇异值分解操作(SVD)可得$\Theta = U\Sigma V^T$，其中，$\Sigma$的维度也不超过$r$。对于进行奇异值分解操作(SVD)后的$\Theta$中的$\Sigma V^T$的前$r$列即为原始矩阵降维后的矩阵。

### 3.4.6 非监督聚类

通过矩阵降维后，我们可以得到关于原始矩阵的低维矩阵，对于地位矩阵来说，它与原始矩阵具有较强的相关性，因此为了利用分析特征以及降维后的矩阵探索潜在的癌症亚型，我们需要采用一种无监督的聚类方法进行聚类，在本文中，采用 K-Means 聚类以及一致化聚类方法，由于，对于 K-means 等聚类方法来说，不同的聚类个数 K 的对结果的效果不一样，因此本文采用轮廓系数来衡量不同聚类个数 k 带来的效果，其中，轮廓系数越大越好。

# 第四章 LRACluster 在多元组学数据集上的应用

## 4.1 数据集介绍

在对于 LRACluster 在多元组学数据集上的应用时，本文所采用的数据集均从 TCGA 下载，TCGA 为肿瘤基因图谱，TCGA 主要的利用以大规模测序为主的基因组分析技术，通过与工业界等广泛的合作，理解癌症的分子机制，提高人们对于癌症发病的分子基础的认识，从而提高我们对于癌症的认识以及收集关于不同癌症类型的分子特征数据。对于 TCGA 来说，起初是对 GBM 一种癌症进行分析，目前已经增加至 36 种癌症类型的研究。

本文采用的数据集为经过 UCSC[12]处理的来自 TCGA 的数据集，其中包括 11 种癌症类型（BRCA[13], COAD, GBM, HNSC, KIRC, LGG, LUAD, LUSC, PRAD, STAD），数据集包含体细胞突变，拷贝数变异，DNA 甲基化和基因表达等分子特征在不同的样本上的基因表达情况。

对于体细胞突变，拷贝数变异分子特征的基因表达来说，由于大量完整的数据集会降低聚类的稳定性，因此，对于体细胞突变，拷贝数变异分子特征的基因表达本文仅采用包含 500 多个样本的基因表达数据[14]。对于 DNA 甲基化的数据，将采用由 Illumina HumanMethylation450 BeadChip[15]处理的包含 8000 多个样本的基因表达数据。对于 mRNA(基因表达)来说，采用 20000 组 RNA-Seq 的基因表达数据。

综上所述，本文将 LRACluster 应用 TCGA 上 BRCA 的数据集，BRCA 包含两种癌症亚型。另外，将把 LRAClusteru 应用在 TCGA 上 Pan-cancer 数据集：该数据集包含 11 种不同的癌症类型。

## 4.2 对于 LRACluster 性能的探究

### 4.2.1 LRACluster 与 iCluster+的对比

LRACluster 是一种性能很高的用来降低矩阵维度并且进行聚类的方法，本文通过将 LRACluster 应用在包含三种癌症类型的数据集以及包含两种癌症亚型的 BRCA 癌症

数据集。为了对 LRACluster 的性能进行衡量，本文通过将 LRACluster 与 iCluster+分别应用于包含三种癌症类型的数据集上，比较这两种聚类方法的准确率以及轮廓系数。

首先，本文将 LRACluster 算法应用在具有三种不同癌症类型的多元组学数据集上，利用分子特征(mRNA、DNA 甲基化)在样本上的基因表达作为原始数据集，利用 K-Means 聚类，可得结果如下图所示：

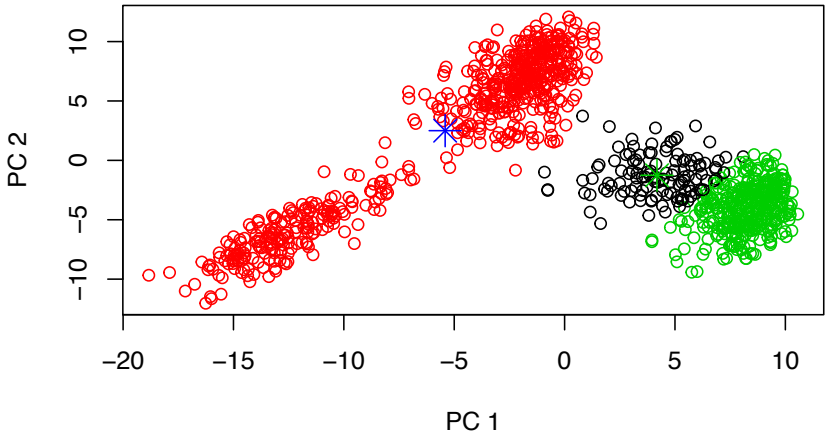


图 4.1：LRACluster 对癌症分类结果

由图可知：样本的聚类情况的集群的分布相对集中，因此可知 LRACluster 在对于癌症分类的上具有较好的效果。

由于对于所有的分子特征(体细胞突变，拷贝数变异，DNA 甲基化和基因表达)在不同的样本上的表达不同，并且数据集较大，因此本文将选取前 100 个分子特征在样本上的基因表达数据，将该数据集应用在 LRACluster 和 iCluster+上，通过设定降维后的维度矩阵为 2 到 10，可得结果如下：

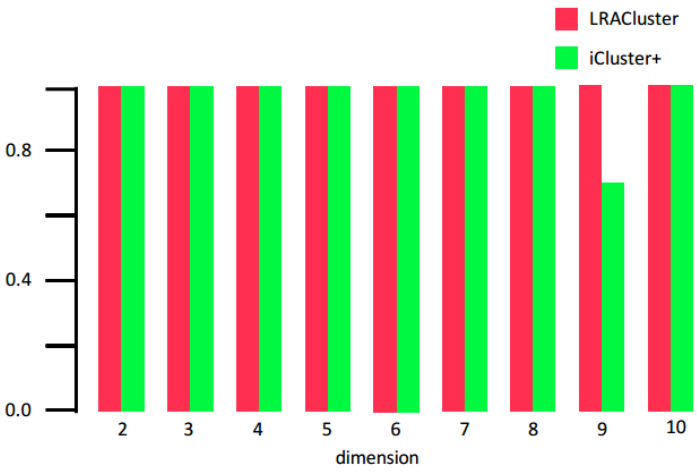


图 4.2：LRACluster 与 iCluster+的准确率对比图

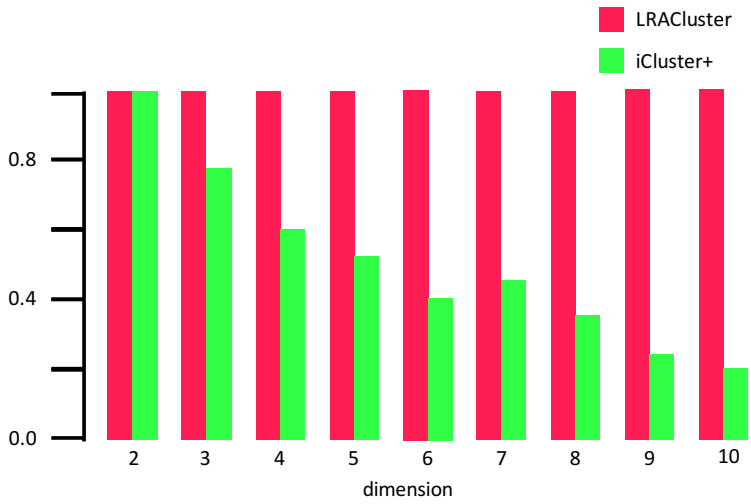


图 4.3：LRACluster 与 iCluster+轮廓系数对比图

由图 4.2，图 4.3 可知：通过对 LRACluster 与 iCluster+的聚类准确率比较来说，两种方法的准确率相近，图中唯一较大的差别为：当维度选择为 9 时，iCluster+的准确率相比 LRACluster 来说较低。对于轮廓系数来说，可以明显看出，LRACluster 比 iCluster+的轮廓系数要高，尤其是当矩阵的维度不断增加时，LRACluser 的轮廓系数趋于稳定，但是 iCluster+的轮廓系数却在不断下降。通过该结果可知，当数据维度以及模型越来越复杂的时候，iCluster+容易陷入优化问题中的局部最优解。相比来说，具有凸性的 LRACluster 的稳定性较好。综上所述，LRACluster 比 iCluster+的聚类和降维效果要好。

## 4.2.2LRACluster 收敛性的探究

对于 LRACluster 来说，收敛性是一个重要的因素。通过动态的改变“解释方差”以及惩罚参数$\mu$可得：LRACluster 可以在很短的迭代次数内收敛，如图 4.4 所示：

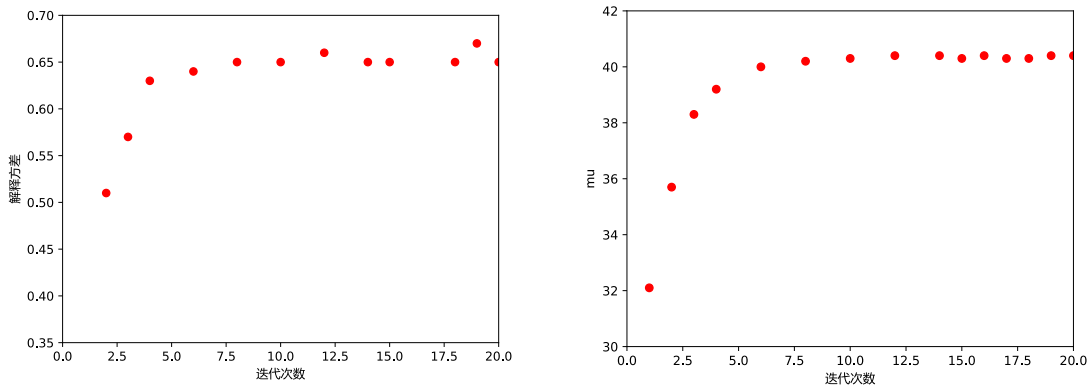


图 4.4 解释方差-迭代次数变化曲线

## 4.3 LRACluster 在 BRCA（乳腺癌）的应用

### 4.3.1 LRACluster 的参数的确定

对于 LRACluster 来说，存在两个重要的参数，降维后的低维矩阵维度 $r$，以及利用低维矩阵进行无监督聚类时的聚类个数选择。为了确定这两个参数，本文采用 BRCA 的数据集，该数据集包含分子特征（体细胞突变，拷贝数变异，DNA 甲基化和基因表达）在样本上的基因表达，BRCA 含有两个潜在的癌症亚型。其中对于矩阵维度 $r$ 本文采用“解释方差”进行衡量，对于聚类个数 $r$ 采用轮廓系数进行衡量。通过绘制矩阵维度在 2 到 10 上解释方差的变化，可得结果如下图所示：

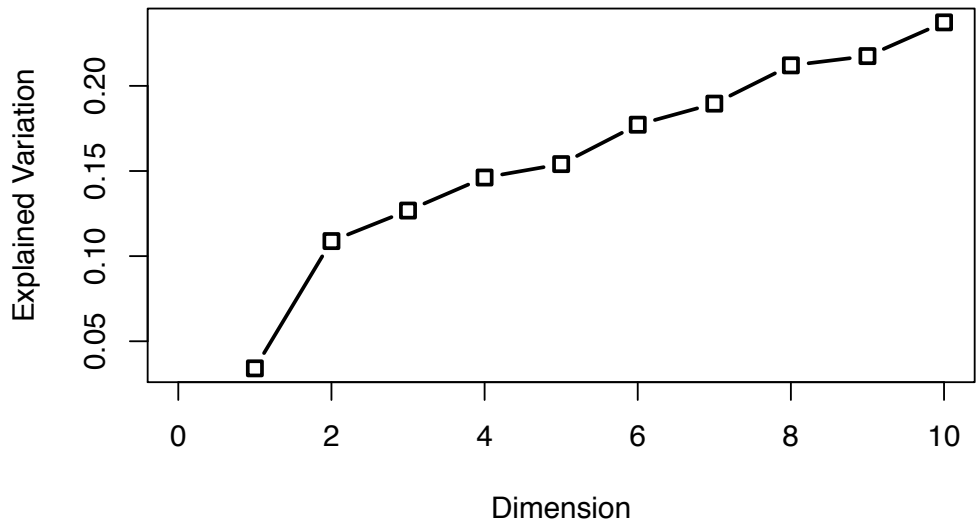


图 4.5 解释方差与矩阵维度的变化关系

由图 4.5 可知：当降维后的矩阵维度 $r=2$ 时，该曲线出现了一个转折点，在转折点之后，虽然解释方差依旧在增长，但是随着矩阵维度的不断增大，解释方差的增长越来越慢，因此对于 LRACluster 在 BRCA 上的 $r$ 应该选择 2。

当降维后的矩阵维度 $r=2$ 时，对于如何选择聚类个数，只需要根据一个最直接的原则：选取轮廓系数最高的聚类个数，那么该非监督聚类的聚类效果最好。因此，本文通过绘制轮廓系数与聚类个数 $c$ 的关系可得结果，如图 4.6 所示：

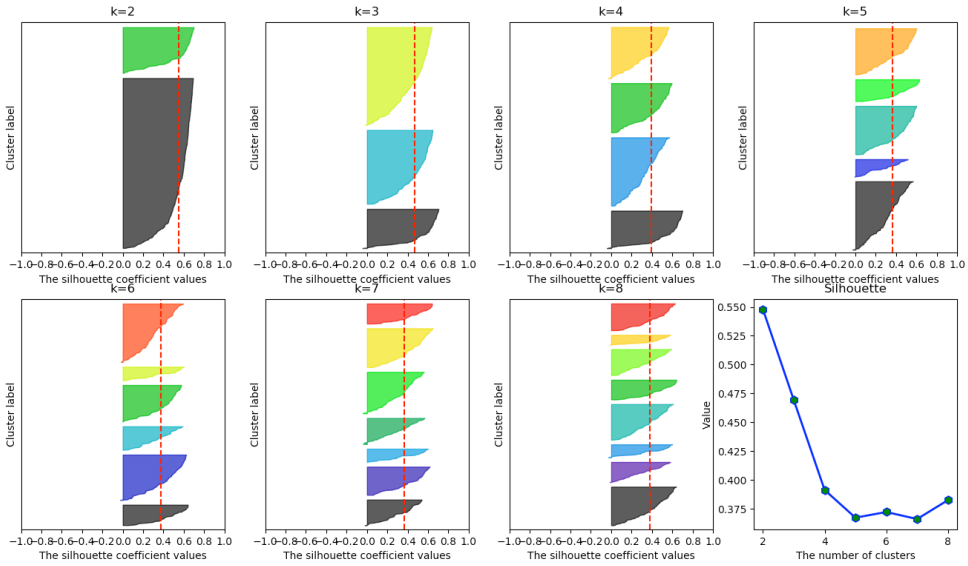


图 4.6 轮廓系数与聚类个数 $c$ 的关系图

如图可知：当聚类个数 $c=2$ 时，聚类的轮廓系数最高，因此选择 $c=2$ 。

### 4.3.2 K-Means 聚类的应用

根据上述可知：当降维后的矩阵维度 $r=2$，聚类个数 $c=2$ 时，本文通过将降维后的矩阵作为输入，对该矩阵进行 K-Means 聚类，可得结果如图 4.7 所示：

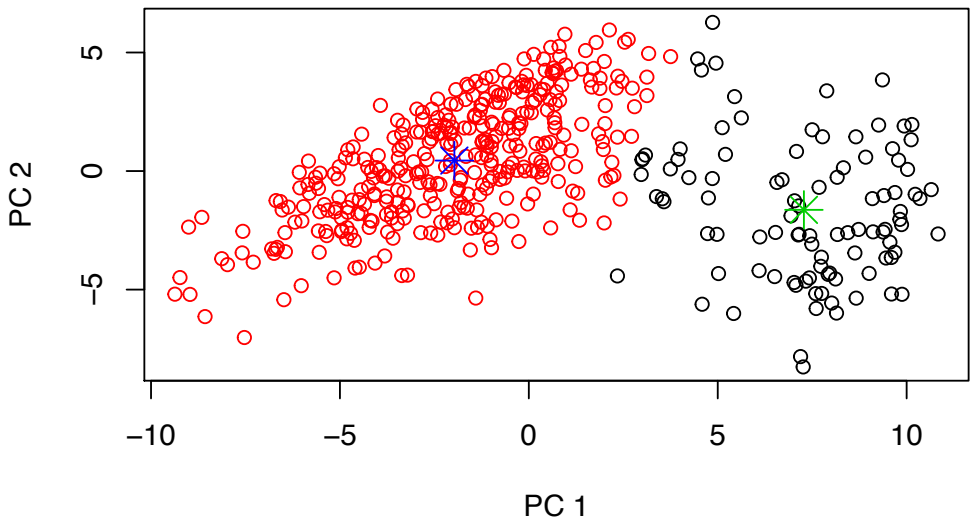


图 4.7 BRCA 的癌症亚类结果

由图 4.7 可知：通过 LRACluster 进行聚类可得，BRCA 存在两个癌症亚型。该结果与 TCGA 数据集的 BRCA 癌症亚型个数相符，因此，LRACluster 是一种有效的聚类方法对于探究高维度的癌症数据集。

### 4.3.3 ConsensusClustering 的应用

根通过利用 LRACluster 降维后，可以得到降维后的低维矩阵，本文将该数据矩阵作为输入，利用 ConsensusClustering 算法探究不同的聚类个数对于轮廓系数的影响，结果如下图所示：

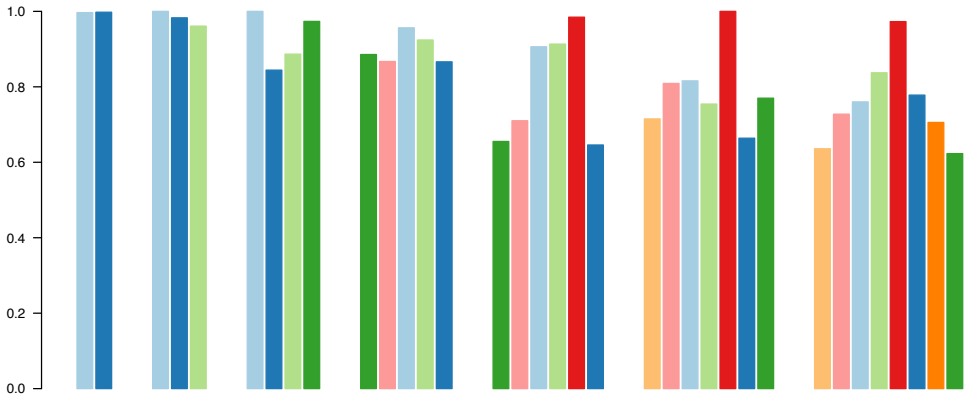


图 4.8 ConsensusClustering 的结果图

由图 4.8 可知：当聚类个数为 2 时，轮廓系数的值为 1，而其他的聚类个数的值均小于 1。该结果与上述 K-means 聚类的实验结果相同，故 LRACluster 是一种有效的聚类方法。

## 4.4 对于 LRACluster 在 Pan-Cancer（泛癌类型）的应用

通过上述实验可知：LRACluster 在对特定的单个的癌症类型的癌症亚型探究具有较好的效果。接下来，本文将使用 LRACluster 探究对于 Pan-Cancer（泛癌类型）的癌症类型的分类效果，绘制“解释方差”与降维后的矩阵维度的关系以及轮廓系数与聚类个数的关系可得结果如下所示：

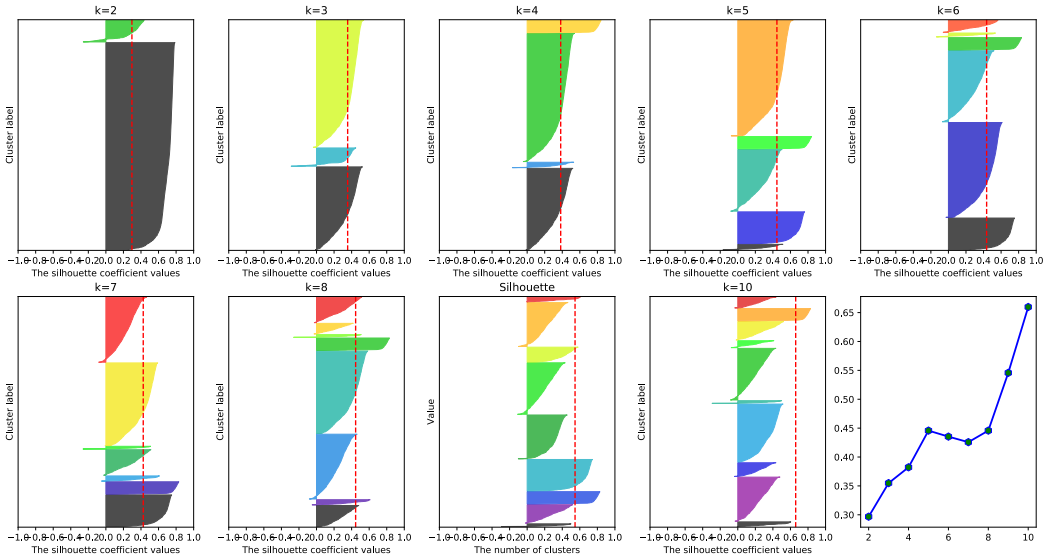


图 4.9 Pan-Cancer（泛癌类型）聚类的轮廓系数变化图

图 4.9：利用降维后维度 $r$=10 的矩阵进行 k-means 聚类，通过计算聚类的轮廓系数，可知当聚类个数 $k$=10 时，轮廓系数最高，故聚类效果最好。

当降维后的矩阵维度选择 10 ，并且聚类个数选择 10 时，通过将 LRACluster 应用在 Pan-cancer 数据集，包含 11 种癌症类型的高维矩阵被降为低秩矩阵，之后利用 k-means 聚类，11 类癌症被分为 10 类矩阵。结果如下所示：

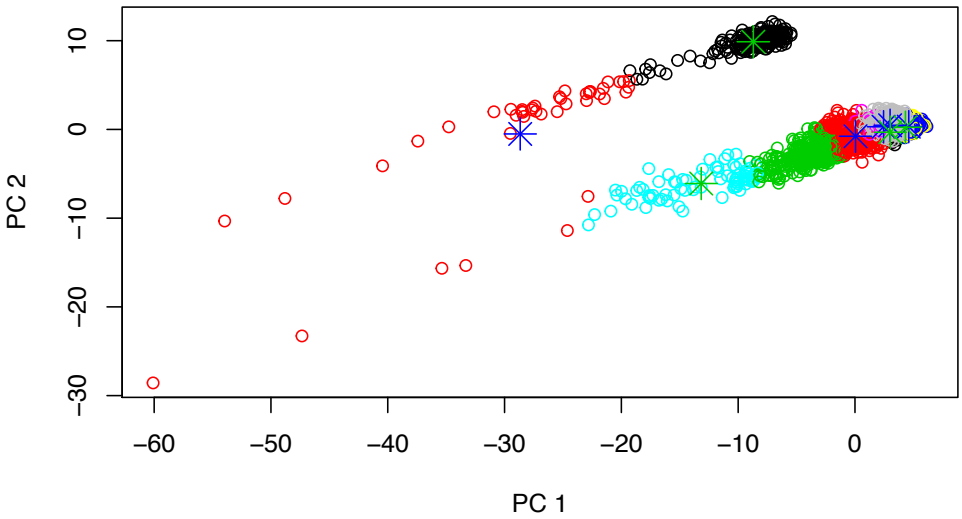


图 4.10 ConsensusClustering 的结果图

表 4.1 非监督聚类的结果

| | BRCA | COAD | GBM | HNSC | KJRC | LGG | LUAD | LUSC | PRAD | STAD | THCA | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 1 | 0 | 0 | 286 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 293 |
| C2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 411 | 412 |
| C3 | 0 | 0 | 41 | 0 | 0 | 451 | 0 | 0 | 0 | 0 | 0 | 492 |
| C4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 231 | 0 | 231 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 293 | 0 | 0 | 293 |
| C6 | 0 | 190 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 194 |
| C7 | 3 | 17 | 0 | 0 | 1 | 0 | 406 | 7 | 0 | 0 | 3 | 437 |
| C8 | 0 | 0 | 0 | 0 | 240 | 0 | 0 | 0 | 00 | 0 | 0 | 240 |
| C9 | 448 | 0 | 1 | 2 | 1 | 0 | 4 | 1 | 0 | 0 | 0 | 457 |
| C10 | 8 | 1 | 0 | 195 | 0 | 0 | 6 | 60 | 0 | 0 | 0 | 270 |
| Total | 460 | 208 | 42 | 484 | 242 | 452 | 418 | 74 | 294 | 231 | 414 | 3319 |

由表 4.1、图 4.10 可知：原本属于同一癌症类型的样本基本都被分为同一类别，这一结果与目前的研究现状相吻合。但是，存在两种脑癌（LGG 和 GBM）却被分成了同一类别。另外，只有 HNSC 被分成两个主要类别（类别 C1 和 C10），类别 C10 中的样本有 40.3%来自 HNSC，其余的样本来自 LUSC（占原始 LUSC 样本的 81.1%），由此可知：不同组织的癌症可能存在着相同的分子特征。

将 LRACluster 分类结果与原始癌症类型进行比较。

通过利用 LRACluster，可将 11 种癌症类型分为 10 种癌症类型，本文通过将 11 种癌症类型的样本在 mRNA、拷贝数变异、DNA 甲基化、体细胞突变等分子特征上的基因表达与利用组织来源进行分类的癌症类型联系，探索这两者之间的关系，如图 4.11 所示：

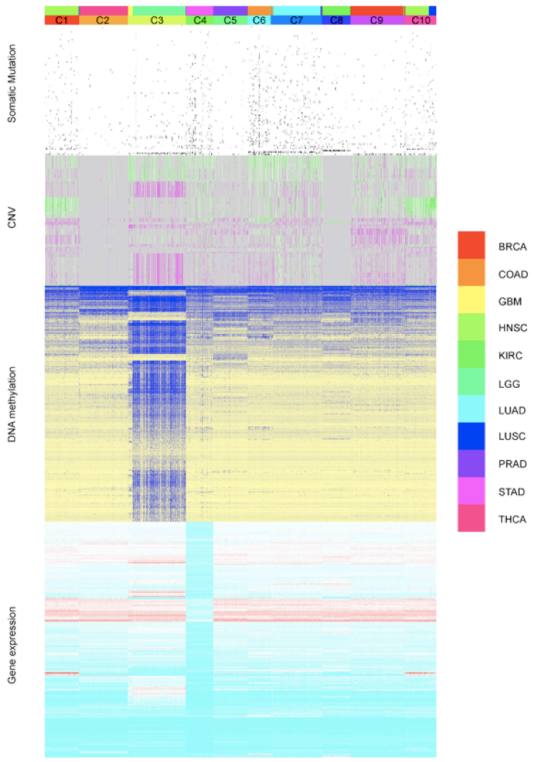


图 4.11 来自 TCGA Pan-cancer 数据集的已鉴定类别相关联的分子特征的热图

顶部颜色条表示已知的癌症类型和由 LRAcluster 分类的类别。

由图 4.11 可知：该结果与利用 LRAcluster 进行分类的结果类似：大多数癌症类型根据其组织来源进行聚类。但是，两种类型的癌症，头颈部细胞癌和肺癌却被份分成了同一类别。

## 4.5 将 LRACluster 分别应用于多种癌症类型

### 4.5.1 概述

本文将 LRAcluster 分别应用于 11 种癌症类型在分子特征上的的数据集。其中：降维矩阵的维度可以根据“解释方差”的变化确定，聚类的个数 $k$ 可以根据“轮廓系数”的变化确定。由表 4.2 可知：不同的癌症的组学数据集具有不一样的癌症亚型个数。并且，对于不同的癌症来说，LRACluster 的结果不同，聚类效果不同。其中， BRCA，LGG，PRAD 和 THCA 等组学数据具有较高的轮廓系数。在 BRCA，LGG，PRAD 和 THCA 中可以探索到一些癌症亚型。对于剩余的 7 种癌症类型，LRAcluster 根据目前的组学数据没有找到明显的癌症亚型，或者是根据 LRACluster 进行分类的效果并不好。下面本文将以 COAD 为例进行实验。

表 4.2 LRACluster 在单个癌症类型的结果

| Cancer | Dimension | Cluster | Silhouette values |
|---|---|---|---|
| BRCA | 2 | 2 | 0.55 |
| COAD | 4 | 4 | 0.40 |
| GBM | 8 | 3 | 0.35 |
| HNSC | 7 | 2 | 0.26 |
| KIRC | 6 | 2 | 0.36 |
| LGG | 2 | 3 | 0.44 |
| LUAD | 5 | 2 | 0.34 |
| LUSC | 5 | 4 | 0.32 |
| PRAD | 2 | 4 | 0.41 |
| STAD | 4 | 3 | 0.37 |
| THCA | 2 | 2 | 0.61 |

### 4.5.2 LRACluster 在 COAD（肺癌）上的应用

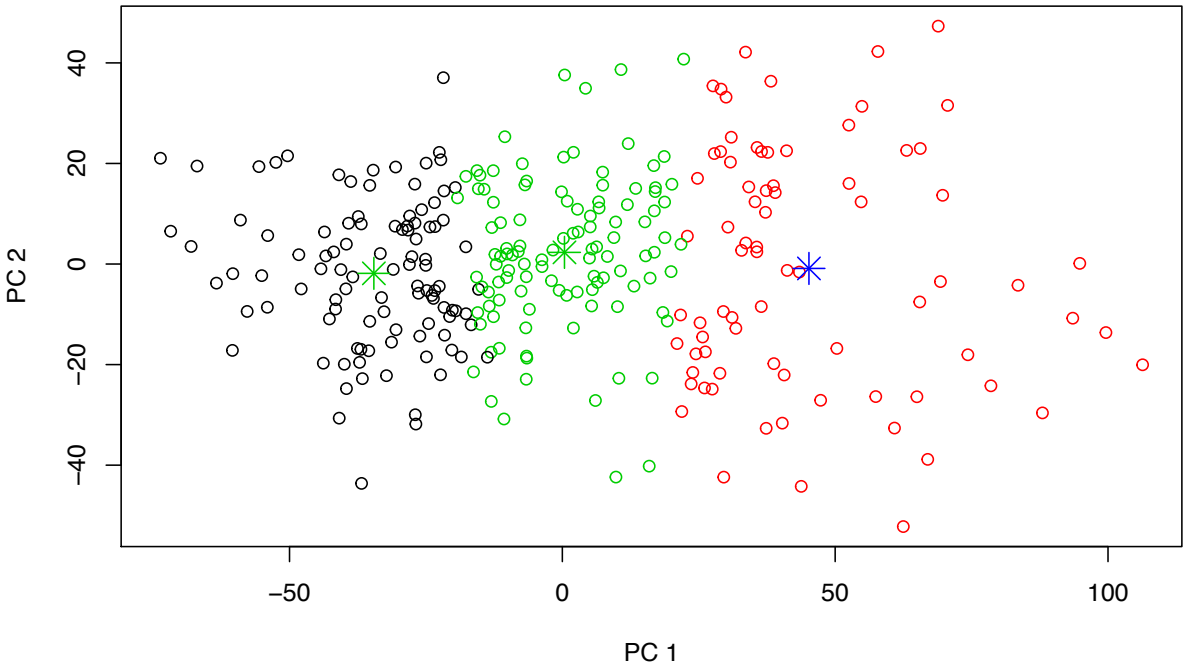


图 4.12　LRACluster 在 COAD 上的应用

由图 4.12 可知：在对于将 LRACluster 应用于 COAD（肺癌）的癌症亚型的探究上时，该算法的分类效果不能达到如期效果。

# 第五章 总结与展望

## 5.1 对于 LRACluster 的讨论

在 LRAcluster 的概率模型中，将正态类型的数据处理成高斯随机变量，并且其方差为 1。对于不同的特征来说，应该具有不同的方差，但是本文将所有变量特征的方差假设成 1，这与主成分分析（PCA）的处理方式很相似。在 LRAcluster 中，似然函数的负对数求和即为主成分分析（PCA）的损失函数。因此，如果输入数据中只有正态类型的数据，那么 LRAcluster 与主成分分析（PCA）算法相似。但是，对于两个算法来说，唯一的不同便是两个算法处理范围，这是因为：LRAcluster 采用 $L_1$ 范数，主成分分析（PCA）采用 $L_0$ 范数。

## 5.2 总结

首先，LRAcluster 的概率模型中的参数数据矩阵与原始矩阵维度相同。其次，对于 LRAcluster 来说，低维矩阵近似的关键步骤是得到与初始矩阵相关性较强的低维矩阵。并且由于 LRA 的似然函数的凸函数性质，可以利用梯度下降法求解优化问题，达到全局最小(最优)的结果。最后，通过对 BRCA、Pan-Cancer 等数据集的结果显示 LRAcluster 是一种运行速度快，分类准确度高，适用于大规模癌症多组学分析的聚类方法。

## 5.3 展望

对于 LRAcluster 来说，该聚类方法没有考虑分子特征之间的关系以及利用稀疏的矩阵来减少矩阵的维度。因此，为了改进 LRAcluster 的分类效果，可以探索分子特征之间的关系，以及不同的分子特征在进行 LRAcluster 时，找到对于 LRAcluster 有代表性的分子特征从而进行分类。

除此之外，LRAcluster 更适合于具有较大协方差的跨组学的分子特征数据集，例如，具有显著相关的 DNA 甲基化与拷贝数变异的数据集。通过对跨组学的的数据集的研究，

可以探究到癌症的驱动因素，通过将这些跨组学的分析特征数据作为输入可以更好的对癌症亚型的分类进行探索。

最后，联合非负矩阵分解(jNMF)[16]是另一种基于癌症多元组学数据集来对癌症亚型进行探索的方法。对于 jNMF 来说，其也将面临损失函数是非凸函数的困难，但是 jNMF 可以分子特征间的关系，以及对于癌症分类来说具有重大影响的分子特征。

# 致 谢

大学四年的生活已然落幕，坐在教学楼撰写毕设论文的最后一章，从对生物信息学的一无所知，到如今的了解一二。从拿到题目时的许多想法到如今的尘埃落定。这几个月，这几年经历了许多，因此想对诸多人表达感激之情！

首先，感谢自己的指导老师——徐悦甡老师。初见徐老师是培训数模的暑期训练，那时便和徐老师结下师生之情，之后在这两年的数模生涯、科研生活中也同徐老师有了许多的交流、合作，毕设过程承蒙徐老师的悉心指导，对论文的撰写、对算法的指导、对论文的改进提出众多意见，希望以后可以与徐老师有更多的学术交流！

其次，感谢实验室的段然师兄，师兄为人谦和，由于自己对于数据集成、大规模数据聚类算法、生物信息学了解甚少。因此时常需要同师兄交流，每次交流师兄都十分耐心，从代码、算法基础、数据集、背景知识等方面给予了我十分巨大的帮助，同段然师兄的交流讨论是一段非常难忘的经历，谢谢师兄！

最后，感激自己的父母，感激父母无条件的支持自己攻读这大学四年的生涯，人生是一场修行，大学仅仅只是一段短跑，希望在将来的读硕士、读博士、工作的职业生涯路上，我得到你们更多的支持，谢谢！

回首看来，大学四年已然过去。有许多感慨，这四年帮助自己的有过许多人，无法一一感谢，只能以此文表达谢意，愿各位老师、师兄、同学前程似锦！

# 参考文献

[1]Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012[J]. International Journal of Cancer, 2015, 136(5):E359-E386.

[2]Li L, Li H. Dimension reduction methods for microarrays with application to censored survival data[J]. Bioinformatics, 2004, 20(18):3406.

[3]Shen R, Olshen A B, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis[J]. Bioinformatics, 2009, 25(22):2906.

[4]Mo Q, Wang S, Seshan V E, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(11):4245.

[5]Lock, E. F. and Dunson, D. B. Bayesian consensus clustering[J]. Bioinformatics, 2013.

[6]Candes EJ, Recht B. Exact Matrix Completion via Convex Optimization. Found Comput Math. 2009;9:717–72.

[7]Monti S, Tamayo P, Mesirov J, et al. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data[J]. Machine Learning, 2003, 52(1-2):91-118.

[8]Wang D, Gu J. Integrative clustering methods of multi-omics data for molecule-based cancer classifications[J]. Quantitative Biology, 2016, 4(1).58-67.

[9]Wu D, Wang D, Zhang M Q, et al. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification[J]. Bmc Genomics, 2015, 16(1):1022.

[10]Hsieh CJ, Olsen PA. Nuclear Norm Minimization via Active Subspace Selection. Proc 31st Int Conf Mach Learn. 2014.

[11]Cai J F, Cand, S, E J, et al. A Singular Value Thresholding Algorithm for Matrix Completion[M]. Society for Industrial and Applied Mathematics, 2010.

[12]Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, et al. The UCSC Cancer Genomics Browser: update 2015. Nucleic Acids Res. 2015; 43(Database

issue):D812–817.

[13]Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications.[J]. World Journal of Clinical Oncology, 2014, 5(3):412-24.

[14]Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43(Database issue):D805–811

[15]Triche T, Jr. IlluminaHumanMethylation450k.db: Illumina Human Methylation 450k annotation data.

[16]Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinforma Oxf Engl. 2011;27:i401–409.