密级 _____

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

基于鲁棒非负矩阵分解的聚类方法研究

作者姓名：杜亮

指导教师：沈一栋 研究员

中国科学院软件研究所

学位类别：博士

学科专业：计算机软件与理论

培养单位：中国科学院软件研究所

2013年4月

Clustering with Robust Nonnegative Matrix Factorization

By

Liang Du

Supervisor:

Professor Yi-Dong Shen

A Dissertation Submitted to
University of Chinese Academy of Sciences
In partial fulfillment of the requirement for the degree of
Doctor of Philosophy
in Computer Software and Theory

Institute of Software
Chinese Academy of Sciences
April, 2013

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明。

签名：_____　　日期：_____

关于论文使用授权的说明

本人完全了解中国科学院软件研究所有关保留、使用学位论文的规定，即：中国科学院软件研究所有权保留送交论文的复印件，允许论文被查阅和借阅；中国科学院软件研究所可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：_____　导师签名：_____　日期：_____

谨以此文献给我最亲爱的妈妈

This dissertation is dedicated to my mother

for her constant love, support and encouragement.

摘　要

聚类分析是数据挖据领域重要的基础性研究问题之一，而非负矩阵分解是一种重要的聚类分析方法。实际数据往往存在质量问题，如误差、噪声、异常点等，导致非负矩阵分解难以全面准确的反映数据的真实特性，给聚类分析带来了困难。本文针对数据的质量问题研究如何提高非负矩阵分解的鲁棒性，以及如何提高基于非负矩阵分解的聚类效果，取得的主要研究成果如下：

- 提出了基于半二次最小化的鲁棒非负矩阵分解方法。针对标准非负矩阵分解对非高斯噪声敏感的问题，本文采用鲁棒的半二次函数度量矩阵的质量，从而得到鲁棒的非负矩阵分解模型；提出适用于不同半二次函数的通用非负矩阵分解算法。带噪声数据的实验表明，本文提出的鲁棒非负矩阵分解的聚类方法可以有效聚类效果。

- 提出了基于稀疏性噪声假设的非负矩阵分解联合聚类算法。本算法从样本与特征关系中抽取一个稀疏错误矩阵，用于刻画非高斯噪声并恢复真实数据；采用绝对值损失函数降低特征与特征以及样本与样本之间噪声关系带来的图正则误差。真实数据集的联合聚类实验表明，本文提出的鲁棒联合聚类算法在性能上优于当前的主流方法。

- 提出了基于半二次最小化的鲁棒联合聚类方法。为了改善稀疏性噪声假设在实际应用中的局限性，本文引入半二次损失函数来分别度量特征与样本关系矩阵的重构误差以及样本与样本、特征与特征关系的图正则误差，从而得到应用面更广的鲁棒联合聚类方法。

- 提出了区间非负矩阵分解方法。实际应用中观测到的数据不可避免的存在一定的误差，给矩阵分解带来了挑战。本文通过均匀分布的区间近似表达数据可能的取值，提出基于区间上下界矩阵的联合非负矩阵分解方法。结合人脸聚类分析和协同过滤两个应用，给出了构造均匀分布对应的区间上下界矩阵方法。实验结果表明区间矩阵分解方法明显优于对应的单值矩阵分解方法。

- 提出了基于加权图正则非负矩阵分解的聚类集成算法。聚类集成的输入数据有基于簇特征和多重关联关系的两种表示形式。当前聚类集成方法仅仅使用其中一种表示形式进行聚类。由于这两种形式在集成任务中存在相关性，本文同时利用两种表示进行聚类集成，一方面利用簇特征表示进行非负矩阵分解，另一方面利用多重关联关系表示进行加权合并辅助聚类。

以上主要研究成果发表在IEEE International Conference on Data Mining (ICDM 2012)和The International Joint Conference on Artificial Intelligence (IJCAI 2013)等国际会议。

关键词： 鲁棒非负矩阵分解，鲁棒联合聚类，区间矩阵分解，加权图正则非负矩阵分解

Abstract

Cluster analysis is a key unsupervised learning technique in data mining research, and Nonnegative Matrix Factorization is one of the important methods in clustering. Due to the presence of noisy data, gross errors, outliers and approximations in real world applications, the ordinary nonnegative matrix factorization methods may fail to accurately capture the intrinsic structure of the data, which poses great challenges to the clustering task. Facing the reality of data sets with quality issues, this thesis aims to improve the robustness of nonnegative matrix factorization and the corresponding quality of clustering.

- We propose a new robust nonnegative matrix factorization method via half quadratic minimization. Unlike the ordinary NMF, we use the half-quadratic loss functions to quantify the quality of matrix factorization. These functions are less increasing than the often used squared error function or have heavier tails than the Gaussian distribution, thus a more faithful factorization result could be expected. To solve the optimization problem for different loss functions, we develop two generic minimization algorithms based on both the multiplicative and the additive half-quadratic reformulations. Clustering results on noisy data sets well demonstrate the effectiveness of the proposed robust NMF method.

- We propose a novel NMF method for co-clustering task by assuming the data set is corrupted by gross but sparse errors. To handle the large and impulsive non-Gaussian noises in the sample-feature relationship matrix, we purposely estimate a sparse error matrix to capture the corrupted parts and recover the cleaned data from these errors; to alleviate the impacts of undesired inter-sample and inter-feature relationship, we use the absolute function to measure the graph regularization error. Experimental results on several real world data sets show that the proposed method usually performs better.

- We propose the robust co-clustering method based on graph regularized nonnegative matrix tri-factorization in the half-quadratic minimization framework. To remedy the limitation of undesired sparse noise assumption and reap the benefits of different half-quadratic loss functions, we further apply the half-quadratic loss functions to measure both the reconstruction error and the graph regularization errors and lead to general robust graph regularized nonnegative matrix tri-factorization.

- We propose the interval-valued matrix factorization methods. The single-value of the data in real world applications would inevitably introduce a certain errors. To capture the uncertainty of the single-valued data, we propose to use a uniform distribution to approximate its possible values; and a joint matrix factorization framework to approximate the intervals of the uniform distribution which leads to interval-valued matrix factorization methods. We also show how to empirically construct a reasonable interval of the uniform distribution in a limited contextual. Experimental results on face analysis and collaborative filtering show that the interval-valued matrix factorization methods perform better than its counterparts.

- We propose a new clustering ensemble algorithm based on weighted graph regularized nonnegative matrix factorization. The input of clustering ensemble can be represented either by clusters or multiple co-associations. Based on the coherence between these two representations, we propose a weighted graph regularized nonnegative matrix factorization method in the joint learning framework. This algorithm uses NMF to estimate an accurate clustering on the cluster-based representation; estimate a consensus co-association relationship through multiple co-association matrices linear combination.

Keywords: robust nonnegative matrix factorization, robust co-clustering, interval-valued matrix factorization, Weighted graph regularized nonnegative matrix factorization

目　　录

第一章　绪论

1.1　引言

随着信息技术的飞速发展，人类收集数据和存储数据的能力得到了极大的提高，数据规模的增长非常迅速。如Google公司每月处理的数据量超过400PB，百度每天大约要处理几十个PB数据，淘宝网每天数千万交易也产生几十TB的数据[2]。很有必要对这些数据进行分析，发现蕴含在数据中的有用信息。正是在这样的背景和趋势下，数据挖掘和机器学习近年来受到了工业界和学术界的广泛关注。

数据挖掘是指从大量数据中找出有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程[44]。机器学习主要考虑如何使机器能够从经验中进行学习，从而不断提高和完善计算机系统自身的性能[74]。

聚类分析是数据挖掘和机器学习领域重要的基础性研究问题，它将样本点划分到不同的簇中，使得特征相似的样本点在相同的簇中。非负矩阵分解[60]作为一种有效的聚类分析方法被广泛用于网络文档自动聚类[9, 106]，社区发现[98, 116]，图像表示和理解[14, 20, 69]，肿瘤亚型发现和归类[12]等。当前非负矩阵分解方法隐式或显式的对数据分布作出一定的假设，如标准非负矩阵分解假设观测数据来自高斯分布，基于Kullback－Leibler失真度的非负矩阵分解假设观测数据来自泊松分布。实际数据中可能出现噪声、异常点、不相关特征、误差等，如图像的遮挡和光照干扰，网页中与主题内容无关的文字，基因表达数据中与类别信息无关的基因等。这些数据质量问题直接影响数据的分布，使得非负矩阵分解模型的假设不完全成立，无法准确的反映实际数据的特征，从而降低了聚类分析的效果。本文针对数据的质量问题研究如何提高非负矩阵分解的鲁棒性，以及如何提高基于非负矩阵分解的聚类分析结果。

1.2　聚类和联合聚类

1.2.1　聚类分析

聚类分析是数据挖掘和机器学习领域中一个基础性问题，其目标是将样本

点划分到不同的簇中使得同一个簇中样本点的距离接近而不同簇中样本点距离较远。聚类分析技术大致可分为数据表示，聚类算法设计和选择，聚类结果评估和验证等环节[105]。聚类方法的分类标准有很多，其中包括：

- 根据聚类算法的输入数据类型分类。数据可以由特征向量来表示，这些特征可以是数值型，类别型或者两者都有；数据还可以由他们之间的关系来表示，这些关系可以通过特征向量之间的相似度或者距离来计算也可以从实际应用中获取；此外数据还可以同时由特征向量和关系来表示。因此聚类算法可以分为数值型算法（Kmeans，NMF[106]等），离散型（如AT-DC[18]）、关系型（如NCut[92]，affinity propagation[34]等）和混合型算法（如GNMF[14]，DNMTF[101]等）。

- 根据输出结果聚类算法分为层次聚类和划分式聚类。根据数据间的邻近关系，层次聚类将数据组织为树状图（dendrogram），其中根节点表示整个数据集，叶子节点表示样本，中间节点表示簇。层次聚类算法可以分为凝聚式（agglomerative）方法和分裂式（divisive）方法，代表性方法包括single-linkage，complete-linkage，average-linkage等。划分式算法是将数据划分到不同的簇中，根据聚类后样本是否仅属于一个簇，又可分为硬划分（如Kmeans，Bregman hard clustering[6]等）和模糊聚类（如Fuzzy Kmeans，Bregman soft clustering[6]等）。

- 根据簇的描述形式聚类算法可分为基于原型的方法（也叫簇代表元）和基于模型的方法。基于原型的方法通常用一个代表元表示簇，用数据到代表元的距离或者相似度来反映聚类分析的好坏，代表性算法有Kmeans，K-medoids等；基于模型的方法可以用一个分布表示簇，用数据由这些分布生成的概率来刻画聚类分析的好坏，代表性的方法有：高斯混合模型（GMM），Bregman聚类等。

1.2.2 联合聚类

一般来讲，可以用一个2D矩阵表示数据，其中一维（如矩阵的行）表示特征，另一维（如矩阵的列）表示样本。传统的聚类算法（如Kmeans，谱聚类和非负矩阵分解等）仅考虑单个维度上（如按列对样本进行聚类）的聚类而忽略另外一个维度存在的簇结构以及两个维度上簇结构之间的关联关系。考虑到特

征簇和样本簇之间的对偶和相互依赖关系，近年来研究人员提出了联合聚类算法，同时将样本和特征划分到不同的簇。常见的联合聚类算法可分为:

- 基于矩阵分解的方法[29]。通过将数据矩阵分解为三个因子矩阵的乘积分别获得了样本的划分矩阵和特征的划分矩阵以及样本簇与特征簇的关联关系。然而数据矩阵仅表示了样本和特征的关系，而忽略了数据间的关系。由于样本和特征分别来自某些流形，我们可以分别构造样本端和特征端的关系图来刻画样本间的关联关系以及特征间的关系，并且采用对偶图正则的非负矩阵三因子分解方法来进行联合聚类[36, 37, 90]，此外该方法还被应用于其他相关任务（如协同过滤[40]，迁移学习[72, 119]等）。为了减少计算代价，文献[82]提出基于列和行分解的联合聚类算法，文献[102]提出基于硬对偶划分（即每个样本和特征只属于一个簇）的图正则非负矩阵分解算法。

- 基于信息论的方法。通过分别将样本和特征看出两种随机变量，其中他们的联合概率分布可以通过数据矩阵来估计。对数据或样本的聚类可以看成对随机变量的压缩过程。Dhillon等在文献[25]中提出量化联合聚类互信息损失函数的方法，文献[5]进一步采用Bregman失真度来度量联合聚类损失并提出泛化的求解方法。

- 基于图划分的方法。通过构造关系图，其中节点表示样本而边的权重表示样本间的相似度，聚类问题可以描述为图分割的问题，如谱聚类[92]。文献[24, 54, 112]提出采用二部图来表示联合聚类的问题并采用谱方法求解。Zhang等人在文献[113]中提出采用局部判别式学习的方式构造数据间和样本间的关系图，利用归一化的数据矩阵表示样本和特征间的关系，并利用联合谱嵌入方法求解联合聚类问题。

1.3　非负矩阵分解

矩阵分解通过将高维数据近似为多个低维因子的乘积使得数据中的某种潜在结构能够更好的被描述出来。常见的方法包括：主成份分析、奇异值分解、矢量量化、非负矩阵分解等。实际应用中非负数据是一种常见的数据类型，如文本数据，图像数据，关系数据、基因表达数据等。非负矩阵分解将原始非负数据近似为两个非负因子的乘积，由于使用了非负约束使得矩阵分解

获得的结果往往对应于基于局部特征的数据表示，有助于对数据本质的刻画。非负矩阵分解技术是数据挖掘和机器学习领域中一个重要方法，根据Google Scholar统计，Lee和Seung发表在《Nature》上的非负矩阵分解论文引用次数高达3780次。

非负矩阵分解最早可追溯到Paatero和Tapper在文献[81]中基于因子分析提出了正矩阵分解（Positive matrix factorization，简称PMF）的概念，并将其应用到环境数据分析，实验结果表明PMF算法优于主成份分析和因子分析两种方法。但是，由于应用领域较偏并且算法实现复杂，PMF并未受到广泛的关注。Lee和Seung于1999年在《Nature》上正式提出了非负矩阵分解的基本概念[60]，并将其应用于获取人脸基于局部特征的表示和文本语义特征抽取。非负矩阵分解的心理学和生理学基础是：对整体的感知由基于对组成整体的部分（局部）的感知构成，直观上讲也就是：整体是由局部构成（如人脸可由眼睛、鼻子、嘴巴等部分构成，文本由语义主题组成）。非负矩阵分解通过非负约束纯加性（only additive, not subtractive）的感知过程刻画出数据的组成部分和数据如何由局部构成（combinations）的本质。针对非负矩阵分解的研究主要集中于下面几个方面：

1.3.1 非负矩阵分解的目标函数

为了量化原始数据和重构结果直接的近似程度，需要引入一个目标函数作为非负矩阵分解的优化目标。目标函数可以是距离函数、相似度度量或者散度，目标函数可以是单一的代价函数也可以是基于相同最优结果的一组代价函数。常用的目标函数包括：平方误差之和（即矩阵的F-范数平方）、Bregman失真度[26]、Minkowski距离（即ℓ_p-范数），超球面损失函数[43]、以及alpha-beta-gamma失真度[21–23, 55]等。不同的数据往往需要不同的损失函数，文献[88]采用推土机距离（Earth Mover’s distance metric，简称EMD）度量直方图数据上的损失；而文献[67]采用广义Kullback-Leibler失真度（基于泊松分布）来刻画计数型数据(如click counts)的近似误差；文献[43, 56]分别采用hypersurface和$L1$损失函数以提高标准非负矩阵分解模型的鲁棒性。

1.3.2 非负矩阵分解求解算法

非负矩阵分解将数据近似为两个非负因子的乘积，到目前为止，尚没有有效的算法同时优化这两个非负因子。当前求解非负矩阵分解的基本路线是通过

交互最小化和分块坐标下降法的思想将非负因子分解成不同的单元，并对每个单元的子问题分别求解。因此，不同求解算法的区别之处在于如何划分非负因子以及如何求解相关子问题。对非负因子的划分主要包括划分为两个因子矩阵，划分为多个非负因子向量以及单独划分非负因子中的每一个元素等三类。根据子问题优化算法可大致分为一阶优化和高阶（主要是二阶）优化算法。不同的优化算法主要区别在于如何选择梯度方向以及如何确定搜索步长。梯度方向选择包括一阶通用梯度、共轭梯度、投影梯度等以及二阶牛顿和拟牛顿法等。步长选择包括基于自适应调整步长的乘法更新公式和基于步长搜索的加法更新机制。代表性算法包括：Lee and Seung在文献[59]中利用辅助函数提出乘法更新公式（multiplicative update rules）来求解，而乘法更新公式事实上可以看成自动重新调整（adaptive rescaled）的梯度下降算法；Lin在文献[66]中提出基于投影梯度和Armijo步长准则的求解算法；Kim等人在文献[51–53]中分别提出了使用投影牛顿法、作用集（active set）方法以及改进的作用集方法Block-Pivot求解非负最小二乘（non-negative least squares，简称NNLS）子问题；Hsieh和Dhillon在文献[49]中提出主动选择单个元素的快速坐标下降算法。此外，这些算法也被扩展到用于求解其他非负矩阵分解模型，如Kuang等人利用牛顿法求解对称非负矩阵分解[58]，Zhang和Yeung利用坐标下降法和辅助函数法分别求解有界的非负矩阵三因子分解问题[116]。

1.3.3　受限的非负矩阵分解模型

标准的非负矩阵分解模型存在两个缺陷：一方面无法单靠非负约束获得唯一解，另一方面无法利用先验知识全面的刻画数据。通过先验知识进一步约束非负矩阵分解的工作主要分为以下四个方面：

- 稀疏非负矩阵分解。尽管非负约束本身会带来一定的稀疏性，通过进一步增加对非负因子稀疏性的约束一方面可以提高非负因子的唯一性，另一方面也可以增强非负矩阵分解获取局部特征的能力。Feng等人[32]利用基向量局部性约束和编码向量的稀疏性约束得到局部非负矩阵分解；Hoyer[47]和Liu[71]等人利用编码向量的ℓ_1范数约束分别提高以平方误差和KL失真度量为损失函数的非负矩阵分解编码的稀疏性；Hoyer[48]进一步通过ℓ_1范数和ℓ_2范数的关系分别显式地度量和提高编码向量和基向量的稀疏性。Pascual-Montano提出[83]引入光滑矩阵来同时平衡基向量和编

码向量的稀疏性。Chen[20]利用局部坐标约束提高编码向量的稀疏性和局部表示能力。

- 正交、对称非负矩阵分解。由于正交约束可以产生严格的聚类解释，Ding等人在文献[29]中提出正交非负矩阵分解和双向正交的非负矩阵三因子分解模型。文献[28]分析了对称非负矩阵分解与核Kmeans以及谱聚类的关系。文献[58]提出基于类牛顿法的方法求解对称矩阵分解问题并将其应用于图聚类问题；而文献[116]提出基于轮换坐标下降法的算法求解对称非负矩阵三因子分解并将其用于社区发现问题。

- 流形非负矩阵分解。标准的非负矩阵分解在欧式空间中测度矩阵的拟合程度。然而在一些实际应用中，数据点可能不是分布在欧式空间中，而从几何角度来看数据点通常分布在嵌入于外围欧式空间的一个潜在的流形体上[1]。因此流形正则的非负矩阵分解要求学习到的模型既能够优化数据在欧式空间的重构质量又可以体现数据中的流形结构。具体来说，图正则方法构造样本（或者特征）间相似关系图用于刻画数据的流形结构，并且要求如果两个样本比较相似其对应的非负因子也要足够接近。常见的方法包括：Cai[14–16]和Gu[38, 39]等人分别提出不同的方式构造关系图（如基于K近邻的关系图，基于局部学习、局部重构以及局部坐标编码等方式构造的关系图）并用于正则化的非负矩阵分解模型；为了同时刻画样本和特征的流形结构，Gu[37, 40]和Wang[90]等人分别提出对偶图正则的非负矩阵分解模型。为了更准确的刻画流形结构，Shen[91]等提出了基于稀疏重构的多流形正则方法，Du[31]和Wang[103]等提出集成流形图正则的非负矩阵分解模型。

- 判别式非负矩阵分解。标准的非负矩阵分解可以看成一种无监督的学习方式。通过结合判别式信息，基本的非负矩阵分解模型可以扩展为判别式的学习模型（如有监督或者半监督的矩阵分解）。通过结合Fisher准则（即最大化类间散度和最小化类内散度），文献[57, 104, 110]提出了判别式的非负矩阵分解模型即相应的求解算法；利用样本中的标签信息，文献[118]提出了有监督关系图正则的非负矩阵分解模型；通过进一步要求标签类别相同的样本对应相同的非负因子，文献[68, 69]提出了带标签约束的非负矩阵分解模型。

1.3.4　基于非负矩阵分解的相关应用

非负矩阵被广泛用于数据挖掘、机器学习和模式识别相关任务中。常见的应用包括：文本挖掘相关任务，如文本聚类[106]、Enron邮件分析[9]、文档自动摘要[97]；图像表示和理解[14, 20, 69]，如人脸识别[57, 88]、表情识别[118]、手写数字识别、纹理分类[88]等；互联网二元（dyadic）数据挖掘，如用基于指数分布（如Weibull分布）的非负矩阵分解刻画用户在网页上的停留时间[67]；计算生物学中microarray数据分析，如基于非负矩阵分解的亚型肿瘤自动归类[12]；遥感数据中的高光谱图像解混（Unmixing）与盲信号分离（Blind source separation）[85]；网络数据分析，如社交网络、Co-athor网络中的社区发现[98, 116]，图数据聚类[58]；基于协同过滤的推荐系统[19, 108, 115]等。

1.4　本文的工作

由于实际数据往往存在质量问题，如误差、噪声、异常点等，现有的代表性非负矩阵分解方法，如标准非负矩阵分解，图正则非负矩阵分解，以及基于对偶图正则非负矩阵三因子分解方法等不能有效的处理噪声数据，影响了聚类和联合聚类的效果。本文针对数据的质量问题，系统深入的研究如何提高非负矩阵分解的鲁棒性，以及如何提高基于非负矩阵分解的聚类分析效果。主要研究内容如下：

- 针对标准非负矩阵分解模型对显著误差和非高斯噪声敏感的问题，本文采用鲁棒半二次损失函数度量矩阵分解的质量，从而得到鲁棒的非负矩阵分解模型。常见的半二次损失函数包括鲁棒统计学中的M估计量、信息论中的相关熵失真度等。与标准非负矩阵分解采用的平方损失函数相比，非高斯噪声和大误差在这些函数上损失较小；与标准非负矩阵分解基于的高斯噪声假设相比，这些函数对应于在长尾数据上有较高概率的分布；为了求解不同半二次损失函数对应的非负矩阵分解问题，我们提出基于半二次最小化的通用求解过程，包括基于乘法形式和加法形式两类算法。我们给出了基于相关熵和休伯损失函数的两个鲁棒非负矩阵分解特例，同时说明当前的鲁棒非负矩阵分解模型也可以看成基于半二次最小化方法的特例。在噪声数据上的聚类实验结果表明本章提出基于相关熵失真度的鲁棒非负矩阵分解模型效果较好。

- 和聚类分析不同，联合聚类算法往往需要同时考虑样本与特征之间的关系，样本与样本之间的关系以及特征与特征之间的关系。同样地，这三部分数据中存在的噪声都会降低联合聚类算法的效果。因此，本文分别从鲁棒数据重构和鲁棒图正则两方面系统性地提高基于非负矩阵分解的联合聚类算法的鲁棒性。具体来讲，为了处理样本与特征关系矩阵中的噪声，我们引入一个稀疏的错误矩阵刻画样本与特征关系中存在的显著误差和非高斯噪声，并从错误矩阵中恢复不含显著误差的关系矩阵；为了进一步刻画恢复的关系矩阵中存在的高斯小噪声本文引入平方误差度量此关系矩阵上的重构误差。为了降低特征与特征以及样本与样本之间噪声关系的影响，我们采用绝对值图正则误差而不是常用的平方图正则误差，此外由于绝对值函数图正则产生的稀疏性还可以产生更紧凑的矩阵分解结果。在真实数据集上的联合聚类实验结果表明，我们提出的鲁棒联合聚类算法要好于当前的主流方法。

- 为了改善稀疏性噪声假设在实际应用中的局限性，我们首先证明本文之前提出的鲁棒联合聚类算法中基于稀疏性噪声的重构误差与休伯损失函数之间的对偶关系，然后进一步提出采用一般的半二次损失函数来分别度量特征与样本关系矩阵的重构误差以及样本与样本，特征与特征关系的图正则误差。我们分别推导出基于半二次优化乘法形式和加法形式的鲁棒图正则非负矩阵三因子分解模型的求解算法并证明其收敛性。

- 常见的矩阵分解方法都采用单值数据作为输入，而在一些实际应用中单值数据无法准确表示和刻画实际问题，因此不可避免地引入了一定的系统误差，这给后续的矩阵分解带来了挑战。本文提出用一个基于均匀分布的区间来近似真实数据可能的取值，并在人脸分析和协同过滤两个应用中提出了经验性方法构造该区间对应的上下界矩阵，本文通过扩展标准的非负矩阵分解和概率矩阵分解提出面向上下界区间的联合矩阵分解方法，即基于区间数据的非负矩阵分解和基于区间数据的概率矩阵分解模型。我们分别在人脸分析（包括人脸识别，聚类和重构）和协同过滤两个应用中系统性地比较了单值矩阵分解模型和对应的区间矩阵分解模型，大量实验结果表明本文提出的区间矩阵分解方法要显著优于标准的单值矩阵分解模型。

- 由于聚类分析的无监督特点，我们通常可以获得对同一数据的多种聚类划分，聚类集成研究如何从这些划分中学习一个鲁棒的一致性聚类。根据输入的多种划分，聚类集成方法构造基于簇特征和基于多重关联关系的两种数据表示形式，并分别基于这两种形式学习最终的一致性聚类。不同于分类集成，聚类集成仍然属于无监督学习，并没有一个统一的标准，因此如何利用这两种表示在聚类集成任务上的内在相关性提高最终的一致性聚类结果是一个很重要的问题。本文利用联合学习的思想提出基于集成图正则非负矩阵分解的聚类集成算法。其中给定加权关联关系时，该算法通过图正则非负矩阵分解提高非负因子的聚类质量；而当非负因子给定时，该算法利用非负因子和关联关系的匹配度学习多重关系的权重。可以看出基于簇特征和多重关联关系的一致性学习在迭代过程中互相提高。需要指出的是，由于聚类集成问题本身也可以看成一个聚类问题，因此本方法不仅可以解决聚类集成的问题，还可以直接用于聚类分析，即通过联合集成图正则和非负矩阵分解的方法一方面提高图正则的鲁棒性，一方面提高聚类效果。

1.5　本文的组织结构

针对非负数据中的非高斯大噪声，第二章研究基于半二次最小化鲁棒非负矩阵分解的聚类方法。针对非负数据中的稀疏大噪声以及噪声关系，第三章研究基于鲁棒非负矩阵分解的联合聚类算法。为了改善稀疏性噪声假设在实际应用中的局限性，第四章研究基于一般半二次损失函数的鲁棒联合聚类模型和算法。针对单值非负矩阵无法准确刻画真实数据的问题，第五章研究基于区间数据的矩阵分解方法。针对单个关系图无法准确刻画真实数据间关系的问题，第六章研究基于集成图正则非负矩阵分解的聚类集成算法。第七章总结本文的工作，并指出一些值得进一步研究的问题。

第二章　基于鲁棒非负矩阵分解的聚类

2.1　引言

在数据挖据和机器学习的许多应用中，我们经常面临高维数据带来的挑战。作为一种常见的数据分析方法，矩阵分解通常寻找两个或者多个低秩矩阵来近似原始数据。因此，矩阵分解技术可以看成是特征抽取和降维方法的一种，其结果可以看作是对原始数据的归约表示。常见的矩阵分解方法包括主成份分析（Principal Component Analysis，简称PCA）、奇异值分解（Singular Value Decomposition，简称SVD）和前面介绍的非负矩阵分解（Nonnegative Matrix Factorization，简称NMF）。与PCA和SVD不同，NMF试图通过两个非负因子来近似原始非负数据，由于非负约束的纯加性，非负矩阵分解可以用于同时获取数据的组成部分以及基于这些部分的混合系数。标准非负矩阵分解模型以及许多相关扩展模型通常采用$L2$范数（也就是平方误差）来度量矩阵分解对原数据的近似程度。标准非负矩阵分解得到了广泛的应用，并且可以证明最小化平方误差等价于最大化基于零均值高斯噪声的概率似然。噪声服从零均值高斯分布这一假设在实际应用中往往并不一定满足，文献[88]指出$L2$范数并不总是最佳的损失函数并提出采用推土机距离（Earth Mover's Distance，简称EMD）度量直方图数据上的矩阵分解质量，而文献[67]提出采用指数分布来描述互联网用户在网页上的停留时间（dwell time）这类二元数据（dyadic data）的生成过程。研究表明最小二乘平方误差对大的、非高斯噪声敏感，有时甚至单个异常点都可能任意的降低数据近似的质量[56]。因此本章的主要研究内容就是如何提高非负矩阵分解模型的鲁棒性，并用于提高噪声数据上的聚类效果。

由于标准非负矩阵分解模型对噪声敏感的原因主要来自于平方误差损失函数或者说高斯噪声假设，我们可以应用鲁棒统计学以及信息论中许多比平方误差在大误差上鲁棒性更好的损失函数度量矩阵分解质量或者采用在长尾（远离数据均值）数据上有着更高概率的分布刻画数据生成过程。我们面临的问题和挑战包括：1）如何选择合适的鲁棒损失函数或鲁棒概率分布。Sandler在文献[88]中指出无论平方误差还是KL失真度在实际应用中有可能是不

适当的，即不同的损失函数在不同的数据和任务上表现并不一致，也就是说如何选择最优的损失函数往往依赖于数据集和具体问题。2）如何求解基于这些损失函数的非负矩阵分解模型。当前的鲁棒非负矩阵分解方法主要集中于针对单个损失函数设计具体的求解算法，如Kuang等在文献[56]中提出针对基于ℓ_1损失函数的L_{21}-NMF和L_1-NMF的乘法更新算法；Zhang等在文献[114]提出针对稀疏大的、非高斯噪声的计算方法；Hamza等在文献[43]中采用梯度下降法求解采用hypersurface函数的非负矩阵分解模型；Cichocki等在文献[22]中采用alpha-beta失真度度量非负矩阵分解的近似程度。为了选择合适的鲁棒损失函数，一个可行的策略是设计通用的优化方法求解全部（或者至少一类）鲁棒损失函数的非负矩阵分解问题，并分别比较不同的损失函数在同一应用和数据集上的效果，然后根据实际问题选择合适的鲁棒非负矩阵分解模型。然而现有的方法对不同的鲁棒损失函数设计不同的求解算法，这一方面使我们很难比较现有鲁棒损失函数（而不是求解算法）的优劣，另一方面也说明如何设计通用的鲁棒学习算法是一个非常困难的问题。

本章中，我们首先分别从基于平方损失的重构误差最小化和基于高斯噪声的概率似然最大化两个角度介绍了标准非负矩阵分解模型，分析其对大的、非高斯噪声敏感的原因，并提出鲁棒非负矩阵分解的基本思想。为了提高非负矩阵模型的鲁棒性，我们提出采用在大的误差上比平方误差损失更鲁棒的一类损失函数，即半二次损失函数，来度量非负矩阵分解的质量。不同于直接优化单个鲁棒非二次（有可能是非凸的）损失函数，我们提出两类基于半二次优化的通用迭代式求解算法。为了检测数据中蕴含的噪声，我们提出基于半二次最小化乘法形式的求解算法，通过分别采用不同的鲁棒损失函数和误差形式（基于矩阵元素，基于样本以及基于特征）来发现数据中的异常元素、噪声样本以及噪声特征，并且通过降低这些噪声元素（样本或者特征）对整个优化问题的影响从而实现鲁棒分解的目的。为了矫正数据中的异常部分，我们提出基于半二次最小化加法形式的通用算法，通过分别采用不同的损失函数修正从原始数据中发现的错误，修正这些错误并恢复“干净”数据，进而对这些恢复后的数据进行非负近似以达到鲁棒分解的目的。我们给出不同损失函数以及相关乘法函数和加法函数的直观解释，讨论当前已有非负矩阵分解的相关扩展和变形与本章提出的鲁棒非负矩阵分解之间的关系，并且进一步讨论今后可能的扩展。在真实数据上的聚类实验结果表明，鲁棒非负矩阵分解模型要普遍优于标准非负矩阵模型，这也说明了非高斯噪声普遍存在于真实数据；本文提出的基于相关

熵度量的非负矩阵分解模型表现较好；然而同时需要指出的是不同的鲁棒非负矩阵分解模型在不同的数据集和任务上表现并不完全一致，这也说明了设计通用算法对比不同鲁棒损失模型的必要性。

本章的结构组织如下。在第2.2.1节，我们介绍本章使用的符号，标准非负矩阵分解模型，以及半二次优化的基本技术。在第2.3节，我们具体介绍鲁棒非负矩阵分解的算法思想，分别阐述基于半二次优化乘法形式和加法形式的通用求解算法并给出具体的鲁棒非负矩阵分解实例，并讨论本章算法和相关算法的关系。在第2.4节，我们比较了本章提出的算法在多个数据集上的聚类效果。最后，在第2.5节，我们给出本章小结。

2.2　预备知识

2.2.1　符号介绍

在这一节，我们简单介绍本章使用的符号习惯。大写字母，比如X和W，表示矩阵。上标T表示矩阵和向量的转置。对于方阵$A \in \mathbb{R}^{d\times d}$，$\mathrm{tr}(A)$代表矩阵$A$的迹，即对角线上元素之和。对于向量$X_i \in \mathbb{R}^d$，$||X_i||$表示$L_2$范数，$||X_i||_1$表示$L_1$范数。具体说来，

$$||X_i|| = (\sum_{j=1}^{d} X_{ij}^2)^{\frac{1}{2}} \qquad \text{and} \qquad ||X_i||_1 = \sum_{j=1}^{d} |X_{ij}| \tag{2.1}$$

表2.2.1总结了其他重要的符号和标记。

2.2.2　非负矩阵分解

给定一个非负数据矩阵$X \in \mathbb{R}^{N\times M}$, 矩阵的行对应于样本而列对应于特征。我们用$X_{i*}$表示矩阵的第$i$行用$X_{*j}$表示矩阵的第$j$列。非负矩阵分解的目的是学习两个非负低维矩阵$U \in \mathbb{R}^{N\times K}$ 和$V \in \mathbb{R}^{M\times K}$，用他们的乘积近似原始数据矩阵$X$，即：

$$X \approx UV^T. \tag{2.2}$$

许多度量标准[26, 59, 88]已经被用来量化非负矩阵分解的质量。Seung与Lee在文献[59]提出两个目标函数，包括平方欧式距离和Kullback-Leibler失真度。标

表 2.1: 第二章符号概要

符号	描述
$\mathbb{R}_+$	非负实数的集合
$\mathbb{R}_M$	M-维实向量空间
N	输入数据点的数目
M	数据点的维度
$X \geq 0$	非负数据矩阵$X \in \mathbb{R}_+^{N\times M}$，即矩阵的每个元素都是非负的$X_{ij} \geq 0$
k	非负因子的个数
$X_{i\cdot}$	X的第i行
$X_{\cdot j}$	X的第j列
U	特征非负因子$U \in \mathbb{R}_+^{N\times k}$
V	数据非负因子$V \in \mathbb{R}_+^{M\times k}$
E	误差矩阵$E \in \mathbb{R}^{N\times M}$表示非负因子乘积$UV^T$与数据矩阵$X$的近似程度
W	权重矩阵$W \in \mathbb{R}_+^{N\times M}$ 表示数据每个元素的权重
S	噪声矩阵$S \in \mathbb{R}^{N\times M}$刻画矩阵每个元素中存在的噪声

准的非负矩阵分解模型可以表示成最小化下面的平方误差之和

$$\begin{aligned} &\min_{U,V} \sum_{i=1}^{N} ||X_{i*} - U_{i*}V^T||^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - (UV^T)_{ij})^2 \\ &\text{s.t.} \quad U \geq 0, V \geq 0. \end{aligned} \tag{2.3}$$

为了更好的理解标准非负矩阵分解模型，我们介绍其对应的高斯噪声概率解释。不失一般性，数据中的元素X_{ij}受到额外噪声的影响，即$X_{ij} = \bar{X}_{ij} + \epsilon_{ij}$，其中$\bar{X}_{ij}$是没有被观测到的真实数据，$\epsilon_{ij}$是附加的噪声。假设噪声服从均值为零方差为$\sigma$的高斯分布，可知$X_{ij} \sim \mathcal{N}(\bar{X}_{ij}, \sigma^2)$，即

$$p(X_{ij}|\bar{X}_{ij}) \sim \exp\left\{-\frac{(X_{ij} - \bar{X}_{ij})^2}{2\sigma^2}\right\}. \tag{2.4}$$

此外假设数据的每个元素是独立同分布（i.i.d.），则数据对数似然函数可以写成

$$\max \quad \log \prod_i \prod_j p(X_{ij}|\bar{X}_{ij}) = -\frac{1}{2\sigma^2}\sum_i \sum_j (X_{ij} - \bar{X}_{ij})^2. \tag{2.5}$$

最大化上述似然函数等价于最小化等式右边的误差平方和$\sum_i \sum_j (X_{ij} - \bar{X}_{ij})^2$，通过假设$\bar{X}$由两个非负低秩矩阵$U, V$的乘积近似，我们就将基于高斯噪声的最大似然估计问题转变为带非负约束的最小化重构误差问题。

式子(2.3)中的标准非负矩阵分解模型是关于联合(U, V)的非凸（non-convex）优化问题，因此很难用非线性优化方法得到全局最优解。然而对于仅关于U或者仅关于V的子问题，式子(2.3)仍然是一个凸优化问题。因此可以通过分块坐标轮换法分别求解相关子问题来得到问题(2.3)的局部最优解。通用的非负矩阵分解求解算法通过迭代下面的乘法更新规则（multiplicative update rules）来求解（细节请参考文献[59]）。

$$U_{ik} = U_{ik} \frac{(XV)_{ik}}{(UV^T V)_{ik}} \tag{2.6}$$

$$V_{jk} = V_{jk} \frac{(X^T U)_{jk}}{(VU^T U)_{jk}}. \tag{2.7}$$

2.2.3　半二次优化

半二次优化是凸优化分析中一种常见的优化技术，已经广泛引用于数据挖据和机器学习问题中[45, 80]。许多数据挖据和机器学习都可以描述为一个求解某个目标函数的优化问题，实际应用中这些目标函数可能是非二次函数，可能是凸函数或者非凸函数。半二次优化的基本思路是利用共轭函数变换引入辅助变量得到一个在扩大定义域上的易处理的增广目标函数来降低优化的复杂度：当辅助变量固定时，增广目标函数是关于原始变量的二次优化问题，这也是该方法中“半二次”的意思。常见的半二次最小化包括基于乘法形式和加法形式[77, 79]，这两种形式的区别在于引入了不同的辅助函数和二次函数，因此对增广目标函数的求解方法也不一样。文献[78, 117]中给出了多种可通过半二次优化求解的目标函数，本文假设待优化的目标函数f满足下列条件[78]：

$H1$　$e \to f(e)$是$\mathbb{R}$上的函数（可以是凸函数也可以是凹函数）　(2.8)

$H2$　$e \to f(\sqrt{e})$是$\mathbb{R}_+$上的凹函数，　(2.9)

$H3$　$f(e) = f(-e), \forall e \in \mathbb{R}$即$f(e)$是对称函数，　(2.10)

$H4$　$e \to f(|e|)$是$\mathbb{R}_+$上的连续增函数，且当且仅当$e = 0$时$f(|e|) = 0$，　(2.11)

$H5$　$e \to f(|e|)$是$\mathbb{R}$上的连续可微函数，且$f'(0) = 0$　(2.12)

$H6$　$\lim_{e \to \infty} f(e)/e^2 = 0$.即，函数$f$比二次函数增长慢　(2.13)

定理2.1.
针对上述函数为目标的优化问题

$$\min \sum_i f(e_i)$$

等价于在扩大的定义域上的优化问题，即

$$\left\{\min \sum_i f(e_i)\right\} = \left\{\min \sum_i \left(\frac{1}{2} b_i e_i^2 + g(b_i)\right)\right\}$$

其中b_i是引入的辅助变量，而$g(b_i)$是$f(e_i)$的共轭函数。扩大后的优化问题可以通过下面的交替最小化方式实现：

$$\min_b \quad \sum_i \left(\frac{1}{2} b_i e_i^2 + g(b_i)\right) \tag{2.14}$$

$$\min_e \quad \left(\sum_i b_i e_i^2\right) \tag{2.15}$$

而第一个子问题的最优解可以通过下面的式子计算

$$b_i^* = \begin{cases} f''(0) & if\ e_i = 0, \\ \frac{f'(e_i)}{e_i} & if\ e_i \neq 0, \end{cases} \tag{2.16}$$

$$\tag{2.17}$$

注意第二个子问题是关于e的二次函数。

证明. 定义$h(e) = -f(\sqrt{e})$，根据假设$H2$可知$h(e)$是一个凸函数，根据假设$H4$可知$h(e)$是$\mathbb{R}_+$上的连续函数。根据凸共轭函数理论，$h(e)$的共轭函数h^*可以定义为$h^*(b) = \sup_{e\geq 0}\{be - h(e)\}$其中$b \in \mathbb{R}$。定义函数$h^*(b)$在$-\frac{1}{2}b$处的函数值为$g(b)$，可得

$$g(b) = h^*(-\frac{1}{2}b) = \sup_{e\geq 0}\{-\frac{1}{2}be - h(e)\} = \sup_{e\geq 0}\{-\frac{1}{2}be^2 + f(e)\}. \tag{2.18}$$

根据Fenchel - Moreau定理，凸共轭函数的共轭函数是其本身，即$(h^*)^* = h$，则函数$(h^*)^*$在点e^2处的函数值为：

$$-f(e) = h(e^2) = \sup_{b<0}\{be^2 - h^*(b)\} = \sup_{b>0}\{-\frac{1}{2}be^2 - h^*(-\frac{1}{2}b)\} = \sup_{b>0}\{-\frac{1}{2}be^2 - g(b)\}. \tag{2.19}$$

根据假设$H6$可知如果$b < 0$则$\lim_{e\to\infty} g(b) = +\infty$，根据假设$H4$可知如果$b = 0$则$\lim_{e\to\infty} g(b) = +\infty$，因此式子(2.18)中的上确界$\sup\{-\frac{1}{2}be^2 + f(e)\}$对应于$b > 0$。最后我们得到：

$$f(e) = \inf_{b\geq 0}\{\frac{1}{2}be^2 + g(b)\}. \tag{2.20}$$

现在考虑什么时候同时达到式子(2.18)的上界和式子（(2.19)）的下界。对于任意$\hat{b} > 0$，定义函数$p_{\hat{b}}(e) = \frac{1}{2}\hat{b}e + h(e)$，将$p_{\hat{b}}(e)$代入式子(2.18)可得$g(b) = \inf_{e\geq 0}\{p_{\hat{b}}(e)\}$。根据假设$H2$可知$p_{\hat{b}}(e)$是关于$e$的凸函数且$p_{\hat{b}}(0) = 0$，根据假设$H6$可知$\lim_{e\to\infty} p_{\hat{b}}(e) = +\infty$。因此可得$p_{\hat{b}}(e)$在某处$\hat{e} \geq 0$有唯一的最小值，且此时式子(2.18)达到上界，即$g(\hat{b}) = -\frac{1}{2}\hat{b}\hat{e}^2 + f(\hat{e})$，同时$f(\hat{e}) = \frac{1}{2}\hat{b}\hat{e}^2 + g(\hat{b})$由此可得式子(2.19)也同时达到下界。函数$p_{\hat{b}}(e)$的最小值可通过其导数计算，即$p'_{\hat{b}}(e) = \hat{b}/2 - f(\sqrt{e})/2\sqrt{e}$，根据假设$H2$可知$p'_{\hat{b}}(e)$是$\mathbb{R}_+$上的增函数。如果$p'_{\hat{b}}(0) \geq 0$也就是说$\hat{b} \geq f''(0^+)$，$p_{\hat{b}}(e)$在$e = 0$时取得最小值；否则当$p'_{\hat{b}}(0) = 0$也即是说$\hat{b} = f(\sqrt{e})/\sqrt{e}$时，函数$p_{\hat{b}}(e)$取得最小值。将此结果代入式子(2.19)并有假设$H3$可得式子(2.16)的结果。证毕。 □

2.3　鲁棒非负矩阵分解

这一节，我们首先介绍鲁棒非负矩阵分解的基本思想。其次，为了检测数据中蕴含的噪声，我们基于乘法形式的半二次最小化通用算法，通过分别采用的不同的损失函数和误差形式来发现数据中的异常元素、噪声样本以及噪声特征，并且通过降低这些噪声元素（样本或者特征）对整个优化问题的影响从而实现鲁棒分解的目的。再次，为了矫正数据中的异常部分，我们基于加法形式的半二次最小化通用算法，通过分别采用不同的损失函数修正在原始数据中发现的错误，并且通过对修正后的“干净”矩阵的非负近似进而达到鲁棒分解的目的。最后我们给出不同损失函数以及相关乘法函数和加法函数的直观解释，讨论当前已有非负矩阵分解的相关扩展和变形与本章提出的鲁棒非负矩阵分解之间的关系，并且进一步讨论今后可能的扩展。

2.3.1　鲁棒非负矩阵分解基本思想

在式子(2.3)和式子(2.5)中我们分别给出了标准非负矩阵分解的最小化重构误差形式和最大化似然概率解释。通过式子(2.3)，我们可以看出标准的非负矩

阵分解模型最小化矩阵每个元素拟合误差的平方之和。假设数据矩阵中存在噪声矩阵ϵ时，我们可以得到下面的带噪声的非负矩阵分解模型：

$$\min_{U,V} \quad \sum_i \sum_j (\bar{X}_{ij} + \epsilon_{ij} - (UV^T)_{ij})^2 = \sum_i \sum_j (E_{ij}^2 - 2E_{ij}\epsilon_{ij} + \epsilon_{ij}^2) \qquad (2.21)$$

可以看出，当矩阵每个元素中的噪声较小（如接近于零）时，噪声对整个目标函数的影响有限（甚至可以忽略），此时带噪声的模型可以看成是对式子(2.3)的近似。然而，当矩阵某些元素的噪声较大时，噪声对整个优化问题的影响也较大，此时获得的矩阵分解结果与实际结果偏差较大；并且当矩阵的某个（些）元素带有很大的噪声（如接近无穷）时，单个噪声元素都可能任意支配整个目标函数，此时学习的模型无法正确刻画数据的整体重构质量。通过式子(2.5)中的概率解释，我们可以看出标准的非负矩阵分解模型建立在噪声服从零均值高斯分布的基础上，其中噪声的量值(magnitude)比较小（噪声偏离期望三倍标准差的概率接近0.3%，即$p(X_{ij} > \bar{X}_{ij} + 3\sigma) + p(X_{ij} < \bar{X}_{ij} - 3\sigma) \approx 0.3\%$），这也就是说高斯分布下数据$X_{ij}$远离真值$\bar{X}_{ij}$的概率非常小。然而，实际应用中的获得的数据往往受到不同的干扰和破坏，数据可能会在较大的程度上偏离真值。

鲁棒矩阵分解的基本目的和特点是：少量异常元素引起的对理想分解结果的偏离较小；较多异常点的存在也不至于引起灾难性的失败。为了提高标准非负矩阵模型对大噪声、非高斯噪声等的鲁棒性，我们可以分别从最小化重构误差和最大化似然概率两个角度进行扩展。从最小化重构误差这一角度出发，矩阵分解中，一个合理的假设是理想的分解模型能够很好的近似非噪声数据矩阵，即矩阵每个元素的拟合误差$E_{ij} = \bar{X}_{ij} - (UV^T)_{ij}$较小。基于这一假设，当大的、非高斯噪声$\epsilon_{ij}$存在时，模型在该元素上的损失$(E_{ij}^2 - 2E_{ij}\epsilon_{ij} + \epsilon_{ij}^2)$主要由噪声决定，并且该元素上的误差大于其他非噪声元素上产生的误差，因此我们可以通过使用对大的拟合误差（大的误差往往由噪声引起）不敏感的损失函数来度量整个重构误差，降低大的拟合误差对整个矩阵分解模型的影响，从而提高模型的鲁棒性。和平方误差相比，对大的误差较为不敏感（即大的误差产生的损失小于平方损失）的函数包括：ℓ_1函数、相关熵度量、Huber函数等。

此外我们可以从最大化似然概率的角度分析如何提高非负矩阵分解的鲁棒性。我们可以采用在长尾数据（偏离均值较远的数据）上获得较高概率的分布来表示数据的生成过程，即提升远离均值数据的概率，降低集中于均值附近数

据的概率。此类常见的分布包括：学生t-分布（不包括高斯分布的特殊情况）、柯西分布、威布尔（Weibull ）分布、帕累托（Pareto）分布等连续概率分布。

实际上，基于最小化重构误差的扩展和基于最大化似然概率的鲁棒性扩展有着一定的内在关联性，如Kong等人在文献[56]中证明最小化ℓ_1函数的重构误差等价于最大化基于拉普拉斯分布的似然概率。

本章的主要工作针对一类能够通过半二次优化技术求解的损失函数（称之为半二次损失函数）提出通用的学习框架，从而得到一类鲁棒非负矩阵分解模型，并且证明当前的工作可以看成该通用算法的实例。

定义矩阵E为真实数据和NMF近似结果之间的拟合误差。在这一章，通过用其他损失函数替换原有的每个元素[1]上的平方拟合误差，可以得到非负矩阵分解的通用优化目标，即

$$\begin{aligned} \mathcal{J}(U,V) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\ell(E_{ij}) \\ \text{s.t.} \quad & U \geq 0, V \geq 0, \end{aligned} \tag{2.22}$$

其中$E_{ij} = X_{ij} - (UV^T)_{ij}$，$\ell(\cdot)$是对大误差不敏感的半二次损失函数。

2.3.2　基于半二次优化乘法形式的噪声检测算法

这里我们先介绍一个通用的基于半二次最小化乘法形式的鲁棒非负矩阵分解求解框架。我们会在以后用这个结果推导出多个鲁棒非负矩阵分解模型的求解算法。

2.3.2.1　基于半二次最小化的通用算法

根据共轭凸函数理论[11]和半二次优化理论[78, 79]以及定理2.1，当拟合误差给定时，以下等式成立

$$\ell(E_{ij}) = \min_{W_{ij}\in\mathbb{R}_+} Q(E_{ij}, W_{ij}) + \phi(W_{ij}), \tag{2.23}$$

[1]需要注意的是这里我们同样可以定义整行或者整列上的误差而不是每个元素上的误差，我们在第2.3.2.5节给出了基于异常样本的鲁棒CIM-NMF模型。

其中$\phi(W_{ij})$是$\ell(E_{ij})$的共轭函数，W_{ij}是对应的辅助变量，并且$Q(\cdot,\cdot):\mathbb{R}\to\mathbb{R}$是关于$E_{ij}$的二次函数。在本节中我们只考虑乘法形式的二次函数[79]，即

$$Q(E_{ij},W_{ij})=\frac{1}{2}W_{ij}E_{ij}^2. \tag{2.24}$$

将式子(2.23)和式子(2.24)代入等式(2.22)，可以得到下面的增广目标函数

$$\begin{aligned}&\min_{U,V}\left\{\mathcal{J}(U,V)=\sum_{ij}\ell(E_{ij})\right\}\\ =&\min_{U,V,W}\left\{\mathcal{J}(U,V,W)=\sum_{ij}[\frac{1}{2}W_{ij}E_{ij}^2+\phi(W_{ij})]\right\}\end{aligned} \tag{2.25}$$

而问题(2.25)可以通过下面的交替最小化方法求解

- 当U和V固定时，问题(2.25)是关于W的凸问题。最优解可以通过下面的式子计算

$$W_{ij}=\begin{cases}\ell''(E_{ij}),\text{if } E_{ij}=0\\ \frac{\ell'(E_{ij})}{E_{ij}},\text{otherwise}\end{cases} \tag{2.26}$$

  需要指出的是W_{ij}仅取决于损失函数$\ell(\cdot)$。在一个合理的模型中，噪声往往会产生非常大的拟合误差。因此对应的权重应该比较小；而绝大多数正常样本的误差应该比较小，相应的权重也应该比较大。因此，W_{ij}可以看成噪声掩码（mask）不断的减小大误差部分对应的权重，降低噪声对整个优化问题的影响，从而提高矩阵分解的鲁棒性。

- 当W固定时，问题(2.25) 可以简化为加权非负矩阵分解。我们可以通过下面的更新公式来获得其局部最优解。

$$U_{ik}=U_{ik}\frac{(W\otimes XV)_{ik}}{(W\otimes(UV^T)V)_{ik}} \tag{2.27}$$

$$V_{jk}=V_{jk}\frac{((W\otimes X)^TU)_{jk}}{((W\otimes(UV^T))^TU)_{jk}}. \tag{2.28}$$

  关于加权非负矩阵分解的推导请见第2.3.2.2节。

2.3.2.2　加权非负矩阵分解

通过给矩阵每个元素赋予一个非负权重，基于平方误差的加权非负矩阵可以表示为下面的优化问题

$$\min_{U,V} \quad \sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}(X_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk})^2 \tag{2.29}$$

$$\text{s.t.} \quad U \geq 0, V \geq 0, \tag{2.30}$$

其中$W \in \mathbb{R}^{N\times M}$是一个非负权重矩阵，用来表示矩阵每个元素的重要性。问题(2.29)同样是一个关于(U,V)是非凸优化问题，我们可以采用类似于最小化问题(2.3)的方法来学习问题(2.29)的局部最优解。

固定V，计算U：当V给定时，关于U的优化问题等价于最小化

$$\mathcal{L}_U(U) = \sum_{i=1}^{N}(X_{i*} - U_{i*}V^T)A_i(X_{i*} - U_{i*}V^T)^T + \text{tr}(\Phi U), \tag{2.31}$$

其中$A_i = \text{diag}(W_{i*}) \in \mathbb{R}^{M\times M}$，$\Phi = [\Phi_{ik}]$是关于非负约束$U_{ik} \geq 0$的拉格朗日乘子。目标函数$\mathcal{L}_U$关于$U_{ik}$的偏导数为

$$\frac{\partial \mathcal{L}_U}{\partial U_{ik}} = -2(X_{i*}A_iV)_k + 2(U_{i*}V^TA_iV)_k + \Phi_{ik}. \tag{2.32}$$

利用KKT条件$\Phi_{ik}U_{ik} = 0$，可得下面的等式

$$[-(X_{i*}A_iV)_k + (U_{i*}V^TA_iV)_k]U_{ik} = 0. \tag{2.33}$$

等式(2.33)可以进一步推出下面的更新规则

$$U_{ik} = U_{ik}\frac{(X_{i*}A_iV)_k}{(U_{i*}V^TA_iV)_k} = U_{ik}\frac{(W\otimes XV)_{ik}}{(W\otimes(UV^T)V)_{ik}}, \tag{2.34}$$

其中$\otimes$是Hadamard运算符，也就是矩阵基于元素的乘法运算，这里我们假设Hadamard运算符比通用的矩阵乘法运算有更高的优先级，即$AB\otimes CD = A(B\otimes C)D$。

固定U，计算V：当U给定时，关于V的优化问题等价于最小化

$$\mathcal{L}_V(V) = \sum_{j=1}^{M}(X_{*j} - UV_{j*}^T)^TB_j(X_{*j} - UV_{j*}^T) + \text{tr}(\Psi V), \tag{2.35}$$

其中$B_j = \mathrm{diag}(W_{*j}) \in \mathbb{R}^{N\times N}$。和问题(2.31)类似，问题(2.35)关于$V$的最优解可以通过下面的更新规则计算

$$V_{jk} = V_{jk}\frac{(X_{*j}B_jU)_k}{(V_{j*}U^TB_jU)_k} = V_{jk}\frac{((W\otimes X)^TU)_{jk}}{((W\otimes(UV^T))^TU)_{jk}}. \tag{2.36}$$

2.3.2.3 相关熵失真度

近年来，信息理论学习(Information Theoretic Learning)中提出了相关熵的概念来处理非高斯噪声和大的噪点。根据Renyi熵[84]的信息势，相关熵是两个随机变量$\boldsymbol{x}$和$\boldsymbol{y}$的通用相似度，定义为

$$V_\sigma(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Expectation}[k_\sigma(\boldsymbol{x}-\boldsymbol{y})], \tag{2.37}$$

其中$k_\sigma(\cdot)$是核函数。实际应用中数据的联合概率密度往往是未知的，我们可以得到相关熵在有限样本集合$\{(x_i,y_i)\}_{i=1}^n$上的估计量：

$$\hat{V}_\sigma(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}k_\sigma(x_i-y_i). \tag{2.38}$$

Liu等人在文献[70]中进一步提出了基于相关熵的失真度（Correntropy Induced Metric, 简称CIM）用于测度样本空间中任意两个向量的距离

$$\mathrm{CIM}(\boldsymbol{x},\boldsymbol{y}) = (k(0) - \frac{1}{n}\sum_{i=1}^{n}k_\sigma(e_i))^{1/2}, \tag{2.39}$$

其中$e_i = x_i - y_i$，本文中我们仅考虑高斯核函数$g(e,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-e^2/2\sigma^2)$。我们在图2.1中画出了CIM损失函数在不同误差上的变化曲线，可以看出当误差较小时CIM损失函数接近于平方误差，当误差增大时CIM接近于绝对值误差而当误差继续增加时CIM函数接近于ℓ_0误差，即此时的损失接近常数1。CIM损失函数在不同误差范围的变化受核参数σ影响。

2.3.2.4 CIM-NMF

将问题(2.3)中的平方误差替换为平方的CIM，我们得到CIM-NMF最小化下面的目标函数

$$\mathcal{J}(U,V) = 1 - \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}g(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk},\sigma),$$

等价于解决下面的优化问题

$$\begin{aligned} \min_{U,V} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}(1-g(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk},\sigma)) \\ \text{s.t.} \quad & U\geq 0, V\geq 0. \end{aligned} \tag{2.40}$$

根据，CIM-NMF对应的优化问题等价于在增大的参数空间(U,V,W)中最小化下面的增广目标函数

$$\min_{U,V,W}\sum_{i=1}^{N}\sum_{j=1}^{M}[W_{ij}(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2+\phi(W_{ij})], \tag{2.41}$$

问题(2.41)可以通过第2.3.2.1节中的通用算法来求解。我们迭代求解关于W和U,V的子优化问题直到算法收敛。

计算W：当U和V给定时，关于W的优化问题的最优解可以通过下面的式子计算

$$\begin{aligned} W_{ij} &= \frac{\frac{d}{dE_{ij}}(1-g(E_{ij},\sigma))}{E_{ij}} \\ &\propto \exp(-\frac{(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2}{2\sigma^2}). \end{aligned} \tag{2.42}$$

计算U和V： 当W给定时，关于U和V的优化问题(2.41)变为加权的非负矩阵分解(2.29)问题。因此，可以直接用于式子(2.27)中的乘法更新规则来学习U和V。

与其他核方法类似核参数的选择将影响算法的性能，本文中我们根据文献[45]的建议经验性的将核参数设置为平均重构误差，也就是说

$$\sigma^2=\frac{\gamma}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2. \tag{2.43}$$

其中γ是可调节参数，如无特殊说明，我们使用$\gamma=1$。根据上述讨论，CIM-NMF的具体算法流程如下：

2.3.2.5　rCIM-NMF

许多应用中我们有关于噪声的额外知识。比如，在DNA数据集如果一条记录（某一行)被破坏，那么该记录的绝大多数元的质量可能比较差，因此可将

Algorithm 1 CIM-NMF 算法描述

输入： 数据矩阵$X \in \mathbb{R}^{N\times M}$, 初始化矩阵$U \in \mathbb{R}^{N\times K}$和$V \in \mathbb{R}^{M\times K}$

输出： U, V 和权重矩阵W

repeat

计算W，其中$W_{ij} = \exp(-\frac{(X_{ij}-\sum_{k=1}^{K} U_{ik}V_{jk})^2}{2\sigma^2})$;

计算U，其中$U_{ik} = U_{ik}\frac{(W\otimes XV)_{ik}}{(W\otimes(UV^T)V)_{ik}}$;

计算V，其中$V_{jk} = V_{jk}\frac{((W\otimes X)^T U)_{jk}}{((W\otimes(UV^T))^T U)_{jk}}$;

根据公式(2.43)计算σ;

until Converges

整个记录（整行）看作异常样本。为了发现这样的噪声模式并且自动的减少对整个模型的影响，我们可以将矩阵重构中在每一行上的损失看成一个整体，并通过采用对大的拟合误差不敏感的损失函数度量整个行上的损失，得到基于样本的鲁棒非负矩阵分解。

这里我们采用基于行（row-based）的相关熵失真度作为损失函数，得到以下的rCIM-NMF：

$$J(U,V) = 1 - \frac{1}{N}\sum_{i=1}^{N} g(||X_{i*} - U_{i*}V^T||, \sigma), \tag{2.44}$$

最小化上面的目标函数等价于求解以下问题：

$$\begin{aligned} \min_{U,V} \quad & \sum_{i=1}^{N}(1 - g(||X_{i*} - U_{i*}V^T||, \sigma)) \\ \text{s.t.} \quad & U \geq 0, V \geq 0. \end{aligned} \tag{2.45}$$

和式子(2.40)中的CIM-NMF类似，我们也可以通过第2.3.2.1节的通用算法求解。类似的，式子(2.45)中的优化问题等价于最小化下面的在扩大的定义域上的优化问题

$$\min_{U,V,\boldsymbol{w}} \quad \sum_{i=1}^{N}[w_i||X_{i*} - U_{i*}V^T||^2 + \phi(w_i)], \tag{2.46}$$

其中w_i是矩阵中第i行X_{i*}的权重。

计算$\boldsymbol{w}$： 根据式子(2.26), 我们可以用下面的公式计算w_i

$$w_i = \exp(-\frac{||X_{i*} - U_{i*}V^T||^2}{2\sigma^2}). \tag{2.47}$$

计算U和V： 当$\boldsymbol{w}$给定时，关于U和V的优化问题等价于下面的加权非负矩阵分解

$$\min_{U,V} \quad \sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}(X_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk})^2, \tag{2.48}$$

其中$W = \boldsymbol{w}\mathbf{1}_M^T$。因此，$U$和$V$可以分别通过公式(2.34)和公式(2.36)来计算。

此外rCIM-NMF中的核参数也可以采用类似于式子(2.43)的方式来计算。rCIM-NMF详细的求解过程见算法2。

Algorithm 2 rCIM-NMF算法描述

输入： 数据矩阵$X \in \mathbb{R}^{N\times M}$, 初始化矩阵$U \in \mathbb{R}^{N\times K}$和$V \in \mathbb{R}^{M\times K}$.

输出： U, V 和权重向量$\boldsymbol{w}$

repeat

计算w_i，其中$w_i = \exp(-\frac{||X_{i*}-U_{i*}V^T||^2}{2\sigma^2}), W = \boldsymbol{w}\mathbf{1}_M^T$;

计算U，其中$U_{ik} = U_{ik}\frac{(W\otimes XV)_{ik}}{(W\otimes(UV^T)V)_{ik}}$;

计算V，其中$V_{jk} = V_{jk}\frac{((W\otimes X)^TU)_{jk}}{((W\otimes(UV^T))^TU)_{jk}}$;

将σ^2设为$\frac{1}{N}\sum_{i=1}^{N}||X_{i*} - U_{i*}V^T||^2$;

until Converges

2.3.2.6 Huber-NMF

在鲁棒统计中，M-估计量是指最小化数据。被广泛应用于机器学习和数据挖掘中的鲁棒任务中。在鲁棒回归中，IRLS通常用于求解M-估计量的问题，另外一种常用的方法是半二次优化，通过乘法形式和加法形式对M-估计量的重新形式化，原始问题可以通过迭代最小化一个增广目标函数来求解。常见的M-估计量包括L_p 函数，L1-L2函数，Huber函数。如何选择一个通用的合适的M-估计量是比较困难的。本章中，考虑到Huber函数和传统的L2函数以及鲁棒的L1函数之间的关系，我们使用Huber函数来测度矩阵的近似程度。

$$\ell_{\text{huber}}(e) = \begin{cases} e^2 & \text{if } |e| \le c \\ 2c|e| - c^2 & \text{if } |e| \ge c \end{cases}, \tag{2.49}$$

其中c是截断参数用于折中L_2-norm和L_1-norm。Huber-NMF可以形式化成下面的优化问题：

$$\min_{U,V} \quad \sum_{i=1}^{N}\sum_{j=1}^{M} \ell_{\text{huber}}(E_{ij}), \tag{2.50}$$

其中$E_{ij} = X_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk}$。再次，Huber-NMF可以通过第2.3.2.1节中的通用算法来求解。

计算W： 给定U和V，W的最优解可以通过下式计算

$$W_{ij} = \begin{cases} 1 & \text{if } |E_{ij}| \leq c \\ \frac{c}{|E_{ij}|} & \text{otherwise} \end{cases} \tag{2.51}$$

计算U和V： 同样的，可以通过式子(2.34)和式子(2.36)最小化问题(2.50)中的目标函数。参考Ding等人在文献[30]中针对鲁棒PCA提出的截断参数估计方法，我们将Huber-NMF中的截断参数经验性的设置为拟合误差的中位数。

$$c = \gamma \text{median}(|E_{ij}|). \tag{2.52}$$

算法5列出了完整的Huber-NMF的计算过程。

Algorithm 3 Huber-NMF算法描述

输入： 数据矩阵$X \in \mathbb{R}^{N\times M}$, 初始化矩阵$U \in \mathbb{R}^{N\times K}$和$V \in \mathbb{R}^{M\times K}$.

输出： U, V 和权重矩阵W

repeat

计算W，其中$W_{ij} = \begin{cases} 1 & \text{if } |E_{ij}| \leq c \\ \frac{c}{|E_{ij}|} & \text{otherwise} \end{cases}$;

计算U，其中$U_{ik} = U_{ik}\frac{(W\otimes XV)_{ik}}{(W\otimes(UV^T)V)_{ik}}$;

计算V，其中$V_{jk} = V_{jk}\frac{((W\otimes X)^TU)_{jk}}{((W\otimes(UV^T))^TU)_{jk}}$;

根据公式(2.52)计算c

until Converges

2.3.3 基于半二次优化加法形式的噪声矫正算法

不同于第2.3.2节中接受的基于乘法形式的优化算法，这里我们介绍一个通用的基于半二次最小化加法形式的鲁棒非负矩阵分解求解框架。我们会在以后

用这个结果推导出多个鲁棒非负矩阵分解模型的求解算法。乘法形式与加法形式的区别在于引入的二次函数形式不同，其对应的鲁棒性解释不同，最有求解过程也不同。

2.3.3.1　基于半二次最小化的通用算法

根据共轭凸函数理论[11]和半二次优化理论[79]，当拟合误差给定时，以下等式成立

$$\ell_{E_{ij}} = \min_{S_{ij}\in\mathbb{R}} Q(E_{ij}, S_{ij}) + \phi_A(E_{ij}). \tag{2.53}$$

其中S_{ij}是引入的辅助变量，$\phi_A(S_{ij})$是$\ell(S_{ij})$加法形式对应的共轭函数，并且$Q(\cdot,\cdot) : \mathbb{R} \to \mathbb{R}$是关于$E_{ij}$的二次函数，本节中我们只考虑加法形式的二次函数[79]，即

$$Q(E_{ij}, W_{ij}) = (E_{ij} - S_{ij})^2. \tag{2.54}$$

将式子(2.53)和式子(2.54)代入等式(2.22)，可以得到下面的增广目标函数

$$\begin{aligned} &\min_{U,V} \left\{ \mathcal{J}(U,V) = \sum_{ij} \ell(E_{ij}) \right\} \\ &= \min_{U,V,S} \left\{ \mathcal{J}(U,V,S) = \sum_{ij} [\frac{1}{2}(E_{ij} - S_{ij})^2 + \phi_A(S_{ij})] \right\} \end{aligned} \tag{2.55}$$

而问题(2.55)可以通过下面的交替最小化方法求解

- 计算辅助变量。根据半二次优化理论[79]，辅助变量可以通过下面的式子计算：

  $$S_{ij} = E_{ij} - \ell'(E_{ij}). \tag{2.56}$$

  可以看出，由于半二次损失函数的导数是减函数，在误差较大时增长较小也就是此时的噪声较大，而给定噪声S后可以从原数据中恢复不含噪声的数据矩阵$X - S$。因此加法形式的半二次最小化可以看成对数据中噪声的抽取和矫正过程。由于大误差对应的噪声被矫正，我们就可以从恢复矩阵中学习合适的矩阵分解模型。

- 求解关于U和V的问题。当S给定时，式子(2.55)等价于求解矩阵的非负分解问题。而该问题可通过第2.3.3.2节中的算法来计算。

2.3.3.2 矩阵的非负因子分解

在这一节，我们针对矩阵（可能还有负数）的非负因子分解问题给出两个常见的乘法更新公式。矩阵的非负因子分解可以形式化为下面的优化问题：

$$\min \quad ||X - S - UV^T||^2 \tag{2.57}$$

$$\text{s.t.} \quad U \geq 0, V \geq 0, \tag{2.58}$$

其中$Xs = X - S \in \mathbb{R}_{\pm}^{N\times M}$，不同于非负矩阵分解，矩阵$Xs$可能含有负数。式子(2.57)是一个非凸优化问题，因此很难用非线性优化方法得到全局最优解。而我们可以采用分块坐标下降法（Block Coordinate Descent，简称BCD）分别求解关于U和V的子问题。本节中我们给出下面两个算法来计算该问题的局部最优解。

算法一 引入Lagrangian乘子$\Phi \in \mathbb{R}^{d\times k}$和$\Psi \in \mathbb{R}^{n\times k}$可得下面的Lagrangian函数

$$\mathcal{L}(U, \Phi) = ||X - S - UV^T||^2 - \text{tr}(\Phi U^T) - \text{tr}(\Psi V^T) \tag{2.59}$$

对目标函数求偏导可得

$$\Phi = -2(X - S)V + 2UV^TV \tag{2.60}$$

$$\Psi = -2(X - S)^TU + 2VU^TU \tag{2.61}$$

利用KKT最优性条件[27]$\Phi_{ij}U_{ij}^2 = 0$，即$[UV^TV - (X - S)V]_{ij}U_{ij} = 0$，我们可以得到下面的更新公式

$$U_{ij} = U_{ij}\sqrt{\frac{[(X - S)V]_{ij}^+}{[UV^TV + ((X - S)V)^-]_{ij}}} \tag{2.62}$$

$$V_{ij} = V_{ij}\sqrt{\frac{[(X - S)^TU]_{ij}^+}{[VU^TU + ((X - S)^TU)^-]_{ij}}} \tag{2.63}$$

算法二

定理2.2.

定义以下的非负二次优化问题

$$\min \quad \frac{1}{2}\boldsymbol{x}^TA\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} \qquad \text{s.t.} \quad \boldsymbol{x} \geq 0 \tag{2.64}$$

其中$\boldsymbol{x}$是d-维非负向量，A是对称半正定矩阵，$\boldsymbol{b}$是任意的d-维向量。分别定义$A^+ = (|A| + A)/2$和$A^- = (|A| - A)/2$并且$A = A^+ - A^-$。可以证明该问题的最优解可以通过以下的式子迭代更新[13, 89]

$$x_i = \left[\frac{-b_i + \sqrt{b_i^2 + 4(A^+\boldsymbol{x})_i(A^-\boldsymbol{x})_i}}{2(A^+\boldsymbol{x})_i}\right] x_i \tag{2.65}$$

可以看出式子(2.57)是分别关于U(或V)的二次函数，因此定理2.64可以直接用来求解问题(2.57)，我们可以得下面的更新公式。

$$U_{ij} = U_{ij}\left[\frac{|((S-X)V)_{ij}| - ((S-X)V)_{ij}}{2(UV^TV)_{ij}}\right] \tag{2.66}$$

$$V_{ij} = V_{ij}\left[\frac{|((S-X)^TU)_{ij}| - ((S-X)^TU)_{ij}}{2(VU^TU)_{ij}}\right] \tag{2.67}$$

2.3.3.3　CIM-NMF(A)

对于式子(2.40)中定义的CIM-NMF，我们也可以通过基于半二次优化加法形式的算法来计算。其中，当U和V给定时，根据式子2.56辅助变量S可以通过下面的公式计算：

$$S_{ij} = E_{ij} - \frac{2E_{ij}}{\sigma^2}\exp(-\frac{E_{ij}^2}{\sigma^2}) \tag{2.68}$$

当S给定时，可以通过第2.3.3.2节的算法计算U和V。和乘法形式的半二次优化算法类似，我们以可以经验性的将CIM失真度中的核参数设置为平均重构误差。算法5列出了完整的计算过程。

Algorithm 4 CIM-NMF(A)算法描述

输入： 数据矩阵$X \in \mathbb{R}^{N\times M}$, 初始化矩阵$U \in \mathbb{R}^{N\times K}$和$V \in \mathbb{R}^{M\times K}$.

输出： U, V 和噪声矩阵S

repeat

根据公式(2.68)计算S;

根据公式(2.62)或(2.66)计算U;

根据公式(2.62)或(2.66)计算V;

根据公式(2.43)计算σ;

until Converges

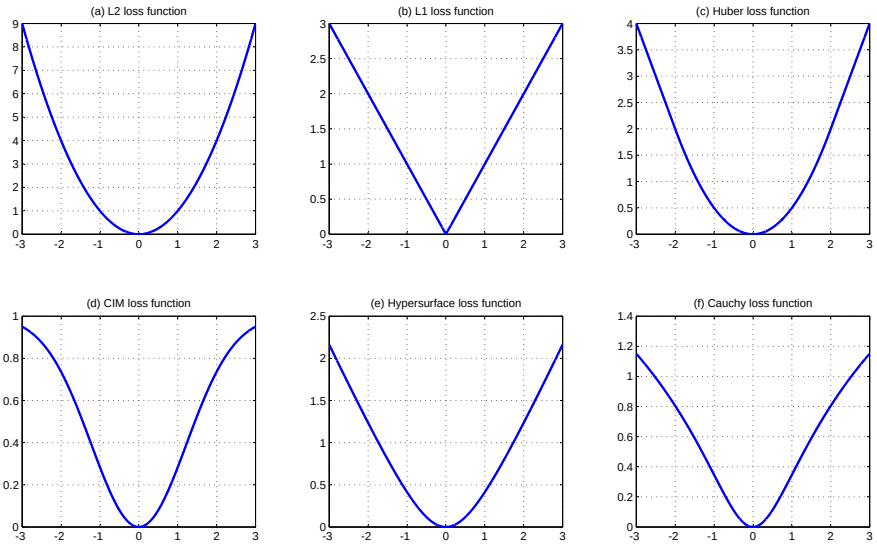


图 2.1: 常见的损失函数

2.3.3.4 Huber(A)-NMF

类似的，我们也可以通过算法5中列出的基于半二次优化加法形式的计算过程求解式子(2.50)中定义的Huber-NMF。

Algorithm 5 Huber-NMF(A)算法描述

输入： 数据矩阵$X \in \mathbb{R}^{N\times M}$, 初始化矩阵$U \in \mathbb{R}^{N\times K}$和$V \in \mathbb{R}^{M\times K}$.

输出： U, V 和噪声矩阵S

repeat

计算S，其中$S_{ij} = \begin{cases} 0 & |E_{ij}| \le c \\ c - c\text{sign}(E_{ij}) & |E_{ij}| > c \end{cases}$；

根据公式(2.62)或(2.66)计算U；

根据公式(2.62)或(2.66)计算V；

根据公式(2.52)计算c

until Converges

2.3.4 讨论和扩展

2.3.4.1 半二次损失函数

为了更好的理解这些函数的表现，我们分别在图2.1，2.2，2.3中画出了部

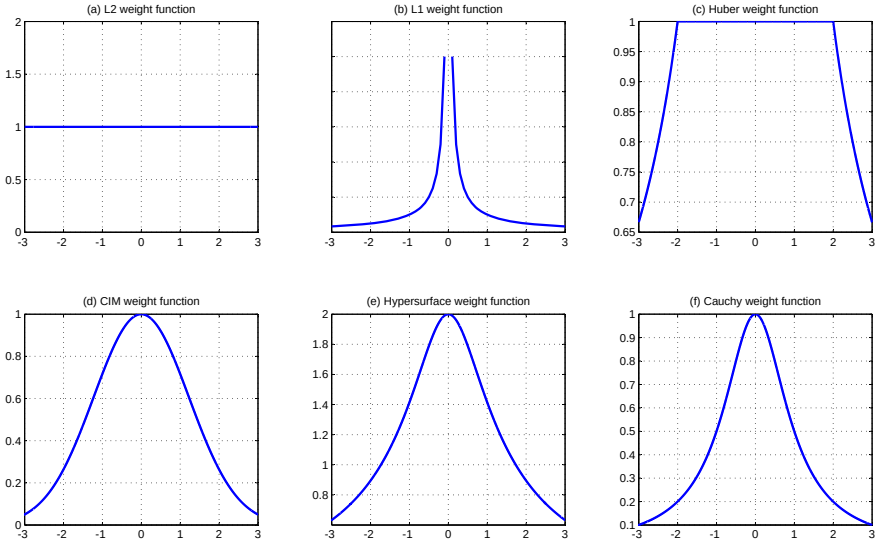


图 2.2: 对应的权重函数

分损失函数和其对应的基于半二次优化乘法形式的权重函数以及半二次优化加法形式的噪声函数。乘法形式通过减小大误差部分的权重降低噪声的影响，加法形式通过从大误差中恢复不含噪声的数据都降低了噪声对整个优化问题的影响，提高了非负矩阵分解的鲁棒性。

从图2.1中可以看出鲁棒损失函数（不包括$L2$函数）对大误差带来的损失都小于$L2$函数，即$\lim_{e\to\infty}\ell(e)/e^2=0$；其中Welsch（也就是CIM）等函数是有界的，限制了大的拟合误差对整个优化问题的影响。而从图2.2中可以看出鲁棒损失函数在大误差上的权重比较小，而从图2.3中可以看出鲁棒损失函数在大误差上的噪声比较明显。

2.3.4.2　和现有算法的关系

在本小节中，我们主要讨论本章提出的算法和现有算法的关系。显然，标准的非负矩阵分解问题可以可以看成加权非负矩阵分解的特例，其中$W=\mathbf{1}^{N\times M}$，并且在优化过程中保持不变。为了提高非负矩阵分解模型的鲁棒性，Kong等人在文献[56]中提出了最小化行上ℓ_1误差的L_{21}-NMF，问题可以写成：

$$\min_{U,V}\quad\sum_{i=1}^{N}||X_{i*}-U_{i*}V^T||, \tag{2.69}$$

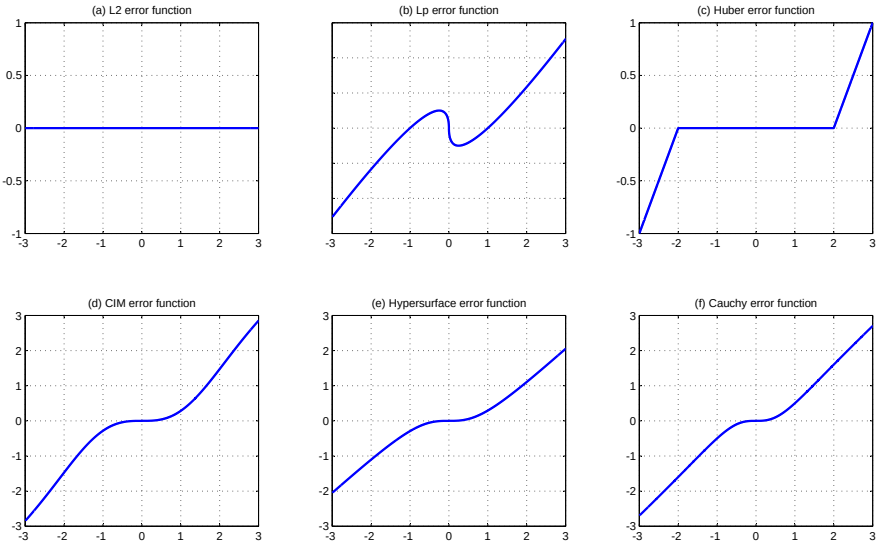


图 2.3: 对应的噪声函数

和最小化矩阵每元上L_1误差的L_1-NMF

$$\min_{U,V} \quad \sum_{i=1}^{N}\sum_{j=1}^{M}|X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk}|, \tag{2.70}$$

此外Hamza和Brady在文献[43]中提出最小化下面的超曲面（hypersurface）损失函数（也叫L_1-L_2 函数）来学习鲁棒非负矩阵分解

$$\min_{U,V} \quad \sqrt{1+\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\sum_{k=1}^{K}U_{ik}V_{jk})^2}-1. \tag{2.71}$$

事实上，L_1和L_1-L_2函数都属于鲁棒统计中的M估计量。这也就是说，$L_{2,1}$-NMF，L_1-NMF和L_1-L_2-NMF都可以通过本章提出的通用求解算法了优化。

另外需要指出的是文献[40, 115]中提出了加权非负矩阵分解的一个特例来解决协同过滤的问题。这些方法使用一个二元权重矩阵来表示用户评分是否缺失，并且在模型学习中这些权重保持不变。然而在本章中，我们利用加权来提高非负矩阵分解的鲁棒性，并且这些权重在学习过程中自动的更新和设置。

此外，本章的方法和Wang在文献[97]中提出的WFS-NMF也有一定的关系。在WFS-NMF中每个样本和每个特征都被赋予了不同的权重用于分别表示样本

和特征的重要性，WFS-NMF可以写成下面的优化问题

$$\begin{aligned} \min_{U,V,W} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}(X_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk})^2 \\ \text{s.t.} \quad & U \geq 0, V \geq 0, W_{ij} = a_i b_j, \sum_{i=1}^{N} a_i^{\alpha} = 1, \sum_{j=1}^{M} b_j^{\beta} = 1. \end{aligned} \tag{2.72}$$

最后我们指出Zhang在文献[114]中提出的鲁棒非负矩阵分解可以看成本章方法的一个特例。

$$\min \quad ||X - S - UV^T||_F^2 + \lambda|S|_1, \quad \text{s.t.} \quad U \geq 0, V \geq 0. \tag{2.73}$$

给定U, V，式子2.73关于S的优化问题等价于最小化

$$\min_{S} \quad ||E - S||_F^2 + \lambda|S|_1, \tag{2.74}$$

其中$E = X - UV^T$，根据文献[114]，S可以通过下面的式子计算

$$S_{ij} = \begin{cases} E_{ij} - \frac{\lambda}{2} & \text{if } E_{ij} > \frac{\lambda}{2} \\ E_{ij} + \frac{\lambda}{2} & \text{if } E_{ij} < -\frac{\lambda}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{2.75}$$

将S代入式子(2.74)可得如下关于U, V的优化问题

$$\min \sum_{i}\sum_{j} \begin{cases} (X_{ij} - (UV)_{ij})^2 & |E_{ij}| < \frac{\lambda}{2} \\ \lambda|X_{ij} - (UV)_{ij}| - \frac{\lambda^2}{4} & \text{otherwise.} \end{cases} \tag{2.76}$$

可以看出上面的式子等价于式子(2.49)中的Huber损失函数。

2.4　实验结果

本节将展示大量的实验结果来证明本章所提出算法的有效性，实验主要集中在对有无明显噪声数据集的聚类和分类结果比较。

2.4.1　对比算法和参数设置

为了比较本章提出的CIM-NMF，rCIM-NMF和Huber-NMF的效果，我们对比的算法包括：（1）Kmeans，代表性的划分式聚类算法；（2）PCA-Km，首

先利用PCA对数据降维，然后采用Kmeans聚类；（3）RPCA-Km，首先利用文献[17]中的鲁棒PCA对数据降维，然后采用Kmeans聚类；(4)NCut[75]，代表性的谱聚类算法；(5)NMF[59]，标准的非负矩阵分解；(6)SR-NMF[114]，基于稀疏性噪声假设的鲁棒非负矩阵分解；(7)L_{21}-NMF[56]，采用L_{21}范数为重构损失的鲁棒非负矩阵分解；(8)WFS-NMF[97]，加权的非负矩阵分解，其中样本和特征都赋予不同的权重。

为了公平的对比这些算法，相关参数设置如下。PCA和RPCA的维度设置为保存99%的方差。RPCA和SR-NMF的正则化参数搜索空间为$\{0.1, 0.2, 0.3, 0.4, 0.5\}$。NCut中的相似度采用高斯核函数并且核参数按照文献[111]中的方法自动设置。对于NMF以及其变种，我们将其维度设置为数据中簇的个数，并且根据文献[29]和[56]的建议采用Kmeans的结果来进行初始化。WFS-NMF的幂参数α和β根据文献[97]的建议设置为0.7。

2.4.2 评价指标

给定聚类算法生成的簇和真实类别标签，我们通过比较下面两个指标来评价聚类算法的性能。

聚类准确性(clustering accuracy, 简称ACC) 聚类准确性是基于类和簇的一一对应关系来评价聚类性能，对于样本$\boldsymbol{x}_i$，p_i和q_i分别为聚类结果和真实标签。ACC可以定义为

$$\mathrm{ACC} = \frac{1}{n}\sum_{i=1}^{n}\delta(q_i, map(p_i)), \tag{2.77}$$

其中n是样本总数，$\delta(x, y)$函数等于1如果$x = y$等于0如果$x \neq y$。而$map(\cdot)$是置换映射函数将簇标签映射到类标签。最佳映射可以通过Kuhn-Munkres算法[73]来获取。ACC是介于$0 \sim 1$之间的值，ACC的值越大说明聚类效果越好。

归一化互信息(normalized mutual information，简称NMI) NMI 是一种外部评价标准，它用来评价算法在一个数据集上的聚类结果与该数据集真实划分的相似程度。用$\mathcal{C}$表示真实标签中类的集合，用$\mathcal{C}'$表示聚类算法获得的簇的集合。它们的互信息定义为

$$MI(\mathcal{C}, \mathcal{C}') = \sum_{c_i \in \mathcal{C}, c'_j \in \mathcal{C}'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}, \tag{2.78}$$

其中$p(c_i)$和$p(c'_j)$分别是样本属于类c_i和簇c'_j的概率，而$p(c_i, c'_j)$是样本同时属于类c_i和簇c'_j的联合概率。实验中我们使用下面的归一化互信息

$$NMI(\mathcal{C}, \mathcal{C}') = \frac{MI(\mathcal{C}, \mathcal{C}')}{\max(H(\mathcal{C}), H(\mathcal{C}'))}, \tag{2.79}$$

其中$H(\mathcal{C})$和$H(\mathcal{C}')$分别是类$\mathcal{C}$和簇$\mathcal{C}'$对应的信息熵。容易验证NMI是介于0 $\sim$ 1之间的值，$NMI = 0$如果$\mathcal{C}$和$\mathcal{C}'$互相独立，$NMI = 1$如果$\mathcal{C}$和$\mathcal{C}'$相同，NMI的值越大说明聚类效果越好。

2.4.3　数据集

表2.2总结了所用数据集的统计特征。数据集的详细描述如下：

- COIL20包含1440张黑色背景的灰度图，这些图像属于20类而每一类包含72张图片。每张图片被处理成32 $\times$ 32像素。
- JAFFE包含213张由10个日本女性提供的7种面部表情（6种基本面部表情和中性）。每张图片被处理成26 $\times$ 26像素。
- CSTR是罗切斯特大学计算机科学系从1991年到2002年发表的技术报告的摘要。该数据集包含476篇摘要，其中分为四个研究领域：自然语言处理（NLP），机器人/视觉，系统和理论。我们通过互信息选择了前1000个单词，该数据集被广泛用于矩阵分解算法的评测[38]。
- WebKB包含了约6000个从四所高校（康奈尔大学，德克萨斯州，华盛顿州，威斯康星州）的计算机科学部门收集的网页。每个网页都标有一个标签：学生，教授，课程，项目，人员，部门，以及其他。我们的实验中使用康奈尔大学的子集。

2.4.4　实验结果

我们首先在无明显噪声的数据集（即数据集通常用于评测非鲁棒学习算法的评测）上对比这些算法的聚类效果，试验结果表明鲁棒算法通常优于标准的非负矩阵分解算法，说明噪声普遍存在于这些数据集中。针对图像遮挡这一特定噪声，我们分别对比了不同算法在含有不同数量的噪声数据上的聚类和分类结果，充分说明本章提出的算法能够有效提高非负矩阵分解模型的鲁棒性。

表 2.2: 第二章实验所使用的数据集

名称	样本数	类别数	特征数
COIL20	1440	1024	20
JAFFE	213	676	10
CSTR	476	1000	4
WebKB	827	4134	7

2.4.4.1 无明显噪声数据集上的实验结果

表2.3和表2.4分别列出了聚类算法在这些数据集上的准确性和归一化互信息。最好的两个结果加黑显示。基于这些实验结果我们观测到：1）加权的非负矩阵分解变种，也就是最后六个算法通常比标准的非负矩阵分解效果好。这些算法都降低了大的拟合误差对应行、列或元素的权重。这也表明，尽管这些数据集通常用于测试非鲁棒算法的结果，精化的权重方案仍然能够提高无权重非负矩阵分解的结果。2）本章提出的算法，即CIM-NMF，rCIM-NMF和Huber NMF在这些数据集上结果通常优于现有的算法。对于图像数据集CIM-NMF和Huber-NMF结果比较好，一个较为合理的解释是噪声分散在整个矩阵；而rCIM-NMF在文本数据集上结果较好，这点与文献[106]中使用的NCW权重方法(normalized cut weighting scheme，简称NCW)类似。

2.4.4.2 遮挡人脸数据集上的实验结果

在这一小节，我们主要通过对比算法在包含噪点数据集上的实验结果来测试算法的鲁棒性。在本实验中我们以ORL人脸数据集为例，该数据集包含40个人对应的400张灰度图。这些图片在不同时间，不同光照条件，不同面部表情以及有无戴眼镜等环境下生成。所有的图片通过手工方式对齐和裁剪。裁剪后的图像由32×32像素描述，每个像素有256灰度值。因此，每幅图片可以通过一个1024维的向量来表示。

为了模拟噪点的产生过程，根据文献[109]中的方法，我们随机选取不同比例($r = 5\%, 10\%, \ldots, 50\%$)的图像，并且遮挡其部分关键脸部特征（如眼睛和嘴巴）。为了减小统计误差，对于不同的噪声比例我们独立的重复20次实验，然后报告在这20次随机实验中的平均性能和标准差。

聚类结果 表2.5、2.6和表2.7、2.8分别列出了不同算法在带有不同比例噪

表 2.3: 第二章的聚类准确性

Data set	COIL20	JAFFE	CSTR	WebKB
Kmeans	0.631	0.657	0.727	0.520
PCA-Km	0.638	0.780	0.749	0.527
RPCA-Km	0.652	0.753	0.703	0.615
Ncut	0.380	0.795	0.714	0.430
NMF	0.651	0.861	0.758	0.557
SR-NMF	**0.671**	0.839	0.777	0.603
$L_{2,1}$-NMF	0.658	0.893	0.771	0.614
WFS-NMF	0.649	0.879	**0.784**	**0.664**
Huber-NMF	0.661	**0.928**	0.765	0.617
rCIM-NMF	**0.695**	0.882	**0.798**	**0.675**
CIM-NMF	0.670	**0.927**	0.761	0.652

表 2.4: 第二章的归一化互信息

Data set	COIL20	JAFFE	CSTR	WebKB
Kmeans	0.743	0.745	0.651	0.042
PCA-Km	0.743	0.821	0.663	0.056
RPCA-Km	0.755	0.831	0.603	0.022
Ncut	0.578	0.833	0.638	0.151
NMF	0.679	0.859	0.665	0.155
SR-NMF	**0.758**	0.862	0.687	**0.176**
$L_{2,1}$-NMF	0.713	0.908	0.681	0.157
WFS-NMF	0.740	0.872	**0.687**	0.013
Huber-NMF	0.743	**0.932**	0.675	0.172
rCIM-NMF	**0.755**	0.897	**0.691**	**0.177**
CIM-NMF	0.753	**0.942**	0.668	0.153

表 2.5: 第二章的ORL人脸数据集上聚类准确性(均值% ± 方差)

$r(\%)$	Kmeans	PCA-Km	RPCA-Km	Ncut	NMF
5	51.1±2.9	50.6±1.2	52.1±2.5	44.2±2.9	54.7±3.2
10	48.7±2.5	48.4±2.6	48.8±1.9	34.6±1.5	51.5±3.0
15	45.5±2.6	45.4±2.0	45.7±2.0	33.8±2.4	49.6±1.7
20	43.3±2.4	43.7±2.4	43.8±2.6	35.0±2.2	45.9±2.5
25	40.5±2.3	41.5±2.3	41.5±1.7	35.4±2.0	44.2±2.8
30	39.1±2.2	39.8±1.7	39.2±1.7	35.1±1.6	42.2±2.8
35	37.1±1.9	37.3±1.7	38.1±2.1	34.4±1.8	38.7±1.9
40	34.6±1.6	35.0±1.7	35.8±1.7	32.8±1.5	36.9±2.2
45	34.1±1.5	34.0±1.4	34.5±1.5	33.3±1.3	36.2±2.1
50	32.1±1.6	32.8±1.5	33.2±1.3	32.3±1.7	34.4±1.8
Avg.	40.6	40.8	41.3	35.1	43.4

点的ORL人脸数据集上的聚类准确性和归一化互信息结果。可以看出，本章提出的算法如CIM-NMF在所有比例上都显著优于其他对比算法。和第二好算法相比，CIM-NMF在聚类准确性指标上相对提高了12.1%，而在归一化互信息上也达到了提高8.5%的提升。对于所有对比的算法，他们的聚类性能都随着噪点比例的增加而降低。有意思的是当噪点比例较低时（如$r = 5\%, 10\%, 15\%$），我们观测到：1）CIM-NMF的结果相对稳定而其他算法性能下降很快；2）所有非负矩阵分解变种都优于标准模型，这也说明这些模型在某种程度上对噪点鲁棒；3）CIM-NMF在这些比例上效果比较接近，这也说明其对噪点的鲁棒性最好。当r继续增大时CIM-NMF仍然表现最好，而其他算法的性能相对接近于标准的非负矩阵分解。所有这些结果都表明CIM-NMF对较大范围的噪点鲁棒性更好。

分类结果 为了进一步验证算法在受破坏数据集上的性能，我们在算法学习到的低维数据表示上利用最近邻分类来测试算法的鲁棒性。对于不同比例的噪声，我们从每一类随机选择3个样本作为训练集剩余的样本用作测试，并且独立重复50次训练集/测试集划分实验，这样我们总共生成和记录了$10 \times 20 \times 50$次实验结果。

表2.9和2.10列出这些算法分类准确性的具体结果。这里我们用基于所有特

表 2.6: 第二章的ORL人脸数据集上聚类准确性-续(均值% ± 方差)

r(%)	SR-NMF	$L_{2,1}$-NMF	WFS-NMF	Huber-NMF	rCIM-NMF	CIM-NMF
5	58.2±2.1	56.0±3.4	56.5±3.0	57.1±2.9	56.7±3.3	**61.0±2.1**
10	54.3±2.7	52.9±2.8	53.6±2.7	55.4±2.2	53.9±2.7	**60.7±3.5**
15	50.5±3.0	49.9±2.6	50.5±2.5	51.8±3.0	50.6±2.8	**58.8±3.3**
20	48.0±2.2	47.1±2.5	47.8±2.1	50.2±2.5	47.8±2.3	**57.2±2.6**
25	46.1±1.8	44.9±2.7	45.5±3.4	47.6±1.8	44.6±2.5	**54.8±3.5**
30	42.7±2.7	42.5±2.2	42.1±2.8	44.5±2.7	42.1±2.5	**51.6±3.2**
35	40.6±2.0	39.4±2.1	39.6±2.4	41.3±2.3	40.6±2.0	**47.1±2.4**
40	38.0±2.0	37.1±1.7	38.7±2.5	38.4±1.6	37.0±2.1	**43.0±2.7**
45	37.1±1.3	37.9±2.5	36.1±3.1	37.6±2.2	36.8±1.7	**42.0±2.8**
50	35.4±2.0	35.0±2.1	35.4±2.4	35.9±2.0	35.0±2.1	**39.7±2.1**
Avg.	45.1	44.3	44.6	46.0	44.5	**51.6**

表 2.7: 第二章的ORL人脸数据集上归一化互信息(均值% ± 方差)

r(%)	Kmeans	PCA-Km	RPCA-Km	Ncut	NMF
5	70.8±1.4	70.6±1.1	72.0±1.2	64.4±2.9	72.1±1.5
10	68.4±1.1	67.9±1.5	69.3±1.0	54.9±1.6	69.5±1.5
15	65.5±1.5	65.4±1.3	66.7±0.9	54.8±2.6	67.7±1.2
20	63.8±1.3	64.0±1.4	64.6±1.3	56.5±2.0	65.3±1.7
25	61.4±1.4	62.0±1.1	63.1±1.3	57.8±2.0	63.4±1.8
30	59.8±1.6	60.4±1.0	61.2±1.0	57.3±1.8	61.4±2.1
35	57.8±1.2	58.6±1.4	60.0±1.2	56.5±1.9	59.5±1.3
40	56.6±1.0	56.8±1.3	57.9±1.2	55.7±1.5	57.5±1.3
45	56.2±1.6	56.1±1.1	57.0±1.4	56.3±1.1	57.2±1.3
50	54.4±1.4	55.1±1.0	56.0±1.3	55.2±1.6	55.8±1.7
Avg.	61.5	61.7	62.8	56.9	63.0

表 2.8: 第二章的ORL人脸数据集上归一化互信息-续(均值% ± 方差)

$r(\%)$	SR-NMF	$L_{2,1}$-NMF	WFS-NMF	Huber-NMF	rCIM-NMF	CIM-NMF
5	74.8±1.0	73.4±1.9	73.6±2.2	74.3±2.0	74.2±1.9	**77.6±1.6**
10	71.9±1.5	70.7±1.5	71.0±1.8	72.4±1.2	70.9±2.0	**77.4±1.6**
15	68.3±1.3	68.1±1.6	68.6±1.8	69.6±1.4	68.2±1.8	**75.7±2.1**
20	66.6±1.7	65.5±1.4	66.1±1.3	67.0±1.4	66.6±1.5	**74.4±1.4**
25	65.0±1.4	63.9±1.6	64.0±1.6	64.9±1.5	63.5±1.8	**72.1±1.9**
30	62.2±1.8	61.7±1.5	62.1±1.5	62.8±1.6	61.6±1.3	**69.9±2.3**
35	60.7±1.5	59.8±1.8	59.3±1.4	60.0±1.9	60.9±1.5	**66.6±1.7**
40	59.0±1.5	58.0±1.0	58.8±1.2	59.1±1.6	57.5±1.3	**63.8±2.0**
45	58.1±1.0	58.1±1.3	58.1±1.4	58.7±1.5	57.6±1.4	**62.7±1.5**
50	56.9±1.5	56.4±1.3	56.6±1.1	56.5±1.6	56.9±1.4	**60.1±1.3**
Avg.	64.4	63.5	63.9	64.5	63.8	**70.0**

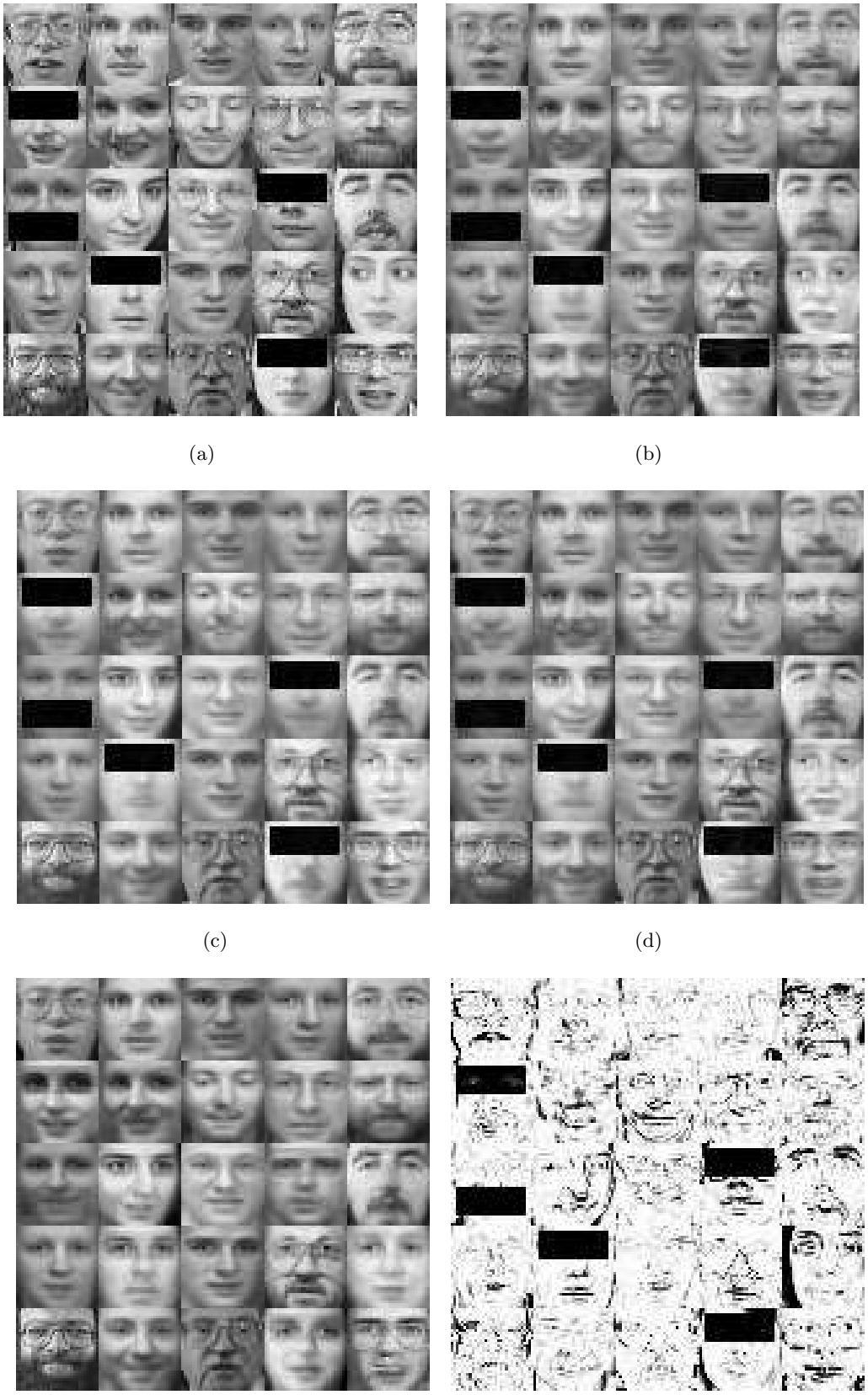
征分类的方法作为一个弱基准算法。NCut可以看作基于图的嵌入式表达。同样CIM-NMF在各种噪点比例下都显著优于其他算法，并且与第二好算法相比在分类准确性上达到了15.4%的相对提高。

重构数据可视化 为了更好的理解本章提出的算法，我们在20%的噪点比例上随机选取25张图片，并且在图2.4显示出多种非负矩阵分解变种对数据矩阵的重构结果，以及CIM-NMF的权重矩阵。可以看出和其他算法相比CIM-NMF重构的数据矩阵更加清楚，并且CIM-NMF得到的权重矩阵的确对于明显的噪点赋予了较小的权重。而SR-NMF和L_{21}-NMF很难区分出这些噪点。

基向量可视化 在本实验中我们在$r = 20\%$的样本上随机注入噪点，并且在图2.5中显示了NMF和CIM-NMF获得的基向量。可以看出NMF中10/40个基向量被遮挡而CIM-NMF中仅有4/40个被噪点破坏。通过对比我们可以看出本章提出的算法获得的基向量由于受噪点影响较小而更加清楚。

2.5 本章小结

在本章中，我们提出了基于半二次最小化的鲁棒非负矩阵分解方法。由于采用的平方误差在大的拟合误差上损失较大以及高斯分布过多的强调数据应集中于均值附近，标准非负矩阵分解对大的、非高斯噪声敏感。为了提高非负矩



(a)　(b)

(c)　(d)

(e)　(f)

图 2.4: (a) 被破坏的原始图像，(b)、(c)、(d)、(e) 通过NMF、SR-NMF、$L_{2,1}$-NMF和CIM-NMF重构得到的图像，(f)CIM-NMF学习得到的权重矩阵

(a)

(b)

图 2.5: (a)NMF学到的基向量，(b)CIM-NMF学到的基向量

表 2.9: 第二章的ORL数据集上分类准确性(均值% ± 标准差)

r(%)	All Features	PCA	RPCA	Ncut	NMF
5	73.3±2.4	73.6±2.3	73.4±2.4	69.9±2.7	73.4±2.2
10	68.8±2.5	69.5±2.4	68.9±2.5	56.9±2.8	70.2±2.5
15	64.6±2.6	65.3±2.7	64.6±2.6	54.0±2.8	65.5±2.6
20	60.8±2.8	61.1±2.6	60.8±2.7	51.4±3.3	61.3±2.7
25	58.1±2.7	58.7±2.8	57.9±2.7	50.0±3.0	59.1±2.7
30	55.0±2.9	55.3±3.0	54.8±2.8	48.1±2.9	55.0±3.1
35	52.2±2.9	52.3±3.1	51.8±2.9	45.9±3.2	52.1±3.3
40	49.3±2.9	49.7±3.0	48.7±2.9	43.8±3.3	49.3±3.1
45	48.2±3.0	48.4±3.0	47.7±3.0	43.8±3.0	47.5±3.2
50	45.7±3.1	45.6±3.1	45.1±3.1	41.9±3.0	44.7±3.1
Avg.	57.6	58.0	57.4	50.6	57.8

表 2.10: 第二章的ORL数据集上分类准确性-续(均值% ± 标准差)

r(%)	SR-NMF	$L_{2,1}$-NMF	WFS-NMF	Huber-NMF	rCIM-NMF	CIM-NMF
5	77.2±2.3	75.8±2.3	75.4±2.6	77.8±2.4	75.4±2.5	**81.1±2.5**
10	72.5±2.4	71.3±2.5	71.9±2.7	73.0±2.6	70.7±2.6	**81.0±2.4**
15	67.8±2.6	67.0±2.5	66.8±2.5	67.8±2.6	66.2±2.7	**79.2±2.8**
20	63.1±2.6	62.2±2.7	62.4±2.4	63.3±2.7	61.9±2.6	**76.2±2.8**
25	60.0±2.8	59.1±2.8	59.7±2.6	60.6±2.9	58.5±2.7	**72.6±2.9**
30	56.2±3.2	55.4±3.1	55.8±2.1	56.8±2.9	54.5±3.4	**68.3±3.1**
35	53.0±3.0	52.2±3.0	52.3±2.6	53.1±3.1	51.9±2.9	**63.1±3.2**
40	50.1±2.9	49.2±3.0	49.6±2.0	50.1±3.3	48.6±3.1	**59.3±3.2**
45	48.1±3.1	47.5±3.4	48.2±2.2	48.7±3.2	47.5±3.0	**56.0±3.3**
50	45.8±3.2	45.0±3.2	45.4±2.8	46.0±3.3	44.4±3.1	**51.7±3.4**
Avg.	59.4	58.5	58.8	59.7	58.0	**68.9**

阵分解的鲁棒性，我们提出采用对大拟合误差比平方误差损失不敏感的损失函数测度数据重构的质量以及采用在长尾数据概率较高的分布刻画数据的生成过程。本章中我们采用可以被半二次优化求解的一类损失函数，即半二次损失函数，来度量矩阵分解的质量。我们分别提出基于半二次优化乘法形式和加法形式的通用求解算法来求解采用那个不同半二次损失函数的鲁棒非负矩阵分解问题。在乘法形式迭代中，我们不断的减小大拟合误差的权重，降低异常点对整个优化问题的影响，从而提高模型的鲁棒性。在加法形式的迭代中，我们不断的分离和估计噪声矩阵并恢复干净的数据矩阵，进而提高模型的鲁棒性。两种通用求解算法直观的解释了基于半二次损失的非负矩阵分解模型的鲁棒性。我

们还分析了现有鲁棒非负矩阵分解算法和本章方法之间的关系。噪声数据集上的聚类实验结果表明，鲁棒非负矩阵分解算法通常优于标准的非负矩阵分解算法，不同的鲁棒算法在不同数据集表现并不完全一致，这也说明了设计通用求解算法的必要性。

此章提出的模型和算法有很多未来的工作可以继续展开，以下列出一些可以展开的方向：

- 第一，提高鲁棒非负矩阵分解的实用性。我们希望能够进一步从半二次优化的角度分析采用其他损失函数（如Bregman失真度，alpha-beta失真度等）的非负矩阵分解模型，从而扩展鲁棒非负矩阵分解的范围。此外由于这些损失函数通常带有一个或两个超参数，鲁棒非负矩阵分解同时也要求模型需要对这些参数的鲁棒性较好，如何分析这些参数对模型的影响并设计有效的参数选择或者自适应方法也是很重要的问题。我们也希望能够利用一些容易获取的领域知识，针对实际问题和数据集快速准确的选择一个合适的鲁棒目标函数。通过这几个问题的研究我们一方面拓宽了可供选择的损失函数和鲁棒的非负矩阵分解模型，另一方面又提高了单个模型对相关参数的稳定性，最后还提高了鲁棒损失函数的选择能力，从而提高鲁棒非负矩阵分解的实际应用价值。

- 第二，提高鲁棒非负矩阵分解局部噪声的识别和修复能力。尽管本章提出的方法比标准非负矩阵分解对大的、非高斯噪声鲁棒性要好，这些鲁棒模型都对噪声的检测和修复都建立在所有数据的基础上（如乘法形式的绝对权重和加法形式的噪声仅依赖于矩阵每个元素拟合误差本身，而相对权重取决于所有元素）。而事实上，噪声往往带有一定的局部性（如图像中的某个像素被破坏，附近的像素也有很大的可能是噪声）。因此我们可以研究如何利用数据间的关系（如特征的空间位置关系）设计鲁棒损失函数超参数设置方法，进一步提高非负矩阵分解的鲁棒性。

- 第三，研究基于半二次最小化的稀疏鲁棒非负矩阵分解以提高鲁棒非负矩阵分解模型从噪声数据中抽取基于局部特征的数据表示（Part-based representation）的能力以及提高鲁棒非负矩阵分解模型最优解的唯一性。标准非负矩阵分解通常通过非负因子的额外约束或正则（如基于$L1$函数的稀疏性正则）来提高分解的稀疏性，而$L1$函数可以看成一种特殊的半

二次损失，因此我们可以进一步考虑采用半二次损失函数分别度量重构误差和非负因子的稀疏性，从而得到稀疏化的鲁棒非负矩阵分解。

第三章　基于鲁棒非负矩阵分解的联合聚类

3.1　引言

在第1.2节，我们介绍了联合聚类的基本思想和当前的主要算法，这里我们详细阐述基于非负矩阵分解的联合聚类算法。和聚类问题类似，样本与特征间的关系对联合聚类也非常重要。Ding等人在文献[29]中提出双向正交的非负矩阵三因子分解模型用于刻画非负因子对样本与特征间关系的近似程度。近年来，大量研究表明样本间和特征间的关联关系对联合聚类也非常重要[37, 113]。一种常用的无监督关联关系是用于刻画样本流形和特征流形几何结构的关系图。Gu和Zhang等人在文献[37, 90]中通过对样本因子和特征因子分别施加图正则提出对偶图正则的（半监督）非负矩阵三因子分解（Dual graph regularized (semi)-nonnegative matrix tri-factorization，简称GNMTF）的联合聚类算法。文献[36]进一步提出通过类似于标准化割（normalized cut）的约束条件使得GNMTF模型是定义良好的，从而避免了GNMTF模型存在的平凡解和尺度变换的问题。文献[102]通过约束样本因子和特征因子为指示矩阵，提出GNMTF模型的快速求解算法。此外，GNMTF模型也被用于求解其他方面的联合聚类问题（如基于高阶、多关系的联合聚类[61, 62, 101]，协同过滤[40]，迁移学习[72, 119]等）。

尽管GNMTF模型受到了广泛关注，这些模型在实际应用中受到了噪声数据带来的挑战。由于GNMTF在数据重构部分通常采用$L2$的平方误差，大的、非高斯噪声产生的大拟合误差可能会支配整个重构误差，也就是说GNMTF的数据重构对大噪声敏感。除了数据重构，样本之间的关系和特征之间的关系，如GNMTF(3.2)中的对偶图正则，对聚类任务也是至关重要的[113]。然而，当这些数据矩阵被噪音和异常点破坏后，构造的图也可能受到污染而无法可靠的表示理想可靠的关联关系。比如，在一个噪音数据集上构造的K近邻关系有可能将两个样本错误的关联起来，两个样本之间的权重关系有可能是不合适的。因此，这些图正则可能会误导矩阵分解模型降低联合聚类效果。

表 3.1: 第三章符号概要

符号	描述
n	输入数据点的数目
d	数据点的维度
X	输入非负数据矩阵，$X \in \mathbb{R}_+^{d\times n}$
k_1	特征簇的数目
k_2	数据簇的数目
$X_{i\cdot}$	X的第i行
$X_{\cdot j}$	X的第j列
$\mathbb{R}_+$	非负实数的集合
F	特征划分矩阵$F \in \mathbb{R}^{d\times k_1}$
H	特征簇与样本簇的关联矩阵$H \in \mathbb{R}^{k_1\times k_2}$
G	数据划分矩阵$G \in \mathbb{R}^{n\times k_2}$
W_F	特征邻接矩阵$W_F \in \mathbb{R}^{d\times d}$
D_F	特征邻接矩阵W_F对应的度矩阵，定义为一个对角矩阵，其对角线上的元素是矩阵W_F每一行的和，即$(D_F)_{ii} = \sum_{i'}(W_F)_{ii'}$
W_G	样本邻接矩阵$W_G \in \mathbb{R}^{n\times n}$
D_G	样本邻接矩阵W_G对应的度矩阵，且$(D_G)_{jj} = \sum_{j'}(W_G)_{jj'}$

3.2 预备知识

我们简单介绍本章使用的符号习惯。对于方阵$A \in \mathbb{R}^{d\times d}$，$\mathrm{tr}(A)$代表矩阵$A$的迹，即对角线上元素之和。$\odot$表示element-wise乘法运算而$\oslash$表示element-wise除法运算. 我们还定义$\mathbf{1}_d \in \mathbb{R}^d$代表每个元素都是1的列向量。表3.2总结了其他重要的符号和标记。

给定一个非负数据矩阵$X \in \mathbb{R}^{d\times n}$, 矩阵的行对应于特征而列表示样本。联合聚类的目标是将样本将特征$\{X_{1\cdot}, \ldots, X_{d\cdot}\}$聚到$k_1$簇$\{\mathcal{C}'_i\}_{i=1}^{k_1}$，同时将样本$\{X_{\cdot 1}, \ldots, X_{\cdot n}\}$ 聚到k_2个簇$\{\mathcal{C}_i\}_{i=1}^{k_2}$。

3.2.1 非负矩阵三因子分解

在非负矩阵三因子分解中，反映特征与样本之间关系的原始数据矩阵被近似为三个非负因子的乘积，如$X \approx FHG^T$其中F和G分别是特征和样本的簇

指示矩阵，而H是特征簇和样本簇的关联矩阵。如果不使用额外的约束，三因子分解问题等价于两个因子的非负矩阵分解问题[29]。因此，Ding等人在文献[29]中提出了正交约束的非负矩阵三因子分解模型：

$$\min ||X - FHG^T||^2 \text{s.t.} F \geq 0, H \geq 0, G \geq 0, F^T F = I, G^T G = I \tag{3.1}$$

Zhang等人提出有界约束和稀疏性约束的非负矩阵三因子分解模型用于社交网络中的社区检测问题。[116]。

3.2.2　图正则非负矩阵三因子分解

为了同时对样本和特征聚类，Gu[37]和Wang[90]等人基于流行假设[8]同时利用样本图正则和对偶特征图正则，提出图正则的非负矩阵三因子分解GNMTF：

$$\mathcal{J} = \sum_{i=1}^{d}\sum_{j=1}^{n}(X_{ij} - (FHG^T)_{ij})^2 + \lambda_F \sum_{i=1}^{d}\sum_{i'=1}^{d} W_{ii'}^F ||F_{i\cdot} - F_{i\cdot}||^2 + \lambda_G \sum_{j=1}^{n}\sum_{j'=1}^{n} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||^2 \tag{3.2}$$

$$= ||X - FHG^T||_F^2 + \lambda_F \text{tr}(F^T L_F F) + \lambda_G \text{tr}(G^T L_G G) \tag{3.3}$$

$$\text{s.t.} \quad F \geq 0, H \geq 0, G \geq 0,$$

其中F和G分别是特征和样本的簇指示矩阵，H反映了样本簇和特征簇之间的关联关系[119]，$L_F = D_F - W_F$是特征图上的图拉普拉斯，而D_F和W_F是特征图的度矩阵和邻接矩阵，样本图Laplacian L_G的定义和特征图类似，λ_F和λ_G是正则化参数。

3.3　鲁棒联合聚类算法

3.3.1　问题形式化

受鲁棒主成份分析（Robust PCA）[17]研究的启发，我们提出一种新的非负三因子分解模型，其中我们假设矩阵的某些元可能被任意的破坏，但这些破坏是稀疏的。我们引入一个噪声矩阵$S \in \mathbb{R}^{d \times n}$来显式的捕捉这些稀疏噪声。因此，为了近似非负数据矩阵X，鲁棒三因子分解的目标可以写成

$$X = FHG^T + S,$$

其中要求$F \in \mathbb{R}^{d\times k_1}, H \in \mathbb{R}^{k_1\times k_2}$和$G \in \mathbb{R}^{n\times k_2}$都是非负矩阵，要求$S$是一个稀疏矩阵。由于$S$的存在，我们可以从稀疏噪声中恢复一个更为真实的数据矩阵，因此可以预期获得一个更正确的分解结果。此外，我们可以通过最小化$X - FHG^T - S$的平方误差来近似处理数据中密集的，数值较小的高斯噪声。

考虑到在图正则中两个本来距离较远的点由于受噪声的影响可能被强制要求比较相似，我们进一步提出采用最小化ℓ_1范数的图正则，它能压制大的、意外的正则错误，在后面的求解算法过程中，我们还发现ℓ_1范数的图正则可以估计更加准确的结构（如样本间的相似性）；另外ℓ_1正则会产生稀疏化的结果，这就使得很多邻近关系样本或特征间的距离接近零，从而产生更紧凑的分解结果，提高联合聚类的效果。

将鲁棒数据重构和鲁棒对偶图正则结合到一个统一的框架，我们提出基于非负矩阵分解的鲁棒联合聚类（Robust Co-Clustering，简称RCC），RCC可以形式化为以下优化问题：

$$\begin{aligned}\mathcal{J}_1 &= ||X - FHG^T - S||_F^2 + \lambda_S||S||_1 \\ &+ \lambda_F \sum_{ii'} W_{ii'}^F||F_{i\cdot} - F_{i'\cdot}||_2 + \lambda_G \sum_{jj'} W_{jj'}^G||G_{j\cdot} - G_{j'\cdot}||_2 \\ \text{s.t.}\quad & F, H, G \geq 0, F\mathbf{1}_{k1} = \mathbf{1}_d, G\mathbf{1}_{k2} = \mathbf{1}_n,\end{aligned} \tag{3.4}$$

其中λ_S、λ_F 和λ_G是正则参数。通过对F和G的每一行增加ℓ_1规范化约束，以及在后面两项上使用ℓ_1范数的误差函数，式子(3.4)中的优化问题是定义良好的并且不再存在GNTMF[36]的尺度转移问题和平凡解的问题。

3.3.2 学习算法

式子(3.4)对于所有变量是个非凸的优化问题，但分别对于单个变量是一个凸问题。下面，我们介绍基于分块坐标下降的迭代算法来最小化式子(3.4)中的目标函数。通过固定其他变量，我们分别跟新S, F, H,和G的值。而通过一序列凸优化问题，我们可以获得关于原问题的一个局部最优解，我们也证明了算法的收敛性。

3.3.2.1　计算S

式子(3.4)关于S的优化问题等价于最小化

$$\mathcal{J}_2 = ||X - FHG^T - S||_F^2 + \lambda_S||S||_1. \tag{3.5}$$

容易验证上面的问题对S的每个元素都是独立的，可以分解成$d \times n$个子问题，并且每个子问题的最优解可以通过所谓的soft-thresholding operator[4]获取，即

$$S_{ij} = \begin{cases} 0 & \text{if } |E_{ij}| \leq \lambda_S/2, \\ E_{ij} - \frac{\lambda_S}{2}\text{sign}(E_{ij}) & \text{otherwise}, \end{cases} \tag{3.6}$$

其中$E_{ij} = (X - FHG^T)_{ij}$。将式子(3.6) 代入式子(3.5)可得

$$\min \mathcal{J}_2 = \begin{cases} E_{ij}^2 & \text{if } |E_{ij}| \leq \lambda_S/2, \\ \lambda_S|E_{ij}| - (\frac{\lambda_S}{2})^2 & \text{otherwise.} \end{cases} \tag{3.7}$$

值得注意的是式子(3.7)在鲁棒统计[42]中也被称为Huber M-估计量，因此不同于式子(3.5)中对噪声S的稀疏性假设（即$|S|_1$是对$|S|_0$的凸近似），最小化基于稀疏性噪声的重构误差可以看成最小化基于半二次优化加法形式的Huber损失。根据这一对偶关系，可以得出RCC算法的鲁棒性不仅仅来自于对稀疏噪声的处理，还可能由于soft-thresholding operator的作用和鲁棒Huber估计量类似。根据这一对偶关系可以看出：1）如果$\lambda_S > 2\max_{ij}|E_{ij}|$，$S$的所有元素变成0，式子(3.5)变成基于F-范数（即平方误差和）的非负三因子分解；2）如果$\lambda_S \to 0$，式子(3.5)类似与误差绝对值之和；3）通过调节权衡参数λ_S，式子(3.5)针对较小的误差（$E_{ij} \leq \lambda_S/2$）采用ℓ_2-范数，而在较大的误差上使用ℓ_1-范数。此外，实验时我们可以根据式子(3.7)和Huber M-估计量的关系指导正则参数λ_S的选择，如设置为拟合误差E的分位数，即$\lambda_S = \text{quantile}(E, r)$，其中$r = \{0.1, \ldots, 0.9\}$。

当前基于非负矩阵三因子分解的方法[29, 36, 37, 90]都采用平方误差测度矩阵所有元素的重构误差，而我们的方法通过权衡参数自动在大误差上使用ℓ_1范数（即绝对值之和）。由于噪音和异常点重构时通常产生较大的拟合误差，我们的方法通过减小大误差减轻了噪音对整个优化问题的影响，提高了三因子分解中数据重构的鲁棒性。

3.3.2.2 计算F

式子(3.4)关于F的优化问题等价于最小化：

$$\begin{aligned}\mathcal{J}_3 &= ||X - FHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2 \\ \text{s.t.} \quad & F \geq 0, F\mathbf{1}_{k_1} = \mathbf{1}_d.\end{aligned} \tag{3.8}$$

由于式子(3.8)中的图正则项采用了ℓ_1-范数，这造成了目标函数不可微，再加上F每行的规范化约束使得问题看起来很难优化。

虽然基于ℓ_1-范数的图正则是凸优化问题，然而，当$||F_{i\cdot} - F_{i'\cdot}||_2 = 0$时其导数不存在，当$||F_{i\cdot} - F_{i'\cdot}||_2 \neq 0$时其导数为

$$\frac{\partial \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2}{\partial F} = \sum_{ii'} \frac{W_{ii'}^F}{2||F_{i\cdot} - F_{i'\cdot}||_2} \frac{\partial ||F_{i\cdot} - F_{i'\cdot}||_2^2}{\partial F} \tag{3.9}$$

引入$\widetilde{W}_{ii'}^F = \frac{W_{ii'}^F}{2||F_{i\cdot} - F_{i'\cdot}||_2}$, $\widetilde{D}_{ii}^F = \sum_{i'} \widetilde{W}_{ii'}^F$，可得关于$F$的辅助函数

$$\mathcal{L}(F) = \text{tr}(\sum_{ii'} \widetilde{W}_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2^2) = \text{tr}(F^T \widetilde{D}^F F - F^T \widetilde{W}^F F), \tag{3.10}$$

可以看出式子(3.9)同样是函数$\mathcal{L}(F)$关于F的导数。因此，我们可以将式子(3.8)中的优化问题重写为

$$\begin{aligned}\mathcal{J}_4 &= \text{tr}(-2F^T P^+ + 2F^T P^- + FQF^T + \lambda_F F^T \widetilde{D}^F F - \lambda_F F^T \widetilde{W}^F F) \\ \text{s.t.} \quad & F \geq 0, F\mathbf{1}_{k_1} = \mathbf{1}_d,\end{aligned} \tag{3.11}$$

其中$P = (X - S)GH^T$，$Q = HG^T GH^T$并且我们引入$P_{ij}^+ = (|P_{ij}| + P_{ij})/2$，$P_{ij}^- = (|P_{ij}| - P_{ij})/2$。

需要指出的是，尽管文献[72, 119]已经提出在非负矩阵三因子分解中利用非负因子F和G的归一化约束得到对应的后验概率解释。然而，为了求解归一化约束的优化问题，当前方法分为两步，即首先求解无归一化约束的非负矩阵三因子分解问题，然后通过归一化投影得到满足约束的结果。这样得到的解虽然满足归一化约束，但是额外的归一化投影无法保证目标函数在归一化前后保持不变或者单调下降，因此归一化投影的可能会造成性能的下降[94]。

为了更好的处理归一化约束，我们引入一个非负因子矩阵$\widetilde{F} \in \mathbb{R}^{d \times k_1}$作为辅助变量，而满足行归一化约束的$F$可以通过$F_{ik} = \frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}$计算。因此对式子

（3.11）满足归一化约束F的优化问题可以通过最小化下面关于$\widetilde{F}$的目标函数获取

$$\mathcal{J}_5(\widetilde{F}) = -2\sum_{ik} P^+_{ik}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} + 2\sum_{ik} P^-_{ik}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}} + \sum_{ikk'} Q_{kk'}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}} + \lambda_F\sum_{ii'k}(\widetilde{L}_F)^+_{ii'}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}\frac{\widetilde{F}_{i'k}}{\sum_s \widetilde{F}_{is}} - \lambda_F\sum_{ii'k}(\widetilde{L}_F)^-_{ii'}\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}\frac{\widetilde{F}_{i'k}}{\sum_s \widetilde{F}_{is}}, \tag{3.12}$$

其中通过$\widetilde{F}$得到的F总是满足归一化约束。

我们可以为式子(3.12)构造下面的辅助函数[37]

$$\begin{aligned} Z(\widetilde{F},\widetilde{F}^t) = &-2\sum_{ik} P^+_{ik}(\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is})(1+\log\frac{\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is}}{\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is}}) \\ &+ 2\sum_{ik} P^-_{ik}\frac{(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})^2 + (\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is})^2}{2\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is}} \\ &+ \sum_{ik}\frac{[(\widetilde{F}^t \oslash (\widetilde{F}\mathbf{1}_{k_1}\mathbf{1}^T_{k_1}))Q]_{ik}(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})^2}{\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is}} \\ &+ \lambda_F\sum_{ik}\frac{[\widetilde{D}^F(\widetilde{F}^t \oslash (\widetilde{F}\mathbf{1}_{k_1}\mathbf{1}^T_{k_1}))]_{ik}(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})^2}{\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is}} \\ &- \lambda_F\sum_{ii'k}\widetilde{W}^F_{ii'}\Bigg\{(\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is})(\widetilde{F}^t_{i'k}/\sum_{k'} \widetilde{F}^t_{i'k'}) \\ &\qquad (1+\log\frac{(\widetilde{F}_{ik}/\sum_s \widetilde{F}_{is})(\widetilde{F}_{i'k}/\sum_{k'} \widetilde{F}_{i'k'})}{(\widetilde{F}^t_{ik}/\sum_s \widetilde{F}^t_{is})(\widetilde{F}^t_{i'k}/\sum_{k'} \widetilde{F}^t_{i'k'})})\Bigg\}. \end{aligned} \tag{3.13}$$

对辅助函数求偏导可得

$$\frac{\partial Z(\widetilde{F},\widetilde{F}^t)}{\partial \widetilde{F}_{ik}} = \sum_{k'}\frac{\partial Z(\widetilde{F},\widetilde{F}^t)}{\partial(\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})}\frac{\partial(\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})}{\partial \widetilde{F}_{ik}} = 0. \tag{3.14}$$

因为$F_{ik}=\frac{\widetilde{F}_{ik}}{\sum_s \widetilde{F}_{is}}$，可知

$$\begin{aligned}\frac{\partial Z(\widetilde{F},\widetilde{F}^t)}{\partial(\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})} &= -2\left\{P^+_{ik'}+\lambda_F[\widetilde{W}^F F^t]_{ik'}\right\}\frac{F^t_{ik'}}{F_{ik'}} \\ &\quad +2\left\{P^-_{ik'}+[FQ]_{ik'}+\lambda_F[\widetilde{D}^F F^t]_{ik'}\right\}\frac{F_{ik'}}{F^t_{ik'}},\end{aligned} \tag{3.15}$$

$$\frac{\partial(\frac{\widetilde{F}_{ik'}}{\sum_s \widetilde{F}_{is}})}{\partial \widetilde{F}_{ik}}=\frac{\delta_{ik'}-F_{ik'}}{\sum_s \widetilde{F}_{is}}, \delta_{ik'}=\begin{cases}1 & \text{if } k=k' \\ 0 & \text{otherwise.}\end{cases} \tag{3.16}$$

令$A=(P^-+F^tQ+\lambda_F\widetilde{D}^F F^t)\oslash F^t$，$C=(P^++\lambda_F\widetilde{W}^F F^t)\odot F^t$，式子(3.14)可以表示为

$$A_{ik}F^2_{ik}+\sum_s[C_{is}-A_{is}F^2_{is}]F_{ik}-C_{ik}=0. \tag{3.17}$$

注意到式子(3.17)仅仅含有变量F。对式子(3.17)求和可得

$$\begin{aligned}&\sum_k A_{ik}F^2_{ik}+\sum_s[C_{is}-A_{is}F^2_{is}]\sum_k F_{ik}-\sum_k C_{ik} \\ =&\sum_s[C_{is}-A_{is}F^2_{is}](\sum_k F_{ik}-1)=0.\end{aligned} \tag{3.18}$$

很明显式子(3.18)的解对于几乎所有的λ_F总是满足归一化约束.

在本章中我们采用谁在文献[94]中提出的算法6计算式子(3.17)的固定点，其中进涉及到element-wise操作并且只需要少数几次（如$nIter=10$）迭代收敛

3.3.2.3 计算H

对式子(3.4)中H的优化等价于最小化

$$\mathcal{J}_6=||X-FHG^T-S||^2_F, \quad \text{s.t.} \quad H\geq 0. \tag{3.19}$$

通过引入Lagrangian乘子$\Lambda_H\in\mathbb{R}^{k_1\times k_2}$并且设置偏导数为0，可以得到

$$\Lambda_H=-2F^T(X-S)G+2F^TFHG^TG. \tag{3.20}$$

使用KKT最优性条件，$(\Lambda_H)_{ij}H^2_{ij}=0$，可以得到以下的更新公式

$$H_{ij}=H_{ij}\sqrt{\frac{[F^T(X-S)G]^+_{ij}}{[F^TFHG^TG]_{ij}+[F^T(X-S)G]^-_{ij}}}. \tag{3.21}$$

Algorithm 6 计算式子(3.17)的不动点

输入： A, C, F^t, $nIter$和k.

输出： F

while $iter \leq nIter$ **do**
  $B = (C - A \otimes F^2)\mathbf{1}_k\mathbf{1}_k^T$;
  $F = (\sqrt{B^2 + 4AC} - B) \oslash (2A)$
  $\mathcal{I} = \{j | B_{j1} < 0\}$
  if $\mathcal{I}$非空 **then**
    $F_{\mathcal{I}\cdot} = \mathrm{diag}(F_{\mathcal{I}\cdot}\mathbf{1}_k)^{-1}F_{\mathcal{I}\cdot}$
  end if
end while

3.3.2.4　计算G

对式子(3.4)中G的优化等价于最小化

$$\mathcal{J}_7 = ||X^T - GH^TF^T - S^T||_F^2 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2$$
$$\text{s.t.} \quad G \geq 0, G\mathbf{1}_{k_2} = \mathbf{1}_n. \tag{3.22}$$

考虑到F和G的对偶性，对式子(3.22)中G的最小化问题可以通过前面提出的对式子(3.8)中F的优化算法来求解。

本章提出的鲁棒联合聚类算法（Robust Co-Clustering，简称RCC）的求解过程如算法7所示。

3.3.3　算法收敛性证明

这一节分析本章提出的迭代算法的收敛性，我们通过Lee和Seung在文献[59]提出的辅助函数和文献[76]中的引理证明上一节提出的算法的收敛性。

定义3.1.
对于任意给定的x和x^t，如果满足以下条件，

$$Z(x, x^t) \geq J(x), Z(x, x) = J(x).$$

就说$Z(x, x^t)$是函数$J(x)$的辅助函数[59]。

Algorithm 7 用于求解式子(3.4)的RCC算法

输入： 数据矩阵$X \in \mathbb{R}_{+}^{d\times n}$, 特征簇的个数$k_1$，样本簇的个数$k_2$，最近邻的个数$p$，特征图正则参数$\lambda_F$，样本图正则参数$\lambda_G$，稀疏噪声正则参数$\lambda_S$。

输出： F, H, S和S.

步骤一：初始化非负因子F, H, G

步骤二：构造特征邻近图W_F和样本邻近图W_G

repeat

根据式子(3.6)更新S;

令$\widetilde{W}_{ii'}^F = \frac{W_{ii'}^F}{2||F_{i\cdot}-F_{i'\cdot}||_2}$, $\widetilde{D}_{ii}^F = \sum_{i'} \widetilde{W}_{ii'}^F$, $P = (X-S)GH^T$, $Q = HG^TGH^T$;

计算$A = (P^- + FQ + \lambda_F \widetilde{D}^F F) \oslash F$;

计算$C = (P^+ + \lambda_F \widetilde{W}^F F) \odot F$;

根据算法(6)更新F;

根据式子(3.21)更新H;

令$\widetilde{W}_{ii'}^G = \frac{W_{ii'}^G}{2||G_{i\cdot}-G_{i'\cdot}||_2}$, $\widetilde{D}_{ii}^G = \sum_{i'} \widetilde{W}_{ii'}^G$, $P = (X-S)^TFH$, $Q = H^TF^TFH$;

计算$A = (P^- + GQ + \lambda_G \widetilde{D}^G G) \oslash G$;

计算$C = (P^+ + \lambda_G \widetilde{W}^G G) \odot G$;

根据算法(6)更新G;

until Converges

引理3.1.

如果Z是函数J的辅助函数，通过求解下面的问题

$$x^{t+1} = \arg\min_x Z(x, x^t).$$

我们可以单调的减小目标函数J[59]。

引理3.2.

对于任意的两个向量$\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$，下面的不等式成立[76]

$$||\boldsymbol{a}||_2 - \frac{||\boldsymbol{a}||_2^2}{2||\boldsymbol{b}||_2} \leq ||\boldsymbol{b}||_2 - \frac{||\boldsymbol{b}||_2^2}{2||\boldsymbol{b}||_2}. \tag{3.23}$$

定理3.1.

算法7使式子(3.4)中RCC目标函数在每次迭代中单调下降并且收敛到一个最优解。

证明. 由于式子(3.4)中的目标函数有下界零，我们只需要证明在每次迭代中对$S^{t+1}, F^{t+1}, H^{t+1}$,和$G^{t+1}$的更新使得目标函数单调下降。通过式子(3.6)获得的S^{t+1}是式子(3.5)对于子问题的全局最优解。根据辅助函数[59]的特点，可以得到$\mathcal{J}_4(F^t) \geq Z(F^{t+1}, F^t) \geq Z(F^{t+1}, F^{t+1}) = \mathcal{J}_4(F^{t+1})$，更新$F^{t+1}$会使得式子(3.11)中的目标函数值单调下降，即$\mathcal{J}_4(F^{t+1}) \leq \mathcal{J}_4(F^t)$，也就是说

$$\begin{aligned}
&||X - F^{t+1}HG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F \frac{||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2} \\
\leq &||X - F^tHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F \frac{||F_{i\cdot}^t - F_{i'\cdot}^t||_2^2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2}.
\end{aligned}$$

因此，这里我们仅需要证明更新F^{t+1}满足$\mathcal{J}_4(F^{t+1}) \leq \mathcal{J}_3(F^t)$。并且基于$\mathcal{J}_4(F^{t+1}) \leq \mathcal{J}_4(F^t)$我们有下面的不等式

$$\begin{aligned}
&||X - F^{t+1}HG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2 \\
&+ \lambda_F \sum_{ii'} W_{ii'} \left(\frac{||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2^2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2} - ||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2 \right) \\
\leq &||X - F^tHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot}^t - F_{i'\cdot}^t||_2 \\
&+ \lambda_F \sum_{ii'} W_{ii'} \left(\frac{||F_{i\cdot}^t - F_{i'\cdot}^t||_2^2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2} - ||F_{i\cdot}^t - F_{i'\cdot}^t||_2 \right)
\end{aligned} \tag{3.24}$$

根据引理(3.2)中的结果，我们知道

$$\begin{aligned}
&\frac{||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2^2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2} - ||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2 \\
&\geq \frac{||F_{i\cdot}^t - F_{i'\cdot}^t||_2^2}{2||F_{i\cdot}^t - F_{i'\cdot}^t||_2} - ||F_{i\cdot}^t - F_{i'\cdot}^t||_2.
\end{aligned} \tag{3.25}$$

结合式子(3.24)和式子(3.25)可以得出下面的结果

$$\begin{aligned}
&||X - F^{t+1}HG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot}^{t+1} - F_{i'\cdot}^{t+1}||_2 \\
\leq &||X - F^tHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot}^t - F_{i'\cdot}^t||_2
\end{aligned}$$

这个不等式表明式子(3.8)中的目标函数$\mathcal{J}_3$在每一次迭代中单调下降。类似的，我们也可以证明式子(3.22)中的目标函数$\mathcal{J}_6$通过更新G^{t+1}也会单调下降。类似

的我们也可以根据辅助函数的特性来证明更新H^{t+1}将使得式子(3.21)的目标函数单调递减。 □

3.4 实验结果

在这一节中，我们在一些实际应用中产生的数据集上进行试验来评测本章提出的联合聚类算法的鲁棒性和有效性。

3.4.1 对比算法和参数设置

我们对比以下的聚类和联合聚类方法：

- Kmeans，代表性的划分式聚类算法。
- 基于规范化割（Normalized Cut）的谱聚类算法[92]，下面简称：NCut。和[37, 90, 113]中使用的参数设置方法相同，两个样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的相似度采用高斯核函数：$w_{ij} = \exp(-\frac{||\boldsymbol{x}_i-\boldsymbol{x}_j||^2}{t})$计算，参数$t$从下列值中选取：$t \in \{1e^{-2}, 1e^{-1}, 1, 1e^{1}, 1e^{2}\}$。
- 非负矩阵三因子正交分解（Orthogonal nonnegative matrix tri-factorizations），以下简称为：ONMTF，该算法在文献[29]中提出，此算法存在特征簇个数k_1，样本簇个数k_2两个参数。
- 图正则非负矩阵分解（Graph Regularized Nonnegative Matrix Facotrization），以下简称为：GNMF，该算法在文献[14, 16]中提出。该算法存在非负因子个数k，最近邻的个数p，图正则参数λ。不同于文献[14]中使用固定参数$p = 5, \lambda = 100$，为了更公平的比较这些算法，我们使用文献[37]中使用的参数设置方法，即从下列值中选取：$p \in \{1, \ldots, 10\}$，$k = c$，$\lambda \in \{0.1, 1, 100, 500, 1000\}$。
- 对偶正则联合聚类算法（Dual Regularized Co-Clustering），以下简称为：DRCC，该算法在文献[37]提出。该算法存在特征簇个数k_1，样本簇个数k_2，特征关系图和样本关系图中最近邻的个数p，特征图和样本图的正则参数λ_F和λ_G。我们采用文献[37]中的方法设置参数搜索范围，即$k_1 = k_2 = c, p = \{1, \ldots, 10\}, \lambda_F = \lambda_G \in \{0.1, 1, 10, 100, 500, 1000\}$。

- 对偶非负矩阵三因子分解（Graph dual regularization non-negative matrix tri-factorization algorithm），以下简称为：DNMTF，该算法在文献[90]提出。DNMTF算法涉及到的参数和上面提到的DRCC算法相同，文献[90]中使用的参数设置方法也与文献[37]相同，因此本实验中我们采用和DRCC相同的方法设置DNMTF的参数。

- IGNMTF，文献[36]中提出的带规范化割约束的GNMTF，该算法的参数和DNMTF相同。

- 局部判别式联合聚类算法（Locally Discriminative Coclustering），以下简称为：LDCC，该算法在文献[113]提出。此算法使用的参数包括样本和特征嵌入维度k，特征关系图和样本关系图中最近邻的个数p，特征图和样本图局部正则参数λ，特征图和样本图正则参数α和β。我们使用文献[113]中使用的参数设置方法，即从下列值中选取：$k \in \{5, 10, \ldots, 50\}, p = \{1, \ldots, 10\}, \lambda = 1, \alpha = \beta \in \{0.1, 1, 10, 100, 500, 1000\}$。

- 本章中提出的鲁棒联合聚类算法，以下简称为：RCC。RCC算法涉及到的参数包括特征簇个数k_1，样本簇个数k_2，特征关系图和样本关系图中最近邻的个数p，特征图和样本图的正则参数λ_F和λ_G，以及稀疏噪音正则参数λ_S。我们采用DRCC相同的方法设置这些参数，即$k_1 = k_2 = c, p = \{1, \ldots, 10\}, \lambda_F = \lambda_G \in \{0.1, 1, 10, 100, 500, 1000\}$，此外基于式子(3.5)稀疏噪音与式子(3.7)Huber损失函数之间的对偶关系，正则化参数λ_S经验性的设置为$\lambda_S = 2r\text{median}(|E_{ij}|)$，其中$r$默认取1。由于重构误差$E_{ij}$在每次迭代中都发生变化，我们也相应的调整$\lambda_S$。

由于每个聚类算法都有一个或者多个参数需要调整，此外这些聚类算法都依赖于初始化（包括算法初始化和后处理初始化），为了公平的比较这些方法，对每个算法我们在不同的参数设置下多次运行这些算法，最后比较每个算法在不同参数设置下能达到最好平均结果。

3.4.2　数据集

本实验中，我们使用五个公开数据集来验证所提算法的有效性。这些数据集包括一个人脸数据集JAFFE[107]，两个手写数字图像数据集MFEA和OPTDIGIT[90]以

及两个基因表达数据集LUNG和GLIOMA[76]。表3.2总结了这些数据集的统计特征。

表 3.2: 第三章实验所使用的数据集

Data sets	# samples	# features	# classes
JAFFE	213	676	10
MFEA	2000	240	10
OPTDIGIT	3823	64	10
LUNG	203	3312	5
GLIOMA	50	4434	4

3.4.3 实验结果

给定聚类算法生成的簇和真实类别标签，我们使用了两个聚类的性能度量来评估。一个指标是准确性（ACC），它是用来测量的百分比正确的标签。第二个指标是归一化互信息（NMI）。这些指标的定义见第2.4.2节。

3.4.3.1 聚类结果

在本节中，我们评估这些算法的聚类效果。由于这些聚类算法都依赖于初始化，在每个参数设置下，我们独立的运行这些实验20次，取其平均值，最后在表3.3和表3.4报告每个算法在所有参数下平均值最高的结果。可以看出，RCC比这些数据集上的聚类准确性和归一化互信息都优于对比的聚类和联合聚类算法。这也说明实际数据往往在某种程度上含有未知的噪声和异常点，这些噪声可能来自于数据产生过程，如图像遮挡、光照等。通过显式的处理这些噪声，我们提出的算法能够提高联合聚类在一般性数据上的效果。

3.4.3.2 参数敏感性和算法收敛性实验

本章提出的RCC算法有四个参数，分别是近邻个数p，正则化参数λ_F, λ_G和λ_S，需要指出的是，当前的GNMTF方法[36, 38, 90]都含有前面三个参数。在本小节中我们以JAFFE数据集为例，在图3.1 (a-c)中给出了RCC算法在这些参数上聚类结果，可以看出RCC在前面三个参数上的结果相对稳定对第四个参数λ_S不

表 3.3: 第三章聚类准确性

	JAFFE	MFEA	OPTDIGIT	LUNG	GLIOMA
Kmeans	0.7239	0.6862	0.7280	0.6911	0.5520
NCut	0.8392	0.7296	0.7855	0.7667	0.5680
ONMTF	0.7777	0.6511	0.6733	0.6931	0.5640
GNMF	0.9371	0.9158	0.7908	0.7697	0.5800
DRCC	0.9305	0.8759	0.7854	0.7719	0.5800
DNMTF	0.9305	0.9255	0.8055	0.8266	0.5800
IGNMTF	0.9256	0.8358	0.7910	0.8162	0.5640
LDCC	0.8249	0.7974	0.7700	0.7574	0.6520
RCC	**0.9765**	**0.9438**	**0.8558**	**0.8808**	**0.6840**

敏感。我们在图3.1(d)中画出了RCC算法的目标函数值随迭代次数的下降关系，可以看出RCC算法收敛较快。

3.5　本章小结

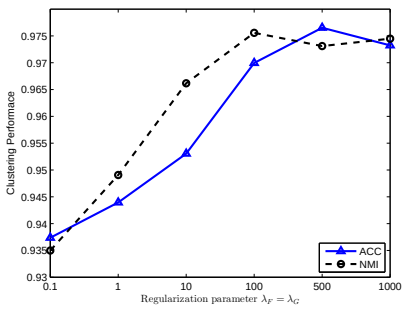
本章中，我们提出一种新的基于非负矩阵分解的鲁棒联合聚类算法（Robust Co-Clustering，简称RCC）。为了处理样本与特征关系矩阵中的噪声，我们引入一个稀疏的错误矩阵刻画样本与特征关系中存在的较大的非高斯噪声，并从中恢复不含大噪声的关系矩阵；为了进一步刻画恢复的关系矩阵中存在的高斯小噪声我们引入平方误差度量此关系矩阵上的重构误差。为了降低特征与特征以及样本与样本之间噪声关系的影响，我们采用绝对值图正则误差而不是常用的平方图正则误差，此外由于绝对值函数图正则产生的稀疏性还可以产生更紧凑的矩阵分解结果。因此，本章提出的算法能有效的处理数据重构中出现的高斯、非高斯以及不可靠图正则等噪声，全面提高GNMTF方法的鲁棒性。我们提出了基于轮换最小化的迭代式算法来学习模型中的参数，并且证明了该算法的收敛性。真实数据集上的联合聚类实验结果表明，我们提出的鲁棒联合聚类算法要好于当前的主流方法。

这里我们列出并讨论一些关于本章的工作值得进一步研究的问题：

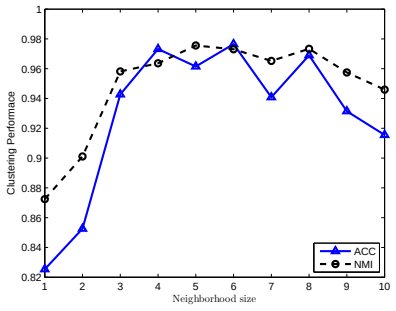
- 本章我们仅考虑基于单个无监督关系的图正则问题，实际应用中数据可

表 3.4: 第三章归一化互信息

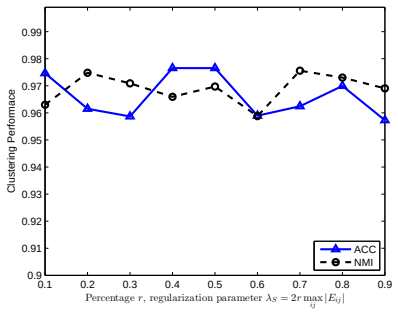
	JAFFE	MFEA	OPTDIGIT	LUNG	GLIOMA
Kmeans	0.7942	0.7016	0.7314	0.5069	0.4710
NCut	0.7974	0.8320	0.7203	0.5982	0.4630
ONMTF	0.8188	0.6189	0.6269	0.5137	0.4405
GNMF	0.9384	0.9051	0.8220	0.6670	0.3523
DRCC	0.9480	0.8257	0.8306	0.5552	0.4841
DNMTF	0.9507	0.9088	0.8361	0.5977	0.5047
IGNMTF	0.9175	0.8528	0.8301	0.5786	0.4880
LDCC	0.8798	0.8400	0.7929	0.5595	0.4979
RCC	**0.9756**	**0.9117**	**0.8611**	**0.6909**	**0.5350**



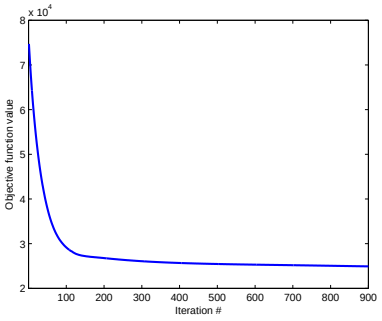
(a)



(b)



(c)



(d)

图 3.1: 参数敏感性结果(a)、(b)和(c)，收敛性曲线(d)

能来自多个子空间或者子流形，而单个关系图仅能部分刻画数据的本质结构，因此我们可以考虑集成流形图正则的鲁棒非负矩阵分解模型。

- 联合聚类算法以及本章提出的鲁棒联合聚类算法往往包含多个需要指定的超参数；并且由于这些模型往往是非凸的，算法的好坏还依赖于参数的初始化。一个好的鲁棒模型和算法不仅可以有效处理数据中的噪声还需要对模型自身的参数和超参数鲁棒。因此如何提高参数初始化的质量和超参数的稳定性对鲁棒联合聚类算法尤为重要。

第四章　鲁棒图正则非负矩阵三因子分解模型和算法

在第二章聚类分析中我们研究了基于半二次损失函数的鲁棒非负矩阵分解，而在第三章联合聚类中研究了一种鲁棒图正则非负矩阵分解方法。为了进一步提高鲁棒联合聚类算法的鲁棒性，我们首先证明之前提出的鲁棒联合聚类算法中基于稀疏性噪声的重构误差与休伯损失函数的对偶关系，然后进一步提出采用一般的半二次损失函数来分别度量特征与样本关系矩阵的重构误差以及样本与样本，特征与特征关系的图正则误差。我们分别推导出基于乘法形式和加法形式的鲁棒图正则非负矩阵三因子分解模型的求解算法并证明其收敛性。

4.1　鲁棒图正则非负矩阵三因子分解模型

首先回顾RCC求解的矩阵分解问题，这里我们不考虑非负因子的归一化约束（事实上，不带归一化约束的非二次图正则模型仍然是定义良好的）。

$$\begin{aligned}\mathcal{J}_1 = &||X - FHG^T - S||_F^2 + \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2 \\ &+ \lambda_S ||S||_1 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2 \\ &\text{s.t.} \quad F, H, G \geq 0.\end{aligned} \tag{4.1}$$

其中稀疏噪声矩阵S可以通过下面的式子求解

$$S_{ij} = \begin{cases} 0 & \text{if } |E_{ij}| \leq \lambda_S/2, \\ E_{ij} - \frac{\lambda_S}{2}\text{sign}(E_{ij}) & \text{otherwise}, \end{cases} \tag{4.2}$$

其中$E_{ij} = (X - FHG^T)_{ij}$。将式子(4.25) 代入式子(4.1)可得

$$\min \mathcal{J} = \begin{cases} E_{ij}^2 & \text{if } |E_{ij}| \leq \lambda_S/2, \\ \lambda_S |E_{ij}| - (\frac{\lambda_S}{2})^2 & \text{otherwise.} \end{cases} \tag{4.3}$$

$$+ \lambda_F \sum_{ii'} W_{ii'}^F ||F_{i\cdot} - F_{i'\cdot}||_2 + \lambda_G \sum_{jj'} W_{jj'}^G ||G_{j\cdot} - G_{j'\cdot}||_2 \tag{4.4}$$

由此我们可以看出RCC的目标函数包含基于Huber损失的重构误差和基于ℓ_1的图正则误差。由于Huber损失函数属于半二次损失函数，我们可以通过加法形

式和乘法形式的半二次优化技术求解。Huber损失函数的半二次优化加法形式可以写成：

$$\ell_{Huber}(E_{ij}) = \sum_{ij} \left\{ (E_{ij} - S_{ij})^2 + g_A(S_{ij}) \right\}, \tag{4.5}$$

其中S_{ij}是辅助变量。S_{ij}可由式子(2.56)计算，即：

$$S_{ij} = \begin{cases} 0 & |E_{ij}| \leq c \\ E_{ij} - c\text{sign}(E_{ij}) & |E_{ij}| > c \end{cases} \tag{4.6}$$

当Huber损失函数中的超参数$c = \frac{\lambda_S}{2}$时，我们可以看出式子(4.6)等价于式子(4.25)。

根据上面讨论，RCC目标函数中最小化基于稀疏性噪声的重构误差和最小化Huber损失函数的对偶关系总是成立（既可以通过半二次优化加法形式又可以通过soft-thresholding算子建立）。根据这一对偶关系，我们可以得出RCC算法的鲁棒性不仅仅来自于对稀疏噪声的处理，还可能由于soft-thresholding算子的作用和鲁棒Huber估计量类似。

根据这一对偶关系我们可以进一步提出基于半二次最小化的鲁棒图正则非负矩阵三因子分解模型，即分别采用鲁棒的半二次损失函数度量重构误差和图正则误差。

$$\begin{aligned} \mathcal{J}_2 \quad &= \sum_{i=1}^{d}\sum_{j=1}^{n} f_E((X - FHG^T)_{ij}) \\ &+ \lambda_F \sum_{i=1}^{d}\sum_{i'=1}^{d} (W_F)_{ii'} f_F(||F_{i\cdot} - F_{i'\cdot}||_2) \\ &+ \lambda_G \sum_{j=1}^{n}\sum_{j'=1}^{n} (W_G)_{jj'} f_G(||G_{j\cdot} - G_{j'\cdot}||_2) \\ &\text{s.t.} \quad F \geq 0, H \geq 0, G \geq 0. \end{aligned} \tag{4.7}$$

其中f_E、f_F和f_G是半二次损失函数。需要指出的是，当f_F和f_G采用非二次函数时，式子(4.7)是定义良好的。

这里我们列出常见的半二次损失函数f：

$$\begin{aligned}
\ell_1\text{-}\ell_2\text{函数：} &\quad f(e)=\sqrt{c+e^2}-1\\
\text{Welsch函数：} &\quad f(e)=\frac{c}{2}[1-\exp(-e^2/c)]\\
\text{Huber 函数：} &\quad f(e)=\begin{cases}\frac{e^2}{2} & |e|\le c\\ c|e|-\frac{c^2}{2} & |e|>c\end{cases}
\end{aligned}$$

其中c是超参数。式子(4.7)中的目标函数同样是关于F,H,G的非负优化问题。因此，我们采用迭代式交互最小化算法求解。和第(二)章鲁棒非负矩阵分解求解算法类似，我们分别给出基于半二次优化乘法形式和加法形式的通用求解算法。

4.2　通用乘法形式的噪声检测算法

有第2.2.3节可知，损失函数的半二次乘法形式可以写成：

$$f(e)=\frac{1}{2}we^2+g_M(w),e\in\mathbb{R},w\in\mathbb{R}_+. \tag{4.8}$$

其中w是辅助变量，g_M是f乘法形式的对偶函数。

分别代入f_E,f_F,f_G对应的乘法形式，并且令$E_{ij}=(X-FHG^T)_{ij},(E_F)_{ii'}=||F_{i\cdot}-F_{i'\cdot}||_2$和$(E_G)_{jj'}=||G_{j\cdot}-G_{j'\cdot}||_2$，问题(4.7)可以写成最小化下面的在扩大的定义域上的目标函数：

$$\begin{aligned}
&\mathcal{J}_3(F,H,G,W_E,W_F,W_G)\\
&=\sum_{ij}(W_E)_{ij}E_{ij}^2+(g_E)_M((W_E)_{ij})\\
&+\lambda_F\sum_{ii'}\left((W_F)_{ii'}(E_F)_{ii'}^2+(g_F)_M((W_F)_{ii'})\right)\\
&+\lambda_G\sum_{jj'}\left((W_G)_{jj'}(E_G)_{jj'}^2+(g_G)_M((W_G)_{jj'})\right)
\end{aligned} \tag{4.9}$$

其中W_E,W_F和W_G是引入的辅助变量，而$(g_E)_M,(g_F)_M$和$(g_G)_M$分别是f_E,f_F和f_G对应的对偶函数。

计算辅助变量 对于半二次损失函数f，乘法形式的辅助变量最优解可以通过下面的函数计算：

$$h_M(e)=\begin{cases}f''(0^+) & \text{if } e=0,\\ f'(e)/e & \text{if } e\neq 0.\end{cases} \tag{4.10}$$

这里列出了常见的损失函数对于的权重函数h_M：

$$\begin{aligned}
\ell_1\text{-}\ell_2 \text{ function:} \quad & h_M(e) = \frac{1}{\sqrt{c+e^2}} \\
\text{Welsch function:} \quad & h_M(e) = \exp(-e^2/c) \\
\text{Huber function:} \quad & h_M(e) = \begin{cases} 1 & |e| \le c \\ c/|e| & |e| > c \end{cases}
\end{aligned}$$

给定具体的损失函数，式子(4.9)中对应的辅助变量可通过下面的式子计算：

$$(W_E)_{ij} = h_E(E_{ij}) \tag{4.11}$$

$$(\widetilde{W}_F)_{ii'} = (W_F)_{ii} h_F((E_F)_{ii}) \tag{4.12}$$

$$(\widetilde{W}_G)_{jj'} = (W_G)_{jj} h_G((E_G)_{jj}) \tag{4.13}$$

其中我们引入$\widetilde{W}_F$和$\widetilde{W}_G$来表示精化的关系矩阵。

计算非负因子F, H, G 当辅助变量给定时，问题(4.9)简化为加权的GNMTF，即

$$\begin{aligned}
\mathcal{J}_4 = & ||W_E \odot (X - FHG^T)||_F^2 \\
& + \lambda_F \mathrm{tr}(F^T \widetilde{L}_F F) + \lambda_G \mathrm{tr}(G^T \widetilde{L}_G G)
\end{aligned} \tag{4.14}$$

where $\widetilde{L}_F = \widetilde{D}_F - \widetilde{W}_F$和$\widetilde{L}_G = \widetilde{D}_G - \widetilde{W}_G$. 需要注意的是上面式子中第一项的权重和第二、三项的权重分别基于当前的估计的精化。和式子(3.2)中的GNMTF相比，式子(4.7)中的目标函数在每次迭代中都估计重构误差和图正则的权重，逐渐的降低大的误差的权重。因此这些乘法形式的辅助变量可以看作是噪声掩码(mask)，可用于屏蔽异常点对矩阵分解的影响。

和文献[39]中的GNMTF算法类似，通过分别将每个变量的偏导设为零，我们可以推出下面的乘法更新公式来最小化式子(4.14)中的目标函数。

$$F_{ij} = F_{ij} \sqrt{\frac{[(W_E \odot X)GH^T + \lambda_F \widetilde{W}_F F]_{ij}}{[(W_E \odot (FHG^T))GH^T + \lambda_F \widetilde{D}_F F]_{ij}}} \tag{4.15}$$

$$H_{ij} = H_{ij} \sqrt{\frac{[F^T(W_E \odot X)G]_{ij}}{[F^T(W_E \odot (FHG^T))G]_{ij}}} \tag{4.16}$$

$$G_{ij} = F_{ij}\sqrt{\frac{[(W_E \odot X)^T FH + \lambda_G \widetilde{W}_G G]_{ij}}{[(W_E \odot (FHG^T))^T FH + \lambda_G \widetilde{D}_G G]_{ij}}} \tag{4.17}$$

算法收敛性分析

本节中，我们证明所提出的通用求解算法的收敛性。我们首先利用一个辅助函数说明更新公式(6.21)使得目标函数(4.14)单调递减。

引理4.1.
对式子(4.14)中F的优化问题等价于最小化：

$$\begin{aligned}\mathcal{J}_7(F) = &\sum_i \mathrm{tr}(F_{i\cdot} HG^T \mathrm{diag}((W_E)_{i\cdot}) GH^T F_{i\cdot}^T) \\ &- 2\mathrm{tr}((W_E \odot X) GH^T F^T) + \lambda_F \mathrm{tr}(F^T(\widetilde{D} - \widetilde{W})F).\end{aligned} \tag{4.18}$$

下面的函数是$\mathcal{J}_7(F)$的辅助函数，此外它还是一个凸函数且其最优解可通过式子(6.21)计算。

$$\begin{aligned}Z(F, F^t) = &-2\sum_{ij}[(W_E \odot X)GH^T]_{ij} F_{ij}^t (1 + \log \frac{F_{ij}}{F_{ij}^t}) \\ &+ \sum_{ij} \frac{\left\{(F_{i\cdot}^t [HG^T \mathrm{diag}((W_E)_{i\cdot}) GH^T])_j + \lambda_F (DF^t)_{ij}\right\} F_{ij}^2}{F_{ij}^t} \\ &- \lambda_F \sum_{ijk} (\widetilde{W}_F)_{ij} F_{ik}^t F_{jk}^t (1 + \log \frac{F_{ik} F_{jk}}{F_{ik}^t F_{jk}^t})\end{aligned}$$

证明. 和[37]类似，我们可以验证$Z(F, F^t) \geq \mathcal{J}_7(F), Z(F, F) = \mathcal{J}_7(F)$，Hessian矩阵$\nabla\nabla_F Z(F, F^t) \succeq 0$。设$\mathcal{J}_7$关于$F$的偏导为零并利用KKT条件，我们就可以得到公式(6.21)用于最小化$\mathcal{J}_7$。 □

定理4.1.
利用公式(6.21)计算F会单调的减小目标函数(4.18)，因此公式(6.21)会使(4.18)收敛。

证明. 利用引理(3.1)和(4.1)，可知$\mathcal{J}_7(F^t) = Z(F^t, F^t) \geq Z(F^{t+1}, F^t) \geq Z(F^{t+1}, F^{t+1}) = \mathcal{J}_7(F^{t+1})$。因此，$\mathcal{J}_7(F)$会单调下降。由于$\mathcal{J}_7(F)$有下界，定理(4.1)的正确性和收敛性得证。 □

与定理(4.1)类似，我们也可以证明公式(4.16)-(6.22)和公式(6.27)-(6.28)的收敛性。

定理4.2.
通用求解算法中由公式(4.11)-(4.13)和公式(6.21)-(6.22)迭代生成的序列 $\{W_E^t, W_F^t, W_G^t, F^t, H^t, G^t\}_{t=1}^{\infty}$ *可以最小化目标函数(4.7)并且收敛。*

证明. 根据半二次乘法形式的最小化辅助函数h_M，也就是$\{Q(e_i, h(e_i)) + g(e_i)\} \leq \{Q(e_i, p_i) + g(p_i)\}$，对于给定的$\{F^t, H^t, G^t\}$，我们可得$J_3(F^t, H^t, G^t, W_E^{t+1}, W_F^{t+1}, W_G^{t+1}) \leq J_3(F^t, H^t, G^t, W_E^t, W_F^t, W_G^t)$。根据定理(4.1)，对于给定的$W_E^{t+1}, W_F^{t+1}, W_G^{t+1}$，公式(6.21)-(6.22)会单调的减小式子(3.5)中的目标函数。由于式子(3.5)是有下界的，因此$t \to \infty$时迭代的整个序列会收敛。 □

4.3 通用加法形式的噪声矫正算法

本节中，我们利用半二次优化的加法形式求解问题(4.7) 有第2.2.3节可知，损失函数f的半二次加法形式可以写成：

$$f(e) = \frac{1}{2}(e-s)^2 + g_A(s), e, s \in \mathbb{R}. \tag{4.19}$$

考虑到图正则部分本身就是加权误差，这里我们仅考虑重构误差上的加法形式，而使用乘法形式的图正则。代入对应的半二次乘法形式和加法形式，问题(4.7)可以重写为最小化扩大的定义域上的优化问题：

$$\begin{aligned}
&\mathcal{J}_5(F, H, G, S, W_F, W_G) \\
&= \sum_{ij}((X - FHG^T - S)_{ij}^2 + (g_E)_A(S_{ij})) \\
&+ \lambda_F \sum_{ii'}^{d} \left((W_F)_{ii'}(E_F)_{ii'}^2 + (g_F)_M((W_F)_{ii'})\right) \\
&+ \lambda_G \sum_{jj'} \left((W_G)_{jj'}(E_G)_{jj'}^2 + (g_G)_M((W_G)_{jj'})\right)
\end{aligned} \tag{4.20}$$

其中S是辅助变量, $(g_E)_A$是损失函数f_E加法形式对应的对偶函数。

计算辅助变量 和乘法形式类似，加法形式中辅助变量的最优解可通过下面的式子计算：

$$h_A(e) = e - f'(e). \tag{4.21}$$

常见的半二次损失函数对应的加法形式辅助变量为：

$$\begin{aligned}
\ell_1\text{-}\ell_2 \text{ function:} \quad & h(e) = e - \frac{e}{\sqrt{c+e^2}} \\
\text{Welsch function:} \quad & h(e) = e - e\exp(-e^2/c) \\
\text{Huber function:} \quad & f(e) = \begin{cases} 0 & |e| \le c \\ e - c\text{sign}(e) & |e| > c \end{cases}
\end{aligned}$$

对于给定的损失函数f，式子(4.20)中的S可由下面的公式计算：

$$S_{ij} = (h_E)_A(E_{ij}) \tag{4.22}$$

对于乘法形式中图正则部分的辅助变量$\widetilde{W}_F$和$\widetilde{W}_G$，我们可以通过式子(4.12)和式子(4.13)计算。

计算非负因子F, H, G 给定辅助变量，问题(4.20)等价于最小化：

$$\begin{aligned}
\mathcal{J}_6 = & ||X - FHG^T - S||_2^2 \\
& + \lambda_F \text{tr}(F^T \widetilde{L}_F F) + \lambda_G \text{tr}(G^T \widetilde{L}_G G)
\end{aligned} \tag{4.23}$$

需要注意的是上面式子中第一项的噪声估计和第二、三项的权重估计分别是基于当前模型的精化。和式子(3.2)中的GNMTF相比，式子(4.23)中的第一项估计当前重构中的噪声，恢复不含噪声的矩阵，并且从恢复的不含噪声的矩阵中学习非负因子；而第二、三项降低噪声图正则的影响。

通过分别将每个非负因子的偏导设为零，我们可以得到下面的乘法更新公式。

$$F_{ij} = F_{ij}\sqrt{\frac{[A^+ + \lambda_F \widetilde{W}_F F]_{ij}}{[A^- + FHG^TGH^T + \lambda_F \widetilde{D}_F F]_{ij}}} \tag{4.24}$$

$$H_{ij} = H_{ij}\sqrt{\frac{[F^T(X-S)G]^+_{ij}}{[F^TFHG^TG + (F^T(X-S)G)^-]_{ij}}} \tag{4.25}$$

$$G_{ij} = G_{ij}\sqrt{\frac{[B^+ + \lambda_G \widetilde{W}_G G]_{ij}}{[B^- + GH^TF^TFH + \lambda_G \widetilde{D}_G G]_{ij}}} \tag{4.26}$$

其中$A = (X-S)GH^T$和$B = (X-S)^TFH$，并且令$A^+_{ij} = (|A_{ij}| + A_{ij})/2$和$A^-_{ij} = (|A_{ij}| - A_{ij})/2$.

关于加法形式通用算法的收敛性可采用类似于第4.2节的方式证明。

4.4 本章小结

为了改善稀疏性噪声假设在实际应用中的局限性，我们首先证明本文之前提出的鲁棒联合聚类算法中基于稀疏性噪声的重构误差与休伯损失函数之间的对偶关系，然后进一步提出采用一般的半二次损失函数来分别度量特征与样本关系矩阵的重构误差以及样本与样本，特征与特征关系的图正则误差。我们分别推导出基于半二次优化乘法形式和加法形式的鲁棒图正则非负矩阵三因子分解模型的求解算法并证明其收敛性。

第五章　区间矩阵分解

矩阵分解方法在解决实际问题时大致分为三步，首先生成、构造、抽取数据的特征，对数据进行表示；然后对原始数据矩阵进行分解以期得到与实际应用相关的模型和数据表示，进而能够更加清晰的展现数据中的结构；最后应用学习得到的模型。其中原始数据表示的质量对后续的矩阵分解影响巨大。传统的矩阵分解技术可以看成基于单值的方法，即输入矩阵的每个元素只有一个数值（也包括缺失）。而在一些实际应用中，单值的数据往往无法准确表示和刻画实际问题，不可避免的引入了某些误差和噪声，这就给后续的矩阵分解带来了挑战。比如，在人脸分析中，矩阵分解需要对人脸进行旋转和对齐使得矩阵的每一列对应人脸的相同位置。这样的对齐要求在实际应用中很难满足，因此实际获得的数据矩阵往往是对人脸对齐的某种近似。再如，在协同过滤中理想的数据是每个元素都真实的反映了用户对物品的评分，而实际评分系统一般仅允许用户在一些离散的数据集合上选择评分，无法真正表示用户的偏好。这些例子说明我们需要表示能力更强的形式刻画实际应用中的数据。

由于真实观测数据中往往存在不可避免的误差和噪声，为了表示和刻画这些不确定性，我们认为通常观测到的数值来自某个概率分布，本章中假设来自某个均匀分布，此时我们可以用均匀分布的上下界来更准确的刻画实际数据。由于实际数据是单值的，我们就需要解决下面三个问题：1）如何利用单值观测数据中有限的上下文信息准确构造单个元素对应的概率分布的参数，也就是均匀分布的上下界？2）给定单个元素的概率分布，如何构造所有原始数据的概率分布和参数？3）由于传统的矩阵分解方法只处理单值矩阵，给定数据可能的上下界区间，如何从这些区间学习准确的矩阵分解模型？

本章我们采用下面的方法分别回答上面的三个问题。1）为了构造单值数据的上下界，我们假设观测到的单值为均匀分布的中值（中心），此时我们只需要准确估计均匀分布的半径一个参数。本章分析了两种数据挖掘应用的数据近似问题：人脸分析和协同过滤，提出了经验性方法来构造观测数据的可能的取值区间。在刻画人脸分析中的近似对齐问题时我们用单个像素周围的像素来估计该半径，而在协同过滤应用中，我们用单个评分对应的用户和物品的评分来估计均匀分布的半径。2）给定每个元素的上下界近似，我们假设所有元素

表 5.1: 第五章符号概要

符号	描述	符号	描述
n	输入数据点的数目	X_{ij}	矩阵X的第(i,j)元
d	数据维度	$I(X_{ij})$	区间矩阵$I(X)$的第(i,j)元
k	低维因子个数	$\boldsymbol{\delta}_{ij}$	X_{ij}对应的区间半径
X	单值数据矩阵$X \in \mathbb{R}^{n\times d}$	$(\boldsymbol{i},\boldsymbol{j})$	观测到的评分集合
$\boldsymbol{\delta}$	数据矩阵X对应的半径矩阵$\boldsymbol{\delta} \in \mathbb{R}^{n\times d}$	$\boldsymbol{j}_i$	第i-th行观测到的评分集合
X^{low}	下界矩阵	$\boldsymbol{i}_j$	第j-th列观测到的评分集合
X^{up}	上界矩阵	V	基矩阵
$I(X)$	基于X的区间数据矩阵	V^{low}	下界对应的基矩阵
$X_{i\cdot}$	X的第i行	V^{up}	上界对应的基矩阵
$X_{\cdot j}$	X的第j列	U	编码矩阵

是独立同分布的，因此我们可以分别用一个下界矩阵和上界矩阵表示所有元素的上下界。3）为了分别近似数据的上下界矩阵，本章提出面向区间数据的联合矩阵分解方法。该方法分别学习上下界矩阵对应的基矩阵并同时学习一个共享的编码矩阵，这样上下界矩阵就可以分别由上下界基矩阵和共享的编码矩阵的乘积计算，这样可以更好的近似数据可能的取值区间并且更好的刻画实际数据。基于人脸分析（包括人脸识别、聚类和重构）和协同过滤的大量实验结果表明本章提出的区间矩阵分解方法要显著优于标准的单值矩阵分解模型。

5.1 基于区间的数据近似

本节我们形式化的描述人脸对齐和协同过滤应用中基于区间的数据近似。表5.1中列出了本章使用的主要符号。我们首先给出区间矩阵的形式化定义，令$X \in \mathbb{R}^{n\times d}$表示实际应用的输入数据，$X_{ij}$为矩阵的一个元素。我们可以通过下面两种等价的定义[65]来刻画单值矩阵X对应的区间矩阵$I(X)$：

定义5.1.
基于中心点和半径的数据表示：我们定义以X_{ij}为中心δ_{ij}为半径的区间表示为

$$I(X_{ij}) =< X_{ij}, \delta_{ij} > \tag{5.1}$$

对矩阵的每个元素，我们都有$I(X) =< X, \boldsymbol{\delta} >$。

定义5.2.
基于上下界的数据表示：我们定义以X_{ij}为中心δ_{ij}为半径的区间对应的上下界分别为$X_{ij}^{low} = X_{ij} - \delta_{ij}$和$X_{ij}^{up} = X_{ij} + \delta_{ij}$，则$X_{ij}$对应的上下界区间表示为

$$I(X_{ij}) = [X_{ij}^{low}, X_{ij}^{up}] \tag{5.2}$$

对矩阵的每个元素，我们都有$I(X) = [X^{low}, X^{up}]$。

实际应用中，我们往往只观测到单值的数据矩阵而不是基于区间的数据。接下来，我们结合协同过滤和人脸对齐两个任务给出如何从观测数据X经验性的构造其对应的区间表示$I(X)$。

5.1.1　人脸分析中的近似对齐

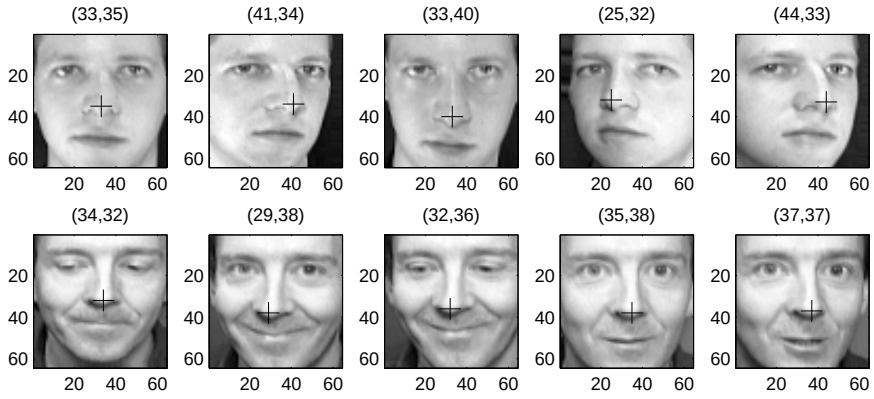


图 5.1: 人脸对齐近似

人脸识别的实际应用中往往需要对人脸图像进行对齐，如多张图像中的相同坐标部分应该对应人脸的相同位置。以图5.1中的鼻尖部分为例，可以看出将这些实际数据精确对齐往往比较困难，第一幅图中坐标$(33, 35)$的鼻尖应该对应于第二幅图坐标$(41, 34)$或者第三幅图的坐标$(33, 40)$。换句话说，即使某个像素并没有精确对齐，它的真实值也应该落在附近的像素。形式上讲某个坐标为$(x, y), x \in \{1, \ldots, d_x\}, y \in \{1, \ldots, d_y\}$的像素可能对应于坐标为$(x + \triangle x, y + \triangle y), 0 \leq \triangle x, \triangle y \leq r$的像素。矩阵分解中，第$i$-th个人脸表示为一个向量$X_{i\cdot} \in \mathbb{R}^{d_x \times d_y, 1}$，用$(x^{(i,j)}, y^{(i,j)})$表示第$i$-th图像中第$j$-th维，即$X_{ij}$，对

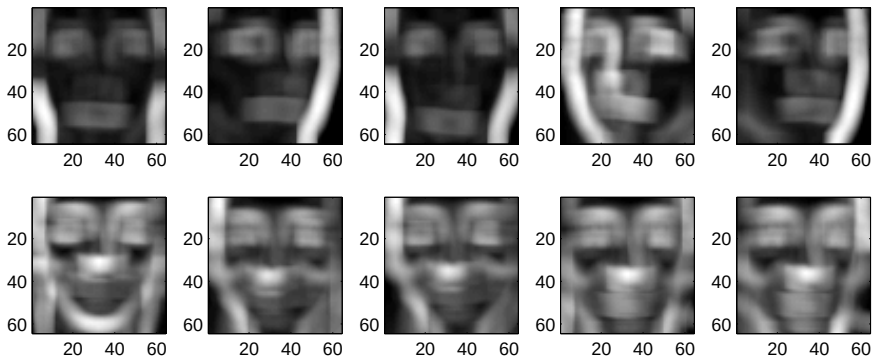


图 5.2: 图5.1对应的半径矩阵$\boldsymbol{\delta}$

应像素的坐标，我们就可以对每个像素X_{ij}定义下面的集合：

$$\mathcal{S}_{ij}^{\mathrm{FA}(r)} = \{X_{ij'} | |x^{(i,j')} - x^{(i,j)}| \leq r \wedge |y^{(i,j')} - y^{(i,j)}| \leq r\} \tag{5.3}$$

$\mathcal{S}_{ij}^{\mathrm{FA}(r)}$表示以$X_{ij}$为中心范围为$r$的像素集合。由于图像没有精确对齐，因此$X_{ij}$可能对应于区间$[\min(\mathcal{S}_{ij}^{\mathrm{FA}(r)}), \max(\mathcal{S}_{ij}^{\mathrm{FA}(r)})]$中的一个值，这也就是之前定义的基于上下界的数据表示。实际应用中，上下界估计并不准确，鲁棒性较差，因此我们用集合$\mathcal{S}_{ij}^{\mathrm{FA}(r)}$的标准差来刻画数据中的变化。根据定义5.1，令$X_{ij}$为$I(X_{ij})$的中心，其半径定义为

$$\delta_{ij}^{\mathrm{FA}(r)} = \alpha \cdot \mathrm{std}(\mathcal{S}_{ij}^{\mathrm{FA}(r)}) \tag{5.4}$$

其中$\alpha \in \mathbb{R}^+$是一个尺度系数。根据定义5.2就可以很容易的计算区间的上下界。图5.2中给出了图5.1每幅图像每个像素对应的半径，其中较亮的部分半径较大反之半径较小。从图5.2可以看出，眼睛和鼻子部分的半径较大，也就是说这些部分对图像对齐误差敏感，而这些误差会降低传统的单值矩阵分解模型的性能。

5.1.2 协同过滤中的评分近似

协同过滤中，评分度实际上是对用户真实偏好程度的某种近似。如某系统将用户对物品的评分分为$[1, 2, 3, 4, 5]$，而某个用户u对两个物品a和b的评分介于2和4之间，并且用户更喜欢a，假设用户的连续偏好程度分别为3.4和2.8，由于评分系统自身约束，用户u只能将a和b都评为3，此时用户的偏好关系被忽视。这个例子说明评分度实际上应该是连续的区间，这样才可能包含用户真实

的偏好度。简单来说，第i-th用户对第j-th物品的评分度X_{ij}同时受用户自身和物品自身影响，因此我们定义和X_{ij}相关的评分集合为：

$$\mathcal{S}_{ij}^{\mathrm{CF}(r)} = \{X_{i'j'} | (i' = i \vee j' = j) \wedge (i', j') \in (\boldsymbol{i}', \boldsymbol{j}')\} \tag{5.5}$$

可以看出$\mathcal{S}_{ij}$实际上是通过评分矩阵X的第i-th行和第j-th列观测到的评分构造。同样的，我们也可以根据相关评分集合$\mathcal{S}_{ij}$的标准差，估计每个评分的半径$\delta_{ij}^{\mathrm{CF}(r)}$，即：

$$\delta_{ij}^{\mathrm{CF}(r)} = \alpha \cdot \mathrm{std}(\mathcal{S}_{ij}^{\mathrm{CF}(r)}) \tag{5.6}$$

其中$\alpha \in \mathbb{R}^+$同样是一个尺度系数。实际应用中，某个用户对不同物品的评分以及不同用户对某个物品的打分差别往往较大，也就是说我们需要定义较大的半径来近似评分矩阵。根据中心的和半径，我们也可以很容易构造上下界区间矩阵。图5.3和图5.4中给出了一个单值评分矩阵和它对应的区间评分矩阵。

	m_1	m_2	m_3	m_4	m_5
u_1		1	4		5
u_2	3		1	2	
u_3		1			4
u_4	5				
u_5		1	4	2	
u_6		3		2	5

图 5.3: 单值评分矩阵X

	m_1	m_2	m_3	m_4	m_5
u_1		[0.6,1.4]	[3.5,4.5]		[4.8,5.2]
u_2	[2.8,3.2]		[0.5,1.5]	[1.5,2.5]	
u_3		[0.7,1.3]			[3.5,4.5]
u_4	[4.5,5.5]				
u_5		[0.4,1.6]	[3.7,4.3]	[1.8,2.2]	
u_6		[2.7, 3.3]		[1.4, 2.6]	[4.2, 5.8]

图 5.4: 图5.3对应的上下界区间矩阵$I(X)$

5.2 区间矩阵分解

本节中我们提出基于区间数据的矩阵分解技术（Interval-valued Matrix Factorization，简称IMF）。IMF是基于上下界区间表示的矩阵分解方法，即输入数据是$I(X) = [X^{\text{low}}, X^{\text{up}}]$。我们将现有的单值矩阵分解模型扩展为对上下界矩阵的联合分解。首先，我们假设矩阵的每个元素X_{ij}都是来自以$[X_{ij}^{\text{low}}, X_{ij}^{\text{up}}]$为区间的均匀分布，即：

$$X_{ij} \sim \text{uniform}(X_{ij}^{\text{low}}, X_{ij}^{\text{up}}) \tag{5.7}$$

基于这一假设，我们有

$$E(X_{ij}) = \frac{1}{2}(X_{ij}^{\text{low}} + X_{ij}^{\text{up}}) \tag{5.8}$$

因此，我们提出下面的联合矩阵分解来估计和近似区间数据的上下界，即：

$$X^{\text{low}} \to UV^{\text{low}} \quad X^{\text{up}} \to UV^{\text{up}} \tag{5.9}$$

在联合矩阵分解中，我们采用同一个矩阵U来表示每一个样本，用$V^{\text{low}}, V^{\text{up}}$来刻画对上下界的近似。对上下界$X^{\text{low}}$和$X^{\text{up}}$的重构误差可以通过下面的式子计算：

$$\hat{X}^{\text{low}} \leftarrow UV^{\text{low}} \quad \hat{X}^{\text{up}} \leftarrow UV^{\text{up}} \tag{5.10}$$

根据式子(5.8)和式子(5.10)，我们就可以用下面的式子重构原始单值数据X：

$$\hat{X} \leftarrow \frac{1}{2}(\hat{X}^{\text{low}} + \hat{X}^{\text{up}}) \tag{5.11}$$

5.2.1 区间非负矩阵分解

根据式子(2.3)和式子(5.10)，我们可得采用平方误差的区间非负矩阵分解（Interval-valued Nonnegative Matrix Factorization，简称I-NMF）：

$$\mathcal{L}_{\text{I-NMF}} = ||X^{\text{low}} - UV^{\text{low}}||^2 + ||X^{\text{up}} - UV^{\text{up}}||_F^2 \tag{5.12}$$

$$\text{s.t.} \quad U \geq 0, V^{\text{low}} \geq 0, V^{\text{up}} \geq 0 \tag{5.13}$$

和标准非负矩阵分解类似，我们通过下面的乘法更新公式最小化式子(5.12)中的目标函数。

$$U_{ij} = U_{ij}\frac{[X^{\text{low}}(V^{\text{low}})^T + X^{\text{up}}(V^{\text{up}})^T]_{ij}}{[UV^{\text{low}}(V^{\text{low}})^T + UV^{\text{up}}(V^{\text{up}})^T]_{ij}} \tag{5.14}$$

$$V_{ij}^{\text{low}} = V_{ij}^{\text{low}}\frac{(U^T X^{\text{low}})_{ij}}{(U^T UV^{\text{low}})_{ij}} \tag{5.15}$$

$$V_{ij}^{\text{up}} = V_{ij}^{\text{up}}\frac{(U^T X^{\text{up}})_{ij}}{(U^T UV^{\text{up}})_{ij}} \tag{5.16}$$

同样，式子(5.12)中的目标函数是有下界的，而乘法更新公式每次迭代都使目标函数单调下降，因此基于上面更新公式的I-NMF算法收敛。

可以看出，标准非负矩阵分解模型将原始单值数据分解为两个低维因子的乘积：一个基(basis)矩阵一个编码(encoding)矩阵。不同的是I-NMF采用联合矩阵分解的框架学习两个基因子$V^{\text{low}}, V^{\text{up}}$用于分别近似上下界矩阵，而学习一个编码矩阵$U$用于刻画原始数据的低维表示，我们可以直接利用I-NMF学习的编码因子U做相关分析。

5.2.2　区间概率矩阵分解

本小节，我们首先回顾标准概率矩阵分解（Probabilistic Matrix Factorization，简称PMF）[86, 87]，然后介绍基于区间数据的概率矩阵分解（Interval-valued Probabilistic Matrix Factorization，简称I-PMF）。

5.2.2.1　PMF

协同过滤中，PMF模型假设评分来自于某种高斯分布，即

$$p(X_{ij}|i, j, U, V, \sigma^2) = \mathcal{N}(X_{ij}|U_{i\cdot}V_{\cdot j}, \sigma^2) \tag{5.17}$$

假设U, V分别来自零均值的球面高斯分布：

$$p(U|\sigma_1^2) = \prod_{i=1}^{n}\mathcal{N}(U_{i\cdot}|\mathbf{0}, \sigma_U^2 I), \quad p(V|\sigma_1^2) = \prod_{j=1}^{d}\mathcal{N}(V_{j\cdot}|\mathbf{0}, \sigma_V^2 I). \tag{5.18}$$

而最大化U, V的后验概率等价于下面的最小化重构误差

$$\mathcal{L}_{\text{PMF}} = \sum_{i=1}^{n}\sum_{j=1}^{d}W_{ij}(X_{ij} - U_{i\cdot}V_{\cdot j})^2 + \lambda(||U||^2 + ||V||^2) \tag{5.19}$$

其中W_{ij}是指示变量用于表示(i,j)是否缺失，而$\lambda=\sigma^2/\sigma_1^2$。PMF可以通过梯度下降法计算局部最优解：

$$\frac{\partial \mathcal{L}_{\text{PMF}}}{\partial U_{i\cdot}}=\sum_{j\in \boldsymbol{j}_i}(U_{i\cdot}V_{\cdot j}-X_{ij})(V_{\cdot j})^T+\lambda U_{i\cdot} \tag{5.20}$$

$$\frac{\partial \mathcal{L}_{\text{PMF}}}{\partial V_{\cdot j}}=\sum_{i\in \boldsymbol{i}_j}(U_{i\cdot}V_{\cdot j}-X_{ij})U_{i\cdot}^T+\lambda V_{\cdot j} \tag{5.21}$$

5.2.2.2 I-PMF

与式子(5.19)中基于单值矩阵的标准PMF类似，我们可得下面的IMF问题：

$$\mathcal{L}_{\text{I-PMF}}=||X^{\text{low}}-UV^{\text{low}}||^2+||X^{\text{up}}-UV^{\text{up}}||_F^2+\lambda(||U||^2+||V^{\text{low}}||^2+||V^{\text{up}}||^2) \tag{5.22}$$

我们可以很容易的推导出下面的梯度下降公式来最小化式子(5.22)中的目标函数。

$$\frac{\partial \mathcal{L}_{\text{I-PMF}}}{\partial U_{i\cdot}}=\sum_{j\in \boldsymbol{j}_i}[(U_{i\cdot}V_{\cdot j}^{\text{low}}-X_{ij}^{\text{low}})(V_{\cdot j}^{\text{low}})^T+(U_{i\cdot}V_{\cdot j}^{\text{up}}-X_{ij}^{\text{up}})(V_{\cdot j}^{\text{up}})^T]+\lambda U_{i\cdot} \tag{5.23}$$

$$\frac{\partial \mathcal{L}_{\text{I-PMF}}}{\partial V_{\cdot j}^{\text{low}}}=\sum_{i\in \boldsymbol{i}_j}(U_{i\cdot}V_{\cdot j}^{\text{low}}-X_{ij}^{\text{low}})U_{i\cdot}^T+\lambda V_{\cdot j} \tag{5.24}$$

$$\frac{\partial \mathcal{L}_{\text{I-PMF}}}{\partial V_{\cdot j}^{\text{up}}}=\sum_{i\in \boldsymbol{i}_j}(U_{i\cdot}V_{\cdot j}^{\text{up}}-X_{ij}^{\text{up}})U_{i\cdot}^T+\lambda V_{\cdot j} \tag{5.25}$$

给定低维因子$U,V^{\text{low}},V^{\text{up}}$，我们就可以根据式子5.8来预测协同过滤应用中的评分。

5.3 实验结果

本章的实验分为两个部分，我们在第5.1.1节以人脸分析为应用对比本章提出的I-NMF和标准NMF，我们在第5.1.2节以协同过滤为应用对比本章提出的I-PMF和标准PMF。

5.3.1 I-NMF和NMF的对比

我们对比NMF和I-NMF在人脸识别、重构、聚类几个方面的性能。

5.3.1.1　实验设置

数据集本实验使用Olivertti Research Laboratory (ORL)提供的人脸数据集来测试NMF和I-NMF，ORL包含40个人的400张图片（每个人10张图片），我们分别使用分辨率为32×32（ORL32）和64×64（ORL64）两种数据，ORL32中的图像可以由1024个特征来表示，而ORL64的特征个数为4096。

对比算法我们分别实现了基于乘法更新公式的NMF和I-NMF。在利用式子5.4进行区间数据构造时，我们使用$r = 5$以及$\alpha = 2.5$。本文采用简单的最近邻分类器在矩阵分解获得的低维因子U上进行人脸识别，采用Kmeans在矩阵分解获得的低维因子U上进行人脸聚类。以原始数据X上的识别和聚类结果作为基准。人脸识别实验中，我们随机选取50%的样本做训练剩下的做测试，并重复该过程10次最后报告平均结果；而人脸聚类和重构在整体数据集上进行，由于NMF和I-NMF都是非凸方法，我们独立运行100次实验最后报告平均结果。

评价指标 本实验采用标准的F1指标评估人脸识别的效果，并采用第2.4.2节定义的ACC和NMI评价聚类结果；采用下面的重构误差（Reconstruction Error，简称RE）评估人脸重构的效果：

$$\mathrm{RE}(X, \hat{X}) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{d}(\hat{X}_{ij} - X_{ij})^2}{n \times d}}, \tag{5.26}$$

其中$\hat{X}$是重构单值数据，NMF可以直接由U, V乘积计算，而I-NMF可通过式子5.8计算。需要指出的是对于F1，ACC和NMI其值越大说明效果越好而对于RE数值越小结果越好。

5.3.1.2　实验结果

结果随非负因子维度k的变化 矩阵分解模型中因子的维度k对模型影响较大，然而一方面最佳的因子个数往往依赖于实际任务和数据集，另一方面也缺乏统一合理的方法来设置该参数，因此本实验给出了k取不同值$[20 : 2 : 40]$时的结果。图5.5中分别给出了不同算法在k取不同值时人脸识别、聚类和重构的实验结果。 可以看出I-NMF算法在ORL32和ORL64两个数据集上都显著优于标准NMF算法。

结果随区间半径的变化 本实验测试式子(5.4)中区间半径$\delta_{ij}^{\mathrm{FR}(r)}$对I-NMF算法的影响，图5.6中给出了半径系数$\alpha$取不同值时的结果，其中本实验因子维度

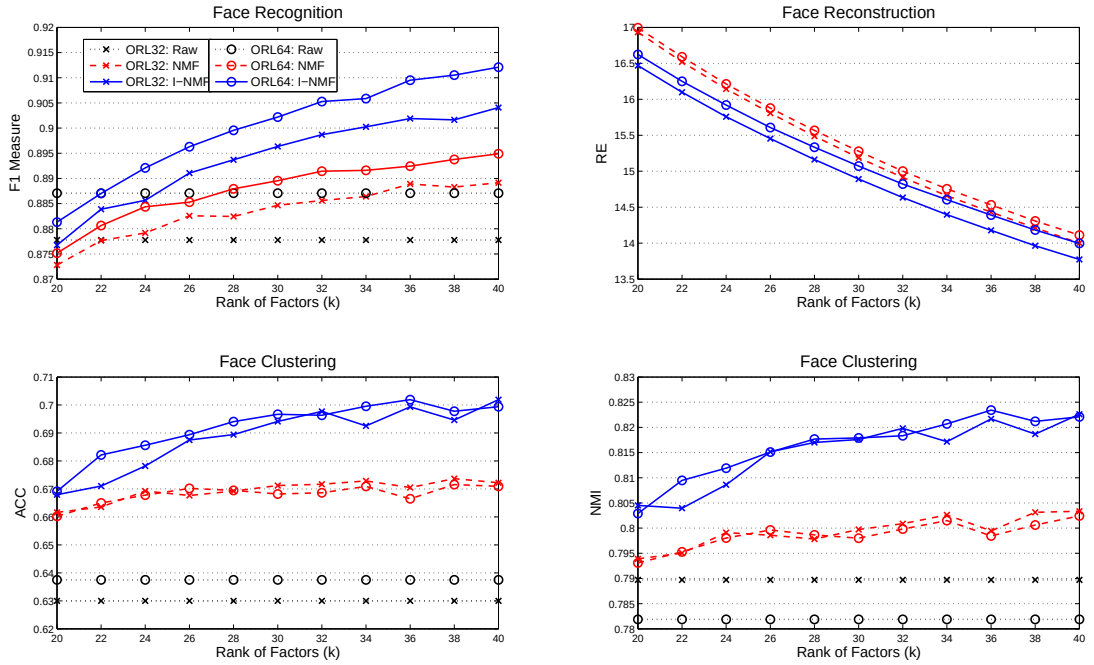


图 5.5: 第五章人脸分析实验中的结果随非负因子维度的变化曲线

设置为30，每个实验我们独立运行I-NMF算法20次并报告平均结果。 可以看出I-NMF在一个相对较大的系数范围内都优于标准非负矩阵分解。

5.3.2 I-PMF和PMF的对比

本实验对比PMF和I-PMF在协同过滤问题上的预测准确性。

5.3.2.1 实验设置

数据集 本实验使用协同过滤中常用的两个数据集。MovieLens[1]数据集包含用户对电影的评分，这里我们使用的子集MovieLens-100K是943个用户在1682个电影上的100，000个评分记录。Netflix[2]数据集是Netflix电影推荐竞赛中使用的数据集，这里我们随机选择一个子集Netflix-100K包括1319个用户在1999个电影上的100，000个评分记录。

[1] http://www.grouplens.org/system/files/ml-data_0.zip

[2] http://archive.ics.uci.edu.cn/datasets/Netflix+Prize

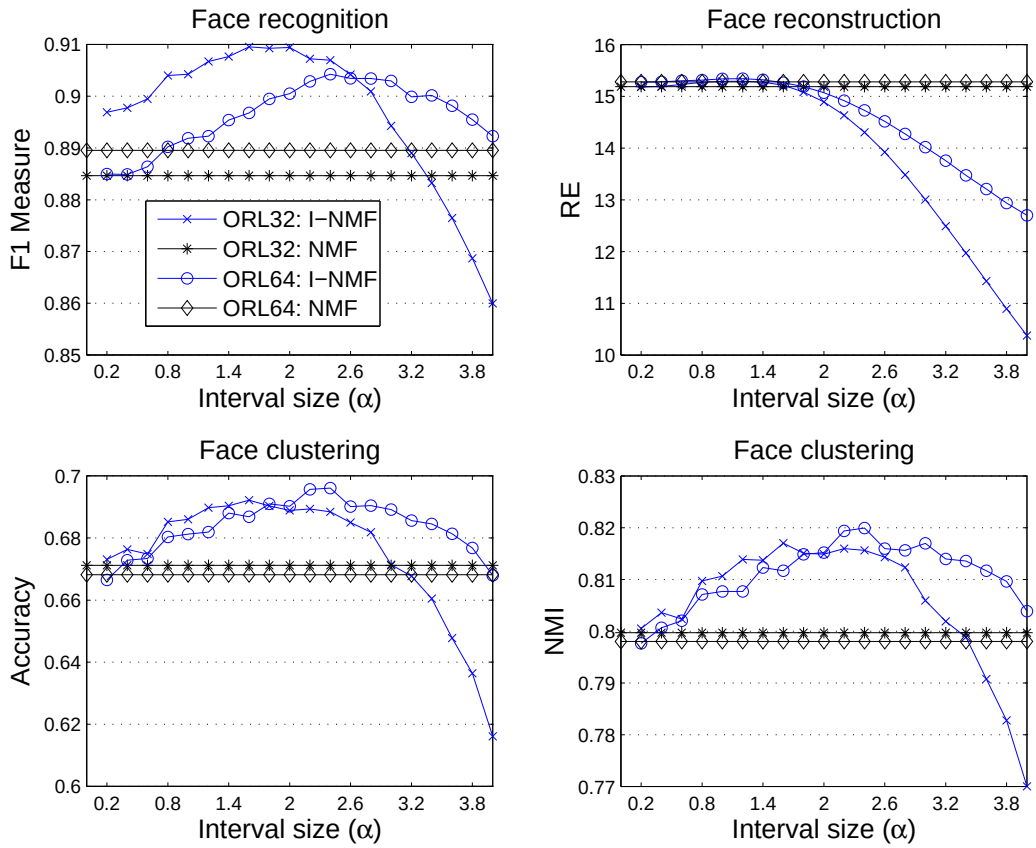


图 5.6: 第五章人脸分析实验中的结果随区间半径系数的变化曲线

对比算法 由于PMF,I-PMF都是基于模型(model based)的推荐方法，我们以常见的基于模型的推荐算法PLSA(Probabilistic Latent Semantic Analysis)[46]作为基准。原始数据被分成两个部分，训练集X^{train}和测试集X^{test}，而基于区间的数据由X^{train}构造，这里我们设$\alpha = 1$。

评价指标 本实验分别采用预测误差和排序准确性两类常见指标评估协同过滤算法的效果。预测误差包括MAE(Mean Absolute Error)和RMSE(Rooted

Mean Squared Error)，其定义如下：

$$\mathrm{MAE}(X^{\mathrm{test}}, \hat{X}^{\mathrm{test}}) = \frac{\sum_{X_{ij}\in X^{\mathrm{test}}} |X_{ij}^{\mathrm{test}} - \hat{X}_{ij}^{\mathrm{test}}|}{|X^{\mathrm{test}}|} \tag{5.27}$$

$$\mathrm{RMSE}(X^{\mathrm{test}}, \hat{X}^{\mathrm{test}}) = \sqrt{\frac{\sum_{X_{ij}\in X^{\mathrm{test}}} (X_{ij}^{\mathrm{test}} - \hat{X}_{ij}^{\mathrm{test}})^2}{|X^{\mathrm{test}}|}} \tag{5.28}$$

对某个用户u在一组电影上的的预测出用户对电影的偏好后，本实验使用NDCG(Normalized Discounted Cumulative Gain)来评估排序效果，其定义如下：

$$\mathrm{NDCG}_N(X^{\mathrm{test}}, \hat{X}^{\mathrm{test}}) = \frac{1}{|\boldsymbol{u}|} \sum_{u\in\boldsymbol{u}} Z_u \sum_{l=1}^{N} \frac{2^{\hat{X}_{um_l}} - 1}{\log(1+l)} \tag{5.29}$$

其中N是排序指标关注的位置，$\boldsymbol{u}$是一组用户，m_l是排序列表中第l个对应的电影，而Z_u是归一化系数，实验中我们评估前5个和前20个的排序效果，即$NDCG_5$和$NDCG_{20}$。需要指出的是MAE和RMSE越小说明结果越好而NDCG_N越大效果越好。

5.3.2.2 实验结果

结果随训练集比例的变化 数据缺失是协同过滤问题中面临的最大挑战，因此本实验在不同比例$(10\%, 20\%, \ldots, 90\%)$的训练集（对应不同比例的缺失率）上评估算法的效果。对于给定的训练集比例，我们随机独立运行6次实验，其中第一次用于参数选择，如PLSA中的隐变量个数和PMF/I-PMF模型中的正则参数，然后用给定的参数训练后面5次并记录平均结果。图5.7中分别给出了不同算法在不同比例训练集上的预测误差和排序准确性的实验结果。可以看出I-PMF算法在MovieLens-100K和Netflix-100K两个数据集上都显著优于标准PMF算法。

结果随区间半径的变化 本实验测试式子(5.6)中区间半径$\delta_{ij}^{\mathrm{CF}(r)}$对I-PMF算法的影响，图5.8中给出了半径系数$\alpha$取不同值时的结果，其中本实验中训练集比例设为50%，每个实验我们独立运行I-PMF算法50次并报告平均结果。可以看出I-PMF在$\alpha = 1$时要明显好于PMF，这也说明此时的区间大小已经能够很好的对评分记录进行近似。

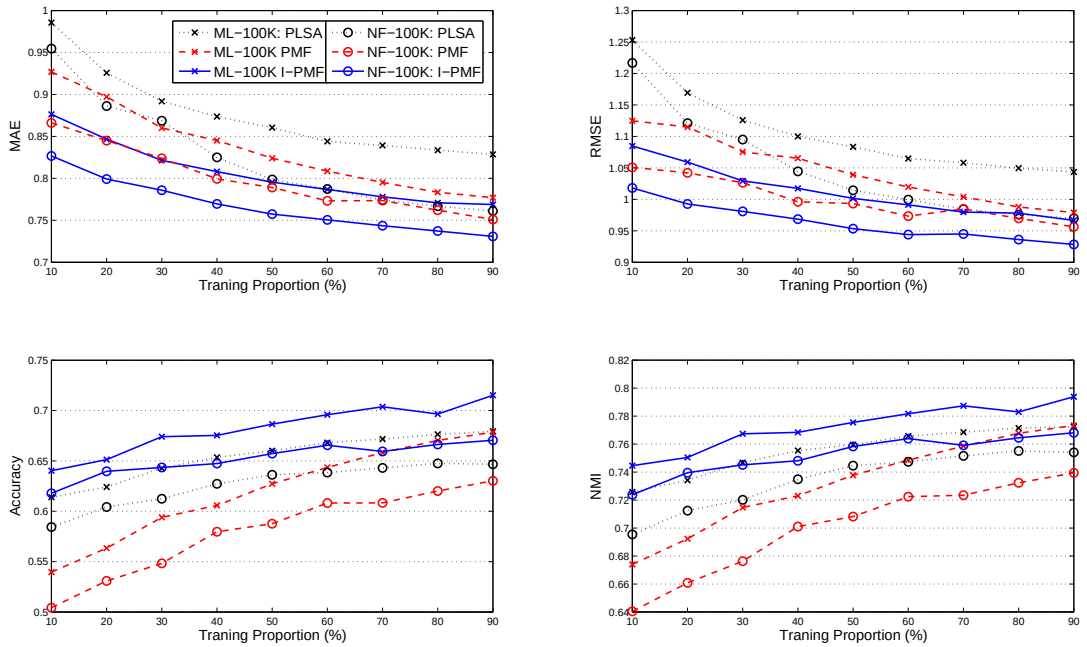


图 5.7: 第五章协同过滤实验中的结果随训练集比例的变化曲线

5.4　本章小结

由于真实观测数据中存在的系统误差和不确定性，本章提出用一个区间来近似和刻画真实数据可能的取值，此时数据可以由区间的上下界对应的均匀概率分布生成，为了分别近似数据的上下界矩阵，本章提出面向区间矩阵的联合矩阵分解方法。该方法分别学习上下界矩阵对应的基矩阵并同时学习一个共享的编码矩阵，这样上下界矩阵就可以分别由上下界基矩阵和共享的编码矩阵的乘积计算，这样可以更好的近似数据可能的取值区间并且更好的刻画实际数据。本章分析了两种数据挖掘应用的数据近似问题：人脸分析和协同过滤，提出了经验性方法来构造观测数据的可能的取值区间，得到其上下界矩阵。基于这些实际应用和数据区间，本章通过联合矩阵分解的方式分别扩展了标准的非负矩阵分解和概率矩阵分解，分别得到基于区间数据的非负矩阵分解模型(I-NMF)和基于区间数据的概率矩阵分解模型(I-PMF)。本章分别在人脸分析（包括人脸识别，聚类和重构）和协同过滤两个应用中系统性的比较了单值矩阵分解模型和对应的区间矩阵分解模型，大量实验结果表明本章提出的区间矩阵分解方法要显著优于标准的单值矩阵分解模型。

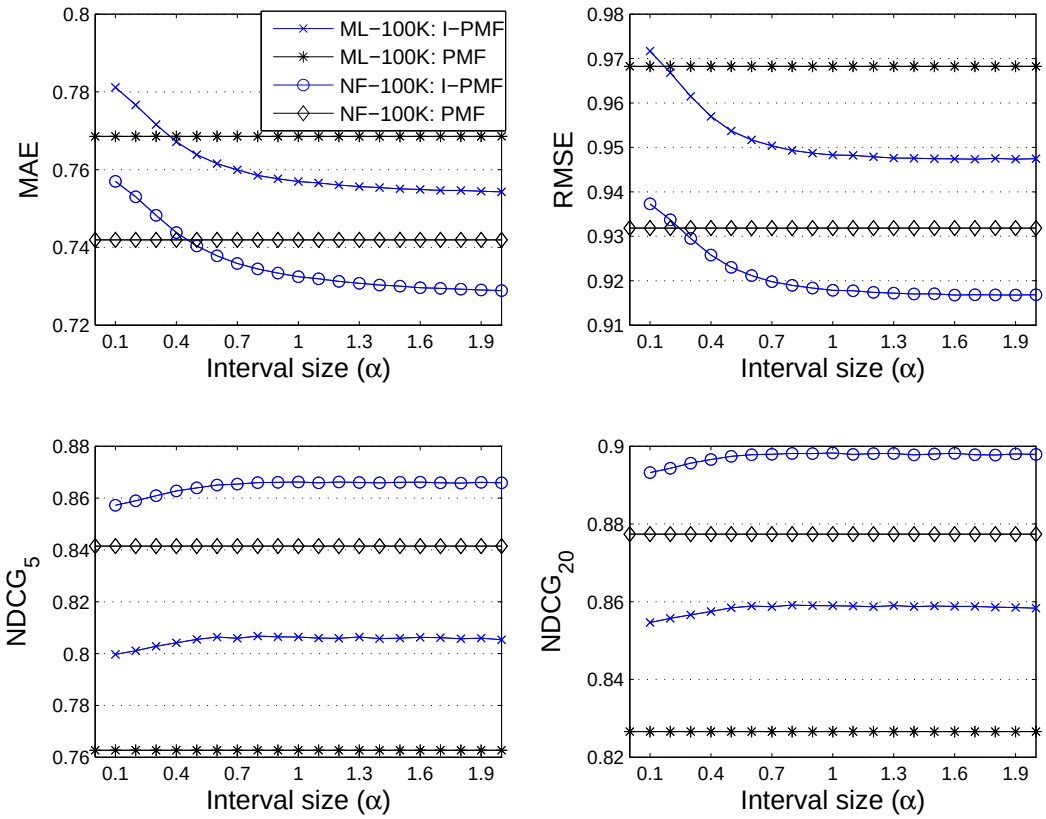


图 5.8: 第五章协同过滤实验中的结果随区间半径系数的变化曲线

第六章　基于加权图正则非负矩阵分解的聚类集成

聚类分析是一种无监督的学习方式，实际应用中对某一数据的聚类分析往往得到多种聚类划分，由于并没有一个统一的标准，为了找到数据中有意义的划分就需要从多种聚类划分中学习一个最终的聚类划分。聚类集成[93](Clustering ensemble)通常是指将多个聚类划分(base clusterings)最终合并成一个聚类结果(final clustering)。和聚类分析相比，聚类集成往往可以从大量聚类结果中得到一个一致的、稳定的、鲁棒的聚类划分，可以提高聚类分析的效果。从聚类的角度看，聚类集成可以看成是以多个聚类结果作为输入的一个新的聚类问题。因此，聚类集成同样面临聚类问题本身面临的困难和挑战。从集成学习的角度看，与分类集成不同，聚类集成问题依然是一种无监督的学习，并没有统一的标准，因此无法利用监督信息指导聚类集成。

聚类集成相关文献研究两方面的问题：1）如何从同一个数据集中生成多种划分结果(base clusterings)?2）如何合并这些划分结果最后生成高质量的、稳定的、一致的聚类(consensus clustering)？对同一数据集的多种划分可以通过很多方式获得：同一个算法随机运行多次（许多聚类算法需要随机指定初值）；同一个算法不同的参数设置；分别利用不同的算法生成；在不同的特征子空间生成等。本章中，我们仅考虑第二个问题，也就是说我们的输入数据是一组划分结果，而用于生成多种划分结果的算法以及数据的原始表示是未知的。

聚类集成方法研究的重点又包括：1）如何根据输入的一组划分，构造合适的数据表示形式?2）如何构造或者学习一致性函数(concensus function)？这里一致性函数是指函数的输出可以产生单个聚类划分，因为输入数据是一组聚类划分，我们称聚类集成得到的是一致性划分，一致性函数可以是显式的也可以是隐式的(implicit)函数。对聚类集成输入数据的表示主要分为两种：一种是基于簇特征(cluster-based features)的数据表示[93]，也就是用每个簇作为一个特征，这种表示可以反映样本是否属于某个簇，类似于用文本中出现的单词组成的向量(bag of words)表示文本数据；另一种是基于多重关联关系的数据表示[99](multiple co-association matrices)，也就是每一个聚类划分构造一个关联矩阵反映样本之间的关系，如两个样本是否属于同一个簇。现有的方法主要利用其中一种形式的数据表示构造和学习一致性函数。对于聚类集成任务而言，

针对这两种表示形式分别构造的一致性函数之间有着内在的相关性（目的相同）。如：如果在基于多重关联矩阵的一致性关系中两个样本关联度高，那么在基于簇特征构造的一致性函数中这两个样本的关联度也应该比较高；反之亦然。由于聚类集成问题中没有没有统一的标准，如何利用这两种表示形式在聚类集成任务上的内在相关性提高一致性学习的效果是聚类集成研究的一个重要问题。

本章中我们利用非负矩阵分解学习基于簇特征数据表示的一致性函数，也就是矩阵分解得到的非负低维因子；此外，我们通过非负线性加权的多重关联关系构造一致的关联关系，这里我们需要学习线性权重。基于这两种一致性函数在聚类集成任务上的内在相关性，我们利用联合学习的思想提出基于加权图正则非负矩阵分解的聚类集成算法。在该算法中，当多重关联关系权重给定时，我们得到一致的关联关系，利用基于该关系的图正则非负矩阵分解提高非负因子的一致性（基于非负因子的聚类质量）；而当非负因子给定时，我们利用非负因子和每个关联关系的匹配度来估计每个关系对应的加权系数。可以看出基于簇特征和多重关联关系的一致性函数学习在迭代过程中互相提高，因此既可以提高非负因子上的聚类集成(final clustering)效果，又可以利用线性权重选择好的输入聚类(base clustering)。在大量数据集上的实验也验证了本算法在这两个方面的优势。

6.1 相关工作

本节我们简要介绍聚类集成的相关工作。当前的主要方法分别通过聚类集成数据表示的两种形式构造和学习一致性函数。

基于簇特征数据表示的一致性函数学习方法主要分为两类。第一类是基于图的方法，这类方法首先根据基于簇特征的表示构造一个关系图，然后利用图切割的方法生成最终的一致性聚类划分。代表性工作包括：Strehl等人在文献[93]中提出三种图方法包括基于簇相似度的划分算法（Cluster-based Similarity Partition Algorithm，简称CSPA），超图划分算法（Hyper-Graph Partition Algorithm，简称HGPA）和meta-clustering算法（MCLA）。其中CSPA构造一个二元相似关系图并用METIS[50]进行图分割；HGPA定义一个超图，其中每一个超边(hyperedge)表示一个簇，然后利用HMETIS[50]算法对超图进行分割。Fern等人在文献[33]中提出利用二部图的数据表示，并利用谱方法进行图分割。

Al-Razgan等提出划分一个加权相似度图[3]。第二类是基于概率模型的方法，这类方法用概率分布刻画样本数据的生成过程（类似于文本挖掘中的图模型）。Topchy等人在文献[95, 96]中提出基于混合多项分布的集成方法；而Wang等人在文献[100]中提出类似于隐形狄利克雷分布（Latent Dirichlet Allocation，简称LDA）[10]的生成概率模型进行聚类集成。

基于多重关联关系的一致性函数学习主要分为两步，首先将多重关系合并为一个关联关系（既可以看成是基于关系的表示，也可以看成是基于特征的表示），然后利用现有的聚类方法生成最终的一致性聚类。主要方法包括基于平均关联关系的学习和加权关联关系的学习。首先根据多重关联关系构造平均关联关系（此时每个输入的聚类划分权重相同），然后利用Kmeans[41]或者NMF[64]来进行聚类集成。为了更准确的刻画不同关联关系对一致性函数的影响，文献[63]在NMF框架下学习加权的关联关系；文献[99]提出学习基于Bregman失真度的加权关联关系，然后利用谱聚类对加权关联关系进行分割。

6.2　预备知识

本节介绍聚类集成问题中的数据表示方法和非负矩阵分解方法及其在聚类集成中的应用。

6.2.1　聚类集成中的两种数据表示

给定一个包含n个样本的数据集和一组m个聚类方案$\boldsymbol{\mathcal{P}} = \{\mathcal{P}^1, \mathcal{P}^2, \ldots, \mathcal{P}^m\}$。每个聚类方案是对$n$个数据的一种划分结果$\mathcal{P}^c, c = 1 \ldots m$，我们可以用$\{\mathcal{C}_1^c \ldots \mathcal{C}_k^c\}$或者标签向量$\mathcal{P}^c \in \mathbb{R}^n$表示将$n$个数据划分为$k$个簇的聚类方案，需要指出的不同划分$P^c$中簇的个数可能是不一样的。这样我们就可以构造基于簇特征的数据表示[93]：对于每个聚类划分$\mathcal{P}^c$，我们可以构造二元成员指示矩阵$H^c \in \mathcal{R}^{n \times k}$，其中$H^c$每一列对应一个簇而每一行对应一个样本，并且如果第$i$个样本在第$c$个划分中属于第$j$个簇，我们有$H_{ij}^c = 1$，否则$H_{ij}^c = 0$。将所有指示矩阵连起来就得到基于簇特征的数据表示，即$X = (H^1, H^2, \cdot, H^m)$，矩阵$X$的每一行对应样本而每一列对应所有划分中某个簇，$X$可以看成对聚类集成输入数据新的表示方式。此外，我们还可以根据划分$\mathcal{P}^c$构造一个$n \times n$的关联矩阵W^c，如果样本i和样本j被划分到同一个簇则$W_{ij}^c = 1$，否则$W_{ij}^c = 0$。我们给出这两种聚类集成数

据表示的例子6.1。

例子6.1.
假设我们有5个样本和2个聚类结果，每个聚类分别将这些样本聚成3个簇。此时，我们有2个划分$\mathcal{P}^1 = [1,1,2,3,3]$和$\mathcal{P}^2 = [2,3,3,1,1]$。下面给出了这些样本基于簇特征的数据表示$X$和多重关联矩阵表示，即基于第一个划分$\mathcal{P}^1$的关联矩阵$W^1$和基于第二个划分$\mathcal{P}^2$的关联矩阵$W^2$。

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad W^1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad W^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

6.2.2 图正则非负矩阵分解

我们首先简单回顾前面几章介绍的非负矩阵分解，这里我们分别给出了以平方欧式距离和KL失真度作为损失函数的非负矩阵分解：

$$\mathcal{L}_{NMF-Euc}(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij} - (UV^T)_{ij})^2 \tag{6.1}$$

$$\mathcal{L}_{NMF-KL}(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij}\log\frac{X_{ij}}{(UV^T)} - X_{ij} + (UV^T)_{ij}) \tag{6.2}$$

上面的非负矩阵分解模型仅考虑样本和特征关系的重构误差，忽略了数据中潜在的结构信息，如流形结构。流形学习认为数据通常分布在嵌入于外围欧式空间的一个潜在的流形体上。实际应用中我们可以定义数据之间关联性的函数，构造数据间的邻接关系图，并用此关系图近似数据分布流形的真实结构。最近，Cai等人在文献[14]中提出了图正则的非负矩阵分解模型和算法（Graph regularized Nonnegative Matrix Factorization，简称GNMF）。图正则的基本思想是希望矩阵分解得到的低维因子能够尽可能的刻画和保持由邻接关系图反映的数据间的流形结构。也就是说，如果样本i和样本i'关联度高（也可以认为是样本与样本在流形上的相近程度），我们希望学习得到的低维因子中对应的$U_{i\cdot}$和$U_{i'\cdot}$距离较近（相似度高）。文献[14]中针对平方欧式距离和KL散度的非

负矩阵分解提出不同的图正则模型包括：GNMF-Euc和GNMF-KL定义如下：

$$\mathcal{L}_{GNMF-Euc}(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij}-(UV^T)_{ij})^2+\sum_{i=1}^{n}\sum_{i'=1}^{n}W_{ij}||U_{i\cdot}-U_{i'\cdot}||^2 \quad (6.3)$$

$$\mathcal{L}_{GNMF-KL}(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij}\log\frac{X_{ij}}{(UV^T)}-X_{ij}+(UV^T)_{ij}) \quad (6.4)$$

$$+\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{l}W_{ij}(U_{il}\log\frac{U_{il}}{U_{i'l}}+U_{i'l}\log\frac{U_{i'l}}{U_{il}}) \quad (6.5)$$

6.3　基于加权图正则非负矩阵分解的聚类集成

6.3.1　问题形式化

我们在第6.2.1节定义了基于簇特征的数据表示，即矩阵$X=(H^1,H^2,\ldots,H^m)$。本章中我们用非负矩阵分解学习两个非负因子U和V，并用U,V的乘积近似$X\approx UV^T$，基于非负矩阵分解和聚类问题之间的关系，本章采用非负因子U作为聚类集成最终的一致性函数（给定U之后我们就可以得到最终的一致性聚类），其中V可以理解代表簇(meta-cluster)。

和GNMF类似，我们也可以通过图正则的方法得到更好的聚类集成结果。这里我们面临的主要问题是：由于聚类集成的输入数据不包含原始数据，如何定义和准确构造邻接关系图刻画数据间的流形结构？考虑到我们可以从多个聚类方案中构造多重关联关系，我们利用这些关系来构造一个合适的邻接关系图。然而一方面我们无法准确的判断和选择哪个聚类方案生成的关联关系最好；另一方面由于这些关联关系图是通过聚类算法得到的，可能所有这些关联关系图都无法准确刻画数据间的真实关系。为此本文通过线性合并多重关联关系来生成一个一致性关系图$\hat{W}$用于刻画样本间的真实关系，即：

$$\hat{W}=\sum_{c=1}^{m}\boldsymbol{\alpha}_c W^c,\quad \sum_{c=1}^{m}\boldsymbol{\alpha}_c=1,\boldsymbol{\alpha}_c\geq 0, \quad (6.6)$$

其中$\boldsymbol{\alpha}$是线性合并系数。给定上面的一致性关系定义我们还需要解决如何得到混合系数$\boldsymbol{\alpha}$，可以看出当$\boldsymbol{\alpha}$中只有一个元素为1其他元素都为零时，一致性关系等于从多重关联关系中选择一个；而当$\boldsymbol{\alpha}_c=\frac{1}{m}$时一致性关系等于平均多重关联关系。

在聚类集成问题中，标准非负矩阵分解生成的一致性函数U和一致性关联关系$\hat{W}$有着很强的内在关联性，即：如果在一致性关系中样本i和样本j关联度高，那么在一致性函数中样本i和样本j关联度也应该比较高；反之，如果如果在一致性函数中样本i和样本j关联度高，那么在一致性关系中样本i和样本j关联度也应该比较高。为此，本章采用联合学习的方式同时估计一致性函数U和一致性关联关系$\boldsymbol{\alpha}$，具体来讲我们提出下面的基于一致性关联关系和一致性函数U的加权图正则函数：

$$\mathcal{R}_{\text{Euc}} = \frac{1}{2}\sum_{c=1}^{m}\sum_{i=1}^{n}\sum_{i'=1}^{n}\boldsymbol{\alpha}_c W_{ii'}^c||U_{i\cdot} - U_{i'\cdot}||^2 \tag{6.7}$$

$$= \sum_{c=1}^{m}\boldsymbol{\alpha}_c(\sum_{i} D_{ii}^c U_{i\cdot}U_{i\cdot}^T - \sum_{ii'} W_{ii}^c U_{i\cdot}U_{i'\cdot}^T) \tag{6.8}$$

$$= \sum_{c=1}^{m}\boldsymbol{\alpha}_c(\text{tr}(U^T D^c U) - \text{tr}(U^T W^c U)) = \text{tr}(U^T(\sum_{c=1}^{m}\boldsymbol{\alpha}_c L^c)U) \tag{6.9}$$

$$\mathcal{R}_{\text{KL}} = \frac{1}{2}\sum_{c=1}^{m}\sum_{i=1}^{n}\sum_{i'=1}^{n}\boldsymbol{\alpha}_c W_{ii}^c(\text{kl}(U_{i\cdot}||U_{i'\cdot}) + \text{kl}(U_{i'\cdot}||U_{i\cdot})) \tag{6.10}$$

$$= \sum_{c=1}^{m}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{l}\boldsymbol{\alpha}_c W_{ii}^c(U_{il}\log\frac{U_{il}}{U_{i'l}} + U_{i'l}\log\frac{U_{i'l}}{U_{il}}) \tag{6.11}$$

其中$D_{ii}^c = \sum_{i'} W_{ii'}^c, c = 1, \ldots, m$是关联矩阵对应的度矩阵，$L^c = D^c - W^c$是图拉普拉斯矩阵。kl表示KL失真度。

给定上面的加权图正则函数，我们就可以定义加权图正则非负矩阵分解模型如下：

$$\begin{aligned}\mathcal{L}_{\text{WGNMF-Euc}} = \quad & ||X - UV^T||^2 + \lambda\text{tr}(U^T(\sum_{c=1}^{m}\boldsymbol{\alpha}_c L^c)U) \\ \text{s.t.} \quad & U \geq 0, V \geq 0, \sum_{c=1}^{m}\boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0\end{aligned} \tag{6.12}$$

如果我们使用KL失真度损失，WGNMF变成最小化下面的目标函数：

$$\begin{aligned}\mathcal{L}_{\text{WGNMF-KL}} &= \sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij}\log\frac{X_{ij}}{\sum_{l=1}^{k}U_{il}V_{jl}} - X_{ij} + \sum_{l=1}^{k}U_{il}V_{jl}) \\ &+ \frac{\lambda}{2}\sum_{c=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{l=1}^{k}\boldsymbol{\alpha}_c W_{ij}^{c}(U_{il}\log\frac{U_{il}}{U_{jl}} + U_{jl}\log\frac{U_{jl}}{U_{il}}) \\ \text{s.t.}\quad & U\geq 0, V\geq 0, \sum_{c=1}^{m}\boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0 \end{aligned} \tag{6.13}$$

其中$\lambda \geq 0$是正则参数用于控制非负因子U的平滑程度。

本章提出的WGNMF和GNMF的关系可以解释为：GNMF的目标函数是仅采用单个图的正则方法，因此GNMF可以看成WGNMF的特例（即$\boldsymbol{\alpha}$中只有一个元素为1其他都为零）；而WGNMF基于一致性函数U和一致性关联关系$\hat{W}$在聚类集成任务中的内在一致性，采用联合学习的方式同时估计U和$\hat{W}$。

文献[35, 62, 103]将多个关系图线性合并成一个关系图的方法称为集成流形正则(Ensemble Manifold Regularization, 简称EMR)，因子我们也将本章的工作描述为集成流形正则的非负矩阵分解。

6.3.2　学习算法

在本节中，我们将探讨如何求解式子(6.12)和式子(6.13)中的优化问题。可以很容易的发现目标函数是关于(U, V, α)的非凸优化问题，是分别关于单个变量的凸问题。由于没有闭式解，找到全局最小值是不现实的。和前面几章的求解思路类似，我们采用分块交互方法来优化。当目标函数中两个变量给定时，关于第三个变量的子问题很容易求解。我们分别给出$\mathcal{L}_{\text{WGNMF-Euc}}$和$\mathcal{L}_{\text{WGNMF-KL}}$的计算过程。

6.3.2.1　最小化$\mathcal{L}_{\text{WGNMF-Euc}}$

给定$\boldsymbol{\alpha}$，计算U, V 当$\boldsymbol{\alpha}$给定时，利用矩阵性质$\text{tr}(X) = \text{tr}(X^T)$以及$\text{tr}(YX) = \text{tr}(XY)$，并移除与$U, V$不相关的项，关于$U$, V的子优化问题简化为：

$$\min_{U,V} \quad -2\text{tr}(X^TUV^T) + \text{tr}(VU^TUV^T) + \lambda\text{tr}(U^TLU) \tag{6.14}$$

$$\text{s.t.} \quad U\geq 0, V\geq 0, \tag{6.15}$$

令ϕ_{ik}和ψ_{kj}分别是非负约束$U_{ik} \geq 0$, $V_{kj} \geq 0$对应的拉格朗日辅助变量，引入$\Phi = [\phi_{ik}], \Psi = [\psi_{kj}]$，问题(6.14)的拉格朗日函数为

$$\mathcal{L} = -2\text{tr}(X^T U V^T) + \text{tr}(V U^T U V^T) + \lambda \sum_{c=1}^{m} \boldsymbol{\alpha}_c \text{tr}(U^T L^c U) + \text{tr}(\Phi U^T) + \text{tr}(\Psi V^T) \tag{6.16}$$

函数$\mathcal{L}$关于基矩阵V和编码矩阵U的偏导可以分别写为

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV + 2UV^T V + 2\lambda(\sum_{c=1}^{m} \boldsymbol{\alpha}_c D^c)U - 2\lambda(\sum_{c=1}^{m} \boldsymbol{\alpha}_c W^c)U + \Phi \tag{6.17}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2X^T U + 2VU^T U + \Psi \tag{6.18}$$

利用KKT条件$\phi_{ik}U_{ik} = 0$和$\psi_{kj}V_{kj} = 0$，我们可以得到下面的两个等式：

$$-(XV + \lambda(\sum_{c=1}^{m} \boldsymbol{\alpha}_c W^c)U)_{ik}\Phi_{ik} + (UV^T V + \lambda(\sum_{c=1}^{m} \boldsymbol{\alpha}_c D^c)U)_{ik}\Phi_{ik} = 0 \tag{6.19}$$

$$-(X^T U)_{kj}\Psi_{kj} + (VU^T U)_{kj}\Psi_{kj} = 0 \tag{6.20}$$

由这些等式可以得到下面的更新公式：

$$U_{il} = U_{il}\frac{(XV + \lambda(\sum_{c=1}^{m} \boldsymbol{\alpha}_c W^c)U)_{il}}{(UV^T V + \lambda(\sum_{c=1}^{m} \boldsymbol{\alpha}_c D^c)U)_{il}} \tag{6.21}$$

$$V_{il} = V_{il}\frac{(X^T U)_{il}}{(VU^T U)_{il}} \tag{6.22}$$

给定U, V，计算$\boldsymbol{\alpha}$ 当U和V给定时，问题$\mathcal{L}_{\text{WGNMF-Euc}}$简化为关于$\boldsymbol{\alpha}$的优化问题：

$$\min_{\boldsymbol{\alpha}} \quad \sum_{c=1}^{m} \boldsymbol{\alpha}_c \text{tr}(U^T L^c U), \text{ s.t. } \quad \sum_{c=1}^{m} \boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0 \tag{6.23}$$

上面的问题是一个线性优化问题，其最优解是：

$$\boldsymbol{\alpha}_c = \begin{cases} 1, & \text{if } c = \arg\min_c \text{tr}(U^T L^c U) \\ 0, & \text{otherwise} \end{cases} \tag{6.24}$$

此时$\boldsymbol{\alpha}$一个极端稀疏的情况，为了避免该平凡解，我们提出在式子6.23中增加正则项，也就是求解下面的问题：

$$\min_{\boldsymbol{\alpha}} \quad \sum_{c=1}^{m} \boldsymbol{\alpha}_c \text{tr}(U^T L^c U) + \lambda_2 ||\boldsymbol{\alpha}||^2, \text{ s.t. } \quad \sum_{c=1}^{m} \boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0 \tag{6.25}$$

其中$\lambda_2 \geq 0$是正则化参数，当$\lambda_2 = 0$原问题有平凡解，当$\lambda_2 \to \infty$时，每个关联关系得到相同的权重，这同样不是我们期望的，因此我们需要设定一个合理的正则参数。这样当U, V给定时我们就需要求解一个关于$\boldsymbol{\alpha}$的二次优化问题。该二次优化问题一般可以通过下面三种方法求解。第一，采用通用二次优化工具包（如CVX[11]），当变量较多时该求解器比较耗时收敛较慢；第二，该问题是一个单形体约束的凸优化问题，当变量个数较大时，可以用EMDA算法[7]来计算。第三，当变量个数很大时，我们可以利用文献[62]中提出的类似于支持向量机序列最小化的分块下降法来优化。

6.3.2.2　最小化$\mathcal{L}_{\text{WGNMF-KL}}$

给定$\boldsymbol{\alpha}$，计算U, V 当$\boldsymbol{\alpha}$给定时，关于U, V的优化问题变成：

$$\begin{aligned}\mathcal{L} = &\sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij}\log\frac{X_{ij}}{\sum_{l=1}^{k}U_{il}V_{jl}} - X_{ij} + \sum_{l=1}^{k}U_{il}V_{jl}) \\ &+ \frac{\lambda}{2}\sum_{c=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{l=1}^{k}\boldsymbol{\alpha}_c W_{ij}^{c}(U_{il}\log\frac{U_{il}}{U_{jl}} + U_{jl}\log\frac{U_{jl}}{U_{il}}) \\ &+ \mathrm{tr}(\Phi U^T) + \mathrm{tr}(\Psi V^T)\end{aligned} \tag{6.26}$$

此时，该问题是基于混合拉普拉斯图正则的GNMF-KL问题，我们可以采用类似于文献[]中的方法来求解，这里我们给出更新U时用到的等式系统：

$$\left(\sum_j v_{jl}I + \lambda(\sum_c L^c)\right)U_l = \begin{bmatrix} u_{1l}\sum_j(x_{1j}v_{jl}/\sum_l u_{1l}v_{jl}) \\ u_{2l}\sum_j(x_{2j}v_{jl}/\sum_l u_{2l}v_{jl}) \\ \cdots \\ u_{nl}\sum_j(x_{nj}v_{jl}/\sum_l u_{nl}v_{jl}) \end{bmatrix} \tag{6.27}$$

该等式可以通过矩阵求逆运算或者一些快速迭代算法来求解。变量V可以由下面的式子计算：

$$V_{jl} = V_{jl}\frac{\sum_i(X_{ij}U_{il}/\sum_l U_{il}V_{jl})}{\sum_i U_{il}} \tag{6.28}$$

给定U, V，计算$\boldsymbol{\alpha}$ 当U, V给定时，我们可以得到关于$\boldsymbol{\alpha}$类似的二次优化问

题：

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \sum_{c=1}^{m} \boldsymbol{\alpha}_c \left[\sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{l=1}^{k} W_{ij}^{c} (U_{il} \log \frac{U_{il}}{U_{jl}} + U_{jl} \log \frac{U_{jl}}{U_{il}}) \right] + \lambda_2 ||\boldsymbol{\alpha}||^2 \\ \text{s.t.} \quad & \sum_{c=1}^{m} \boldsymbol{\alpha}_c = 1, \boldsymbol{\alpha}_c \geq 0 \end{aligned} \tag{6.29}$$

Algorithm 8 基于加权图正则非负矩阵分解的聚类集成算法

输入： 聚类划分$\{\mathcal{P}^i\}_{i=1}^m$，聚类集成中簇的个数$k$，图正则参数$\lambda$和二次优化正则参数$\lambda_2$

输出： 聚类集成结果V和多重关联关系的混合系数$\boldsymbol{\alpha}$

1: 分别构造m个关联矩阵$\{W^c\}_{c=1}^m$和基于簇特征的数据表示X
2: 初始化U, V，并令$\boldsymbol{\alpha} = [1/m, 1/m, \ldots, 1/m]^T$
3: **repeat**
4: 根据$\boldsymbol{\alpha}$计算加权的加权图正则拉普拉斯，
5: 根据式子(6.21)（对应于$\mathcal{L}_{\text{WGNMF-Euc}}$）或者式子(6.27)（对应于$\mathcal{L}_{\text{WGNMF-KL}}$）计算$U_{il}$，
6: 根据式子(6.22)（对应于$\mathcal{L}_{\text{WGNMF-Euc}}$）或者式子(6.28)（对应于$\mathcal{L}_{\text{WGNMF-KL}}$）计算$U_{il}$，
7: 通过二次优化求解器求解式子(6.25)（对应于$\mathcal{L}_{\text{WGNMF-Euc}}$）或者式子(6.29)（对应于$\mathcal{L}_{\text{WGNMF-KL}}$）中的二次优化问题，计算最优解$\boldsymbol{\alpha}$，
8: **until** convergence

上述算法每一步都使目标函数单调下降，并且目标函数有下界，所以算法8收敛。

6.4 实验结果

本节中我们在大量真实数据集上评估本章提出的聚类集成算法的有效性。

6.4.1 实验设置

为了生成聚类集成问题的输入数据，和文献[100]相同，我们随机运行Kmeans算法200次将每20个聚类结果分成一组作为聚类集成问题的输入，这样我们运行聚类集成算法10次并且列出其10次的平均结果。

为了说明本章提出的聚类集成方法WGNMF的优势，我们对比下面的主流聚类集成方法：

- Kmeans，代表性的划分聚类算法，也是聚类集成输入数据的生成算法，
- 基于一致性关联矩阵（Consensus matrix）的Kmeans (KC)，其中一致性关联矩阵定义为多重关联关系的平均值，
- 文献[93]分别提出了基于簇相似度的划分算法（Cluster-based Similarity Partitioning Algorithm，简称CSPA），超图划分算法（Hyper Graph Partitioning Algorithm，简称HGPA）和and MetaClustering Algorithm (MCLA)三种算法。我们使用作者提供的Matlab工具包ClusterPack[1]。
- Bayesian Cluster Ensembles (BCE)是文献[100]中提出的一个生成概率模型，该模型同时学习基于特征的数据表示和隐式的一致性函数，
- 文献[99] 通过加权平均的关联关系作为一个一致性函数提出了Generalized Weighted Cluster Aggregation (GWCA) 算法，这里我们使用欧式距离来学习加权的一致性函数，并用谱聚类算法[2]来生成最终的集成聚类，

对于本章提出的聚类集成算法WGNMF，我们分别给出了基于欧式距离和KL失真度的加权图正则非负矩阵分解的结果（WGNMF-Euc和WGNMF-KL）。

6.4.2　数据集

我们使用多种公开数据集来测试聚类集成算法的效果。数据中簇的个数从3个到40个，样本的个数从47个到2340个，而特征维度从4维到21839。这些数据集包括：

- 5个UCI[3]数据集（Iris, Glass, Ecoli, Soybean, Zoo）
- COIL是Columbia提供的一个包含20个物体的图像数据库，每个对象旋转5度拍摄一张图片，每个物体共有72张图片。

[1] www.lans.ece.utexas.edu/~strehl
[2] http://www.cis.upenn.edu/~jshi/software/
[3] http://archive.ics.uci.edu/ml/datasets.html

- ORL是Olivetti实验室提供的人脸数据库，包含40个人的400张图片（每个人10张图片），这些图片包括脸部多种表情（如睁眼/闭眼，笑和不笑，带不带眼镜等）。
- WebACE是WebACE项目中使用的一些文档，包括2340篇文章，这些文章来自于路透新闻的20个类别中。
- Tr11, Tr12是TREC[4]提供的两个文本数据集。

这些数据集常被用于测试聚类和聚类集成方法的有效性。数据概要信息如表6.1所示。

表 6.1: 第六章实验所使用的数据集

数据集	样本个数	特征维度	类别个数
Iris	150	4	3
Glass	214	9	6
Ecoli	336	7	8
Soybean	47	35	4
Zoo	101	18	7
COIL	1440	1024	20
ORL	400	1024	40
WebACE	2340	21839	20
Tr11	414	6429	9
Tr12	313	5804	8

6.4.3 实验结果

我们使用第2.4.2节定义的聚类准确性（ACC）和归一化互信息（NMI）来评价聚类和聚类集成算法的好坏。

聚类集成结果 表6.2、6.3和表6.4、6.5中列出了这些算法在不同数据集的归一化互信息结果。从这些结果中我们可以看出本章提出的WGNMF算法在所

[4] http://trec.nist.gov

表 6.2: 第六章的聚类准确性

	Iris	Glass	Ecoli	Soybean	Zoo
Kmeans	0.81	0.42	0.65	0.72	0.69
KC	0.85	0.49	0.64	0.65	0.67
CSPA	0.87	0.43	0.56	0.69	0.58
HGPA	0.62	0.40	0.51	0.72	0.55
MCLA	0.89	0.46	0.61	0.73	0.74
BCE	0.89	0.49	0.66	0.73	0.74
GWCA	0.89	0.53	0.64	0.73	0.74
WGNMF-Euc	0.89	0.54	0.67	0.75	0.77
WGNMF-KL	0.89	0.51	0.65	0.73	0.73

有数据集上的结果都好于Kmeans，这说明聚类集成算法通过合并多种聚类结果可以有效提高聚类算法的效果和稳定性。此外，WGNMF (WGNMF-Euc或者WNGMF-KL)在其中9个数据集上表现最好而在另外一个数据集也接近最好结果。总体来讲，加权图正则非负矩阵分解能够有效的提高聚类集成的效果（如准确性和归一化互信息结果）。

单个聚类的质量评估 WGNMF在聚类的同时学习多个关联关系的混合系数，而这些权重系数反映了单个关联关系和全部关联关系的一致性，因此我们可以用学习到的权重α来评估聚类集成中多个输入聚类的好坏。

根据WGNMF和GWCA [99]（GWCA也是一种加权的一致性聚类方法，也学习多个聚类的权重）学习得到的权重系数，图6.1给出了前5个权重对应的平均聚类结果和20个聚类的平均结果。可以看出，权重较大的聚类（input base clustering）效果往往比较好。

参数敏感性 本章提出的WGNMF中的正则参数λ用于平衡重构误差和加权图正则误差。表6.6列出了WGNMF在λ取不同值$(0.1, 1, 10, 100)$的聚类集成结果。可以看出，WGNMF在这些参数上的结果相对稳定，这也说明WGNMF对图正则参数不敏感。

6.5　本章小结

本章中我们利用非负矩阵分解学习基于簇特征数据表示的一致性函数，也

表 6.3: 第六章的聚类准确性（续）

	COIL	ORL	WebACE	Tr11	Tr12
Kmeans	0.59	0.50	0.43	0.52	0.47
KC	0.62	0.53	0.46	0.57	0.56
CSPA	0.69	0.58	0.40	0.49	0.54
HGPA	0.55	0.60	0.35	0.47	0.52
MCLA	0.69	0. 60	0.47	0.52	0.57
BCE	0.67	0.51	0.48	0.58	0.58
GWCA	0.58	0.52	0.47	0.58	0.58
WGNMF-Euc	0.69	0.56	0.46	0.60	0.57
WGNMF-KL	0.71	0.60	0.48	0.60	0.60

就是矩阵分解得到的非负低维因子；此外，我们通过非负线性加权的多重关联关系构造一致的关联关系，这里我们需要学习线性权重。基于这两种一致性函数在聚类集成任务上的内在相关性，我们利用联合学习的思想提出基于加权图正则非负矩阵分解的聚类集成算法。在该算法中，当多重关联关系权重给定时，我们得到一致的关联关系，利用基于该关系的图正则非负矩阵分解提高非负因子的一致性（基于非负因子的聚类质量）；而当非负因子给定时，我们利用非负因子和每个关联关系的匹配度来估计每个关系对应的加权系数。可以看出基于簇特征和多重关联关系的一致性函数学习在迭代过程中互相提高，因此既可以提高非负因子上的聚类集成(final clustering)效果，又可以利用线性权重选择好的输入聚类(base clustering)。在大量数据集上的实验也验证了本算法在这两个方面的优势。

表 6.4: 第六章的归一化互信息

	Iris	Glass	Ecoli	Soybean	Zoo
Kmeans	0.69	0.31	0.58	0.72	0.69
KC	0.72	0.33	0.59	0.67	0.70
CSPA	0.71	0.29	0.51	0.63	0.59
HGPA	0.39	0.26	0.40	0.69	0.60
MCLA	0.74	0.32	0.56	0.71	0.74
BCE	0.74	0.35	0.59	0.69	0.70
GWCA	0.75	0.37	0.57	0.71	0.73
WGNMF-Euc	0.74	0.38	0.59	0.73	0.74
WGNMF-KL	0.74	0.37	0.58	0.71	0.73

表 6.5: 第六章的归一化互信息（续）

	COIL	ORL	WebACE	Tr11	Tr12
Kmeans	0.73	0.71	0.53	0.48	0.38
KC	0.74	0.74	0.55	0.58	0.54
CSPA	0.76	0.76	0.51	0.52	0.49
HGPA	0.69	0.77	0.45	0.48	0.43
MCLA	0.78	0.77	0.54	0.52	0.50
BCE	0.77	0.69	0.57	0.60	0.57
GWCA	0.73	0.73	0.56	0.56	0.50
WGNMF-Euc	0.79	0.75	0.58	0.59	0.57
WGNMF-KL	0.80	0.78	0.57	0.60	0.59

表 6.6: 第六章WGNMF对参数λ的敏感性

	Acc				NMI			
	0.1	1	10	100	0.1	1	10	100
Iris	0.89	0.89	0.89	0.89	0.74	0.74	0.74	0.74
Glass	0.53	0.54	0.52	0.52	0.37	0.38	0.37	0.37
Soybean	0.73	0.73	0.73	0.75	0.71	0.71	0.71	0.73
Zoo	0.75	0.75	0.77	0.68	0.73	0.74	0.74	0.70

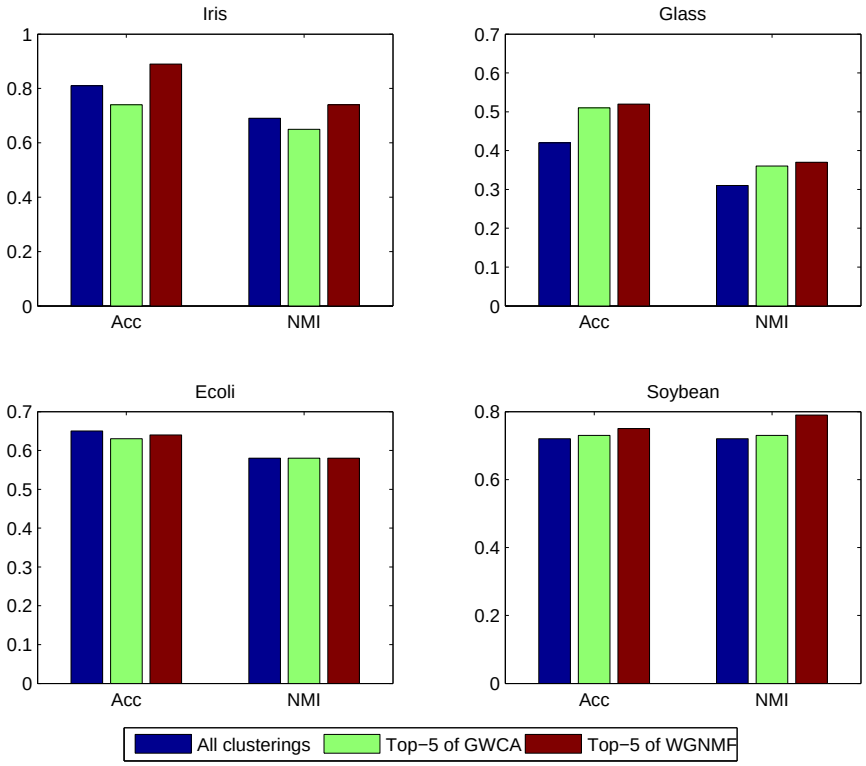


图 6.1: 聚类效果对比：全部聚类vs. Top-5聚类

第七章　结束语

7.1　全文总结

本文开展基于鲁棒非负矩阵分解的聚类方法研究。具体说来，我们提出的鲁棒非负矩阵分解模型和算法的相应贡献如下：

- 第二章提出了基于半二次优化的鲁棒非负矩阵分解算法。该方法通过半二次损失函数度量非负矩阵分解的质量，由于半二次函数在显著误差上损失较小或者在长尾数据上有着较高的概率分布，这样得到的分解结果就可以克服标准非负矩阵分解对显著误差、非高斯噪声敏感的缺点。本章提出基于半二次优化的通用优化框架统一求解基于半二次损失函数的非负矩阵分解问题。通用求解算法分别通过半二次优化的乘法形式和加法形式，将难处理的原始优化问题简化为两步易处理的问题并分别求解。

- 第三章提出了基于非负矩阵分解的鲁棒联合聚类算法。该算法从鲁棒数据重构和鲁棒图正则两方面提高基于非负矩阵分解的联合聚类算法的鲁棒性。具体来讲，该算法引入一个错误矩阵刻画稀疏的显著误差以恢复真实数据，并通过平方误差度量恢复数据上的重构误差，提高数据重构的鲁棒性；该算法通过采用绝对值图正则误差而不是常用的平方图正则误差降低特征与特征以及样本与样本之间噪声关系的影响并产生更紧凑的矩阵分解结果。真实数据集上的联合聚类实验结果表明，本章提出的鲁棒联合聚类算法要好于当前的主流方法。

- 第四章提出了鲁棒图正则非负矩阵三因子分解的模型和算法。为了进一步提高鲁棒联合聚类算法的鲁棒性，我们首先证明之前提出的鲁棒联合聚类算法中基于稀疏性噪声的重构误差与休伯损失函数的对偶关系，然后进一步提出采用一般的半二次损失函数来分别度量特征与样本关系矩阵的重构误差以及样本与样本，特征与特征关系的图正则误差。我们分别推导出基于乘法形式和加法形式的鲁棒图正则非负矩阵三因子分解的求解算法并证明其收敛性。

- 第五章提出区间矩阵分解方法。为了克服实际应用中单值数据引入的误差，该方法用一个基于均匀分布的区间来近似真实数据可能的取值，通过联合矩阵分解准确地近似该区间对应的上下界矩阵。在人脸分析和协同过滤两个应用中提出了经验性方法构造该区间对应的上下界矩阵，大量实验结果表明本文提出的区间矩阵分解方法要显著优于标准的单值矩阵分解。

- 第六章提出加权图正则非负矩阵分解的聚类集成算法。基于聚类集成问题中两种数据表示形式：基于簇特征和基于多重关联关系的表示，在集成任务上的内在相关性，本文利用联合学习的思想提出基于加权图正则非负矩阵分解的聚类集成算法。其中给定加权关联关系时，该算法通过图正则非负矩阵分解提高非负因子的聚类质量；而当非负因子给定时，该算法利用非负因子和关联关系的匹配度学习多重关系的权重。可以看出基于簇特征和多重关联关系的一致性学习在迭代过程中互相提高。需要指出的是，本算法也可以直接用于聚类分析，即通过联合加权图正则和非负矩阵分解一方面提高图正则的鲁棒性一方面提高聚类效果。

7.2 下一步的研究工作

本文中提出模型和算法有很多未来的工作可以继续展开，以下列出一些可以继续展开的方向：

- 本文的工作主要面向一类可以被半二次优化框架统一求解的目标函数（即半二次损失函数）考虑其鲁棒性，我们也可以考虑其他失真度函数（如Bregman失真度，alpha-beta-gamma-失真度）并试图从半二次优化的角度统一求解该类非负矩阵分解优化问题。本文的工作主要建立了针对半二次损失函数的非负矩阵分解通用求解算法，而如何利用不同应用，不同领域，不同任务的上下文知识准确选择目标函数仍然是实际应用亟待解决的问题。

- 鲁棒非负矩阵分解往往包含多个需要指定的超参数，并且由于这些模型是非凸的，算法的好坏还依赖于模型参数的初始化。一个好的鲁棒模型和算法不仅可以有效处理数据中的噪声还需要对模型自身的参数和超参

数鲁棒。因此如何提高参数初始化的质量和超参数的稳定性对鲁棒联合聚类算法尤为重要。

- 本文的工作基于无监督方式提高非负矩阵分解的鲁棒性，如何结合额外的领域知识（如标签信息，像Must-link和Can't-link这类的基于实例的两两约束）有效提高模型鲁棒性就成了一个很有意义的问题。尽管本文提出的方法比标准非负矩阵分解对显著误差、非高斯噪声鲁棒性要好，这些鲁棒模型对噪声的检测和修复都建立在所有数据的基础上（如乘法形式的绝对权重和加法形式的噪声仅依赖于矩阵每个元素拟合误差本身，而相对权重取决于所有元素）。而事实上，噪声往往带有一定的局部性（如图像中的某个像素被破坏，附近的像素也有很大的可能是噪声）。因此我们可以研究如何利用数据间的关系（如特征的空间位置关系）设计鲁棒损失函数超参数设置方法，进一步提高非负矩阵分解的鲁棒性。

- 本文的工作旨在通过利用鲁棒的损失函数来提高模型的鲁棒性，而数据挖掘机器学习中的集成学习也是提高模型稳定性和泛化能力的一个重要手段。不同的是基于重采样（bootstrapping, bagging）的集成学习主要通过不断增加大训练误差的权重来提高学习的效果（如AdaBoost,RankBoost），而鲁棒矩阵分解的主要思想是通过不断减小显著误差的权重来提高模型的鲁棒性，可以看出虽然boosting和鲁棒分解的技术路线相反，但其本质都是一种代价敏感的学习机制。因此，如何集成分解模型与集成学习提高分解模型的鲁棒性也是一个很有意思的问题。

参考文献

[1] 周志华，杨强. 机器学习及其应用*2011*, volume 4.

[2] 李国杰，程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 27(006):647–657, 2012.

[3] M. Al-Razgan and C. Domeniconi. Weighted clustering ensembles. In *Proceedings of 6th SIAM International Conference on Data Mining*, pages 258–269, 2006.

[4] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[5] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dharmendra S Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 509–514. ACM, 2004.

[6] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research (JMLR)*, 6:1705–1749, 2005.

[7] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[8] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems (NIPS)*, 14:585–591, 2001.

[9] Michael W Berry and Murray Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research (JMLR)*, 3:993–1022, 2003.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[12] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences, (PNAS)*, 101(12):4164–4169, 2004.

[13] Deng Cai, Xiaofei He, and Jiawei Han. Locally consistent concept factorization for document clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 23(6):902–913, 2011.

[14] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1548–1560, 2011.

[15] Deng Cai, Xiaofei He, Xuanhui Wang, Hujun Bao, and Jiawei Han. Locality preserving nonnegative matrix factorization. In *Proceedings of the 21st international jont conference on Artifical intelligence (IJCAI)*, pages 1010–1015. Morgan Kaufmann Publishers Inc., 2009.

[16] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *Proceedings of the 8th IEEE International Conference on Data Mining, ICDM'08.*, pages 63–72. IEEE, 2008.

[17] Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the Association for Computing Machinery (JACM)*, 58(3):1–37, 2011.

[18] Eugenio Cesario, Giuseppe Manco, and Riccardo Ortale. Top-down parameter-free clustering of high-dimensional categorical data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(12):1607–1624, 2007.

[19] Gang Chen, Fei Wang, and Changshui Zhang. Collaborative filtering using orthogonal nonnegative matrix tri-factorization. *Information Processing & Management (IPM)*, 45(3):368–379, 2009.

[20] Yan Chen, Hujun Bao, and Xiaofei He. Non-negative local coordinate factorization for image representation. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 569–574. IEEE, 2011.

[21] Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[22] Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

[23] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Csiszar' s divergences for non-negative matrix factorization: Family of new algorithms. In *Independent Component Analysis and Blind Signal Separation*, pages 32–39. Springer, 2006.

[24] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 269–274. ACM, 2001.

[25] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 89–98. ACM, 2003.

[26] Inderjit S. Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 283–290. The MIT Press, 2006.

[27] Chirs Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(1):45–55, 2010.

[28] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proc. SIAM Data Mining Conf*, number 4, pages 606–610, 2005.

[29] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, (KDD)*, pages 126–135. ACM, 2006.

[30] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine Learning (ICML)*, pages 281–288. ACM, 2006.

[31] Liang Du, Xuan Li, and Yi-Dong Shen. Cluster ensembles via weighted graph regularized nonnegative matrix factorization. In *Advanced Data Mining and Applications*, pages 215–228. Springer, 2011.

[32] Tao Feng, Stan Z Li, Heung-Yeung Shum, and HongJiang Zhang. Local non-negative matrix factorization as a visual representation. In *The 2nd International Conference on Development and Learning, 2002. Proceedings*, pages 178–183. IEEE, 2002.

[33] Xiaoli Zhang Fern and Carla E Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning (ICML)*, page 36. ACM, 2004.

[34] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[35] Bo Geng, Dacheng Tao, Chao Xu, Linjun Yang, and Xian-Sheng Hua. Ensemble manifold regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(6):1227–1233, 2012.

[36] Quanquan Gu, Chris Ding, and Jiawei Han. On trivial solution and scale transfer problems in graph regularized nmf. In *Proceedings of the 22th international joint conference on Artificial Intelligence (IJCAI)*, pages 1288–1293. AAAI Press, 2011.

[37] Quanquan Gu and Jie Zhou. Co-clustering on manifolds. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 359–368. ACM, 2009.

[38] Quanquan Gu and Jie Zhou. Local learning regularized nonnegative matrix factorization. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1046–1051, 2009.

[39] Quanquan Gu and Jie Zhou. Neighborhood preserving nonnegative matrix factorization. In *Proceedings of the 20th British Machine Vision Conference (BMVC)*, pages 1–10, 2009.

[40] Quanquan Gu, Jie Zhou, and Chris Ding. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *Proceedings of the 10th SIAM International Conference on Data Mining (SDM)*, pages 199–210, 2010.

[41] Stefan T Hadjitodorov, Ludmila I Kuncheva, and Ludmila P Todorova. Moderate diversity for better cluster ensembles. *Information Fusion*, 7(3):264–275, 2006.

[42] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. Wiley, 2011.

[43] A. Ben Hamza and David J. Brady. Reconstruction of reflectance spectra using robust nonnegative matrix factorization. *IEEE Transactions on Signal Processing (TSP)*, 54(9):3637–3642, 2006.

[44] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[45] Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1561–1576, 2011.

[46] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[47] Patrik O Hoyer. Non-negative sparse coding. In *Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565. IEEE, 2002.

[48] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research (JMLR)*, 5:1457–1469, 2004.

[49] Cho-Jui Hsieh and Inderjit S Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 1064–1072. ACM, 2011.

[50] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359, 1999.

[51] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. Fast newton-type methods for the least squares nonnegative matrix approximation problem. In *Proceedings of SIAM Conference on Data Mining (SDM)*, pages 343–354, 2007.

[52] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[53] Jingu Kim and Haesun Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *2008 IEEE 8th International Conference on Data Mining (ICDM)*, pages 353–362. IEEE, 2008.

[54] Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, 13(4):703–716, 2003.

[55] Raul Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural computation*, 19(3):780–791, 2007.

[56] Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using ℓ_{21}-norm. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM)*, pages 673–682. ACM, 2011.

[57] Irene Kotsia, Stefanos Zafeiriou, and Ioannis Pitas. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *IEEE Transactions on Information Forensics and Security*, 2(3):588–595, 2007.

[58] Da Kuang, Chris Ding, and Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM)*, pages 106–117, 2012.

[59] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems (NIPS)*, 13:556–562, 2001.

[60] Daniel D. Lee and HSebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[61] Ping Li, Jiajun Bu, Chun Chen, and Zhanying He. Relational co-clustering via manifold ensemble learning. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM)*, pages 1687–1691. ACM, 2012.

[62] Ping Li, Jiajun Bu, Chun Chen, Zhanying He, and Deng Cai. Relational multimanifold coclustering. *IEEE Transactions on Systems, Man and Cybernetics, Part B (TSMCB)*, PP(99):1–11, 2013.

[63] Tao Li and Chris Ding. Weighted consensus clustering. In *Proceedings of the 8th SIAM International Conference on Data Mining (SDM)*, pages 798–809, 2008.

[64] Tao Li, Chris Ding, and Michael I Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, pages 577–582. IEEE, 2007.

[65] Eufrásio de A Lima Neto and Francisco de AT de Carvalho. Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, 54(2):333–347, 2010.

[66] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[67] Chao Liu, Hung-chih Yang, Jinliang Fan, Li-Wei He, and Yi-Min Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *Proceedings of the 19th international conference on World wide web (WWW)*, pages 681–690. ACM, 2010.

[68] Haifeng Liu and Zhaohui Wu. Non-negative matrix factorization with constraints. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 506–511, 2010.

[69] Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang. Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(7):1299–1311, 2012.

[70] Weifeng Liu, Puskal P. Pokharel, and Jose C. Principe. Correntropy: properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing (TSP)*, 55(11):5286–5298, 2007.

[71] Weixiang Liu, Nanning Zheng, and Xiaofeng Lu. Non-negative matrix factorization for visual coding. In *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 288–293. IEEE, 2003.

[72] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1033–1039, 2012.

[73] László Lovász and Michael D. Plummer. *Matching theory*. American Mathematical Society, 1986.

[74] Tom M Mitchell. Machine learning. wcb, 1997.

[75] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems (NIPS)*, 2:849–856, 2002.

[76] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *Advances in Neural Information Processing Systems (NIPS)*, 23:1813–1821, 2010.

[77] Mila Nikolova. Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. *Multiscale Modeling & Simulation*, 4(3):960–991, 2005.

[78] Mila Nikolova and Raymond H. Chan. The equivalence of half-quadratic minimization and the gradient linearization iteration. *IEEE Transactions on Image Processing (TSP)*, 16(6):1623–1627, 2007.

[79] Mila Nikolova and Michael K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific Computing*, 27(3):937–966, 2006.

[80] Mila Nikolova, Michael K Ng, and Chi-Pan Tam. Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Transactions on Image Processing (TIP)*, 19(12):3073–3088, 2010.

[81] Pentti Paatero and Unto Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[82] Feng Pan, Xiang Zhang, and Wei Wang. Crd: fast co-clustering on large datasets utilizing sampling-based matrix decomposition. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD)*, pages 173–184. ACM, 2008.

[83] Alberto Pascual-Montano, J.M. Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(3):403–415, 2006.

[84] Jose C Principe, Dongxin Xu, and John Fisher. Information theoretic learning. *Unsupervised adaptive filtering*, 1:265–319, 2000.

[85] Yuntao Qian, Sen Jia, Jun Zhou, and Antonio Robles-Kelly. Hyperspectral unmixing via $l_{1/2}$ sparsity-constrained nonnegative matrix factorization.

IEEE Transactions on Geoscience and Remote Sensing, 49(11):4282–4297, 2011.

[86] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning (ICML)*, pages 880–887. ACM, 2008.

[87] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. *Advances in neural information processing systems (NIPS)*, 20:1257–1264, 2008.

[88] Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1590–1602, 2011.

[89] Fei Sha, Yuanqing Lin, Lawrence K Saul, and Daniel D Lee. Multiplicative updates for nonnegative quadratic programming. *Neural Computation*, 19(8):2004–2031, 2007.

[90] Fanhua Shang, L.C. Jiao, and Fei Wang. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition*, 45(6):2237–2250, 2012.

[91] Bin Shen and Luo Si. Nonnegative matrix factorization clustering on multiple manifolds. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI)*, pages 575–580, 2010.

[92] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):888–905, 2000.

[93] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research (JMLR)*, 3:583–617, 2003.

[94] Meng Sun and Hugo Van Hamme. Large scale graph regularized nonnegative matrix factorization with ℓ_1 normalization based on kullback˘leibler divergence. In *IEEE Transactions on Signal Processing (TSP)*, pages 3876–3880, 2012.

[95] A. Topchy, A.K. Jain, and W. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1866–1881, 2005.

[96] Alexander Topchy, Anil K Jain, and William Punch. A mixture model of clustering ensembles. In *Proceedings of 4th SIAM International Conference on Data Mining (SDM)*, pages 379–390, 2004.

[97] Dingding Wang, Tao Li, and Chris Ding. Weighted feature subset nonnegative matrix factorization and its applications to document understanding. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, pages 541–550. IEEE, 2010.

[98] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery (DMKD)*, 22(3):493–521, 2011.

[99] Fei Wang, Xin Wang, and Tao Li. Generalized cluster aggregation. In *Proceedings of the 21st international joint conference on artificial intelligence (IJCAI)*, pages 1279–1284, 2009.

[100] Hongjun Wang, Hanhuai Shan, and Arindam Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70, 2011.

[101] Hua Wang, Feiping Nie, Heng Huang, and Chris Ding. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pages 774–783. IEEE, 2011.

[102] Hua Wang, Feiping Nie, Heng Huang, and Fillia Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *Pro-*

ceedings of the 22th international joint conference on Artificial Intelligence (IJCAI), pages 1553–1558. AAAI Press, 2011.

[103] Jing-Yan Wang, Halima Bensmail, and Xin Gao. Multiple graph regularized nonnegative matrix factorization. *Pattern Recognition*, 2013.

[104] Yuan Wang and Yunde Jia. Fisher non-negative matrix factorization for learning local features. In *Proceedings of the 4th Asian Conference on Computer Vision (ACCV)*, pages 27–30, 2004.

[105] Rui Xu, Donald Wunsch, et al. Survey of clustering algorithms. *IEEE Transactions on Neural Networks (TNN)*, 16(3):645–678, 2005.

[106] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR)*, pages 267–273. ACM, 2003.

[107] Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing (TIP)*, 19(10):2761–2773, 2010.

[108] Jiho Yoo and Seungjin Choi. Weighted nonnegative matrix co-tri-factorization for collaborative prediction. In *Advances in Machine Learning*, pages 396–411. Springer, 2009.

[109] Xiao-Tong Yuan and Bao-Gang Hu. Robust feature extraction via information theoretic learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1193–1200. ACM, 2009.

[110] Stefanos Zafeiriou, Anastasios Tefas, Ioan Buciu, and Ioannis Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks (TNN)*, 17(3):683–695, 2006.

[111] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems (NIPS)*, 17:1601–1608, 2004.

[112] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management (CIKM)*, pages 25–32. ACM, 2001.

[113] Lijun Zhang, Chun Chen, Jiajun Bu, Zhengguang Chen, Deng Cai, and Jiawei Han. Locally discriminative coclustering. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 24(6):1025–1035, 2012.

[114] Lijun Zhang, Zhengguang Chen, Miao Zheng, and Xiaofei He. Robust non-negative matrix factorization. *Frontiers of Electrical and Electronic Engineering in China*, 6(2):192–200, 2011.

[115] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. Learning from incomplete ratings using non-negative matrix factorization. *Proceedings of the 6th SIAM International Conference on Data Mining (SDM)*, pages 548–552, 2006.

[116] Yu Zhang and Dit-Yan Yeung. Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 606–614. ACM, 2012.

[117] Zhengyou Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1997.

[118] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and WBastiaan Kleijn. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (TSMCB)*, 41(1):38–52, 2011.

[119] Fuzhen Zhuang, Ping Luo, Hui Xiong, Qing He, Y. Xiong, and Zhongzhi Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization. *Statistical Analysis and Data Mining*, 4(1):100–114, 2011.

发表文章目录

[1] **Liang Du** and Yi-Dong Shen. Towards robust co-clustering. The 23rd International Joint Conference on Artificial Intelligence (IJCAI), 2013, (Oral paper, accepted rate 195/1473 = 13.2%).

[2] **Liang Du**, Xuan Li and Yi-Dong Shen. Robust nonnegative matrix factorization via half-quadratic minimization. In Proceedings of IEEE 12th International Conference on Data Mining (ICDM), 2012, pages 201-210. (Regular paper, accepted rate 81/756 = 10.7%).

[3] **Liang Du** and Yi-Dong Shen. Joint clustering and feature selection. The 14th International Conference on Web-Age Information Management (WAIM), 2013, (Full paper, Accepted).

[4] **Liang Du**, Yi-Dong Shen, Zhiyong Shen, Jianying Wang and Zhiwu Xu. A self-supervised framework for clustering ensemble. The 14th International Conference on Web-Age Information Management (WAIM), 2013, (Full paper, Accepted).

[5] **Liang Du**, Xuan Li and Yi-Dong Shen. Cluster ensembles via weighted graph regularized nonnegative matrix factorization. Advanced Data Mining and Applications (ADMA), 2011, pages 215-228.

[6] **Liang Du**, Xuan Li and Yi-Dong Shen. User graph regularized pairwise matrix factorization for item recommendation. Advanced Data Mining and Applications (ADMA), 2011, pages 372-385.

[7] Xuan Li, **Liang Du** and Yi-Dong Shen. Update summarization via graph-based sentence ranking. IEEE Transactions on Knowledge and Data Engineering (TKDE), May 2013, vol.25, no.5, pp.1162-1174.

[8] Xuan Li, **Liang Du** and Yi-Dong Shen. Graph-based marginal ranking for update summarization. In Proceedings of the Eleventh SIAM International Conference on Data Mining (SDM), 2011, pages 486-497.

[9] Zhiyong Shen, **Liang Du**, Xukun Shen and Yi-Dong Shen. Interval-valued matrix factorization with applications. In Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), 2010, pages 1037-1042.

[10] Xuan Li, Yi-Dong Shen, **Liang Du** and Chen-Yan Xiong. Exploiting novelty, coverage and balance for topic-focused multi-document summarization. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM), 2010, pages 1765-1768.

[11] Liang Wu, Alvin Chin, Guandong Xu, **Liang Du**, Xia Wang, Kangjian Meng, Yonggang Guo and Yuanchun Zhou. Who Will Follow Your Shop? Exploiting Multiple Information Sources in Finding Followers. The 18th International Conference on Database Systems for Advanced Applications (DASFAA), 2013, pages 401-415.

简　　历

基本情况

姓　　名：杜亮　　性　　别：男
民　　族：汉族　　籍　　贯：山西省左权县
出生年月：1985年4月　　政治面貌：中共党员

教育状况

2003 年 9 月至 2007 年 7 月，武汉大学国际软件学院，软件工程学士。

2007 年 9 月至 2013 年 7 月，中国科学院软件研究所，硕博连读研究生。

工作经历

无。

研究兴趣

机器学习、数据挖掘、信息检索。

联系方式

通讯地址：北京市中关村南四街四号中科院软件园五号楼

邮编：100190

E-mail: duliang@ios.ac.cn

致　　谢

值此论文完成之际，谨在此向多年来给予我关心和帮助的老师、同学、朋友和家人表示衷心的感谢!

首先要感谢我的导师沈一栋研究员。六年求学期间，沈老师在学习、研究以及生活上都给予了无微不至的关怀和指导，提供了良好的学术氛围和宽松的科研环境，是我完成学业的基本条件。沈老师严谨的治学态度不断求索的开拓精神和刻苦钻研的为学精神是我永远学习的榜样。沈老师待人谦和、乐观向上的人生态度同样值得我学习。在此谨向他致以我深深的谢意。

感谢实验室的师兄杜剑锋博士、沈志勇博士、孙军博士、李炫博士和师弟熊辰炎、邓军、周芃、石磊、汪涵默，感谢你们给予我的各种帮助，和你们的交流和讨论给了我许多启发，你们的优秀也让我受益良多。在这里我要特别感谢沈志勇博士和李炫博士，感谢你们一直以来的无私帮助和支持，和我一起讨论科研中遇到的各种问题。

感谢读博期间的同学和朋友，许智武博士、别晓辉博士、余玉银博士、计算所单书畅博士、部德振、曹东坡、段智全等同学，感谢你们给予我的支持和带给我的欢乐。

感谢我的家人，包括我的父亲、母亲和我的哥哥，感谢他们由始至终的支持、鼓励和关爱，我才得以顺利完成学业。殷殷厚爱，无以为报，希望这篇博士论文能成为我送给他们最好的礼物。

感谢所有曾经给予我关心、支持和帮助的人们。