**大学生创新创业训练计划项目**

**结项报告**

项目名称：基于网络大数据研究的聚类算法及其应用领域

项目类别： 创新训练项目

项目级别： 北京市级

立项年度： 2019年

依托单位： 管理科学与工程学院

主 持 人： 易尹珏

项目成员： 易尹珏、刘晓语、路安仪、张利娟、魏帅

指导教师： 孟凡

2020年 4 月 20 日

**内容摘要：**在如今的数据大爆炸时代，我们无时无刻不接受着海量的数据如文本、图像、视频和链接等等。信息的产生是如此快速而复杂，数据的类型、大小和维度也已远远超出传统手工处理的承受范围。因此，聚类分析，作为一种最常见的无监督学习技术，在处理网络大数据、分类分析数据模式、总结归纳信息知识方面获得了广泛应用。在对现有算法进行深入学习的基础上，本项目针对于目前大多数传统聚类算法仅适用于处理单视角数据、而现实情况中许多数据都是同时由多个角度进行描绘的问题，引入多视角情景，研究一种多视角判别分析方法对数据进行降维处理，将原始空间多视角数据集投影到低维空间，既充分考虑多视角带来的信息多元化，又尽可能降低数据分析的复杂性。与此同时，针对现有聚类算法工具大多存在着时间复杂度高的弊端，尝试引用一种较为创新的单视角聚类算法，用新思路刻画聚类中心。综合考虑上述两个优化目标，在多视角数据上获取初始标签信息，对数据进行降维后获得子空间内新的数据集，并在子空间上利用新聚类算法，多次迭代后输出最终的分类标签，以此逻辑建立起完整的聚类算法模型。随后在Image segmentation data set、Mfeat两个多视角数据集上进行实验，测试其聚类结果，并基于三种评价指标对此聚类模型进行综合评估。

**关键词：**网络 大数据 聚类 多视角

**Abstract:** In today's era of data explosion, we are constantly receiving huge amounts of data such as text, images, videos and links.The generation of information is so fast and complex that the types, sizes, and dimensions of the data are far beyond the reach of traditional manual processing.Therefore, cluster analysis, as one of the most common unsupervised learning techniques, has been widely used in processing network big data, classifying and analyzing data patterns, and summarizing information knowledge.On the basis of an in-depth study of existing algorithm, this project for the present most of the traditional clustering algorithm applies only to deal with single view many data in the data, and the reality are describe problem by multiple angles at the same time, the introduction of multiple points of view, study a multi-view discriminant analysis method to dimension data, more than the original space perspective projection data sets to low dimensional space, fully considering the perspectives of both the diversified information, and reduce the complexity of data analysis.At the same time, in view of the disadvantages of high time complexity in most existing clustering algorithm tools, this paper tries to introduce an innovative single-angle clustering algorithm and describe the clustering center with a new idea.Taking the above two optimization objectives into consideration, the initial label information was obtained on the multi-angle data, and the new data set in the subspace was obtained after dimensionality reduction of the data, and the new clustering algorithm was used in the subspace to output the final classification label after several iterations, so as to establish a complete clustering algorithm model logically.Then experiments were carried out on two multi-perspective data sets, Image segmentation data set and Mfeat, to test the clustering results, and the clustering model was comprehensively evaluated based on three evaluation indexes.

**Key Words:** Network Big data Clustering Multiple perspectives

**目 录**

**[引言 4](#_Toc38751002)**

**[一、研究目标 5](#_Toc38751003)**

**[二、研究过程执行情况概述 5](#_Toc38751004)**

[（一）按计划执行情况 5](#_Toc38751005)

[（二）研究工作进展及成果 5](#_Toc38751006)

**[三、研究过程 5](#_Toc38751007)**

[（一）文献研究与算法学习 5](#_Toc38751008)

[（二）数据收集与整理分析 6](#_Toc38751009)

[（三）多视角降维模型、单视角快速聚类算法建立 6](#_Toc38751010)

[（四）算法的案例实验及结果评估 6](#_Toc38751011)

**[四、研究内容 6](#_Toc38751012)**

[（一）多视角 6](#_Toc38751013)

[（二）基础聚类方法：CFDP 8](#_Toc38751014)

[（三）多视图降维及MVDA算法 11](#_Toc38751015)

[（四）算法总流程 13](#_Toc38751016)

**[五、测试结果分析 13](#_Toc38751017)**

[（一）数据集 13](#_Toc38751018)

[（二）聚类评价指标 14](#_Toc38751019)

[（三）算法实验结果及分析 15](#_Toc38751020)

[（四）优缺点总结 1](#_Toc38751021)9

**[结论 1](#_Toc38751022)9**

**[参考文献 2](#_Toc38751023)0**

# **引言**

聚类分析是一个有着重要作用的研究领域,其中心思想是根据设定的约束或准则，将原本数量较大、相对独立的大数据集样本分为若干类群，“距离”近的对象视为一类，不同类之间的对象“距离”较远。将分散的个体聚类成类，使其各自具有一定功能性，进而发现样本之间潜在的信息和规律。自上世纪八十年代研究人员提出k-means聚类以来，应用于各种情景下的聚类方法相继被提出。经过近半个世纪的发展，网络聚类分析的脉络逐渐清晰，体系逐渐完善，并成为数据挖掘领域的重要组成部分。根据不同的聚类思想，传统的聚类算法大致可分为四类：层次化聚类方法（bottom-up，top-down）、基于划分的聚类方法（k-means就是一个典型的效率极高的基于划分的聚类算法）、基于模型的聚类方法（如统计学、神经网络方法）和基于密度和网格的聚类方法（如DBSCAN、STING等）。

研究面向大数据的聚类算法问题具有重要的现实意义。聚类分析在许多大数据实际领域都有广泛的应用。模式识别方面，聚类常用于手写数字、字符字体识别；机器学习方面，聚类主要用于计算机视觉、图像分割等。在一些现实问题上，市场营销部门可利用聚类进行客户分层，发掘不同客户群的特点以开拓潜在市场份额；流媒体可通过对用户的微博条目进行主题分类，从而总结出近期的热点新闻和大众的聚焦中心。

虽然聚类算法已经得到了极大的发展，聚类的方法也应势而生百花齐放，但迄今还没有一个能完全适用于多种数据结构的聚类算法，即每一种聚类算法都只能应用在满足相应条件的数据集上。目前以k-means聚类为代表的传统聚类算法大多适用于处理单视角数据，但我们应意识到用多个视角去描绘数据、刻画事物的手段变得越来越常见——新闻会同时出现多方报道，一个网站包含文本、图片、其他网页超链接等信息数据。从这些数据中探索有用的知识是数据挖掘的一个重要课题，近年来，逐渐有研究提出创新性的多视角聚类算法，以摆脱传统聚类只能局限于对单视角数据进行聚类的问题。同时，现有聚类算法工具还大多存在着时间复杂度高的弊端。

本项目引入多视角情景，研究一种多视角判别分析方法对数据进行降维处理，将原始空间多视角数据集投影到低维空间，既充分考虑多视角带来的信息多元化，又尽可能降低数据分析的复杂性。与此同时，针对高时间复杂度问题，尝试引用一种较为创新的单视角聚类算法，用新思路刻画聚类中心。综合考虑上述两个优化目标，具体步骤为在多视角数据上获取初始标签信息，对数据进行降维后获得子空间内新的数据集，并在子空间上利用新聚类算法，多次迭代后输出最终的分类标签，以此逻辑建立起完整的聚类算法模型。随后在Image segmentation data set、Mfeat两个多视角数据集上进行实验，测试其聚类结果，并基于三种评价指标对此聚类模型进行综合评估。

# **一、研究目标**

聚类分析是数据挖掘相关领域的重要方法工具，在许多大数据实际领域，例如模式识别、计算机视觉、机器学习也都有着广泛的应用。经过数十年研究，聚类算法已得到多样化的深入发展，但目前的多数聚类算法都基于数据来源是单视角这一基本前提，与现实中大多信息都是被同时由多个角度数据刻画的实际情况不符；此外，面对高维大样本数据处理任务，现有的为追求较高精度的聚类算法工具，大多还存在着时间复杂度高的弊端。基于面向大数据的聚类算法的重要研究意义，针对上述提及的现有算法缺陷，我们将尝试引入多视角数据进行聚类分析，同时在多视角降维的基础之上采用一种新型的快速聚类算法完成单视角上的基础聚类。

# **二、研究过程执行情况概述**

# **（一）按计划执行情况**

1.2019.04-2019.07 结合研究对象，进行全面系统的文献调研，重点对模型理论与算法进行深入研究，同时进行网络大数据数据库的收集工作；

2.2019.08-2019.10 在多视角领域找到研究聚焦点，同时建立起基础聚类模型，并对其进行小型样本实验；

3.2019.11-2020.01 建立起对多视角数据进行降维处理的算法模型，进一步训练和调试单视角上的基础聚类算法，将两阶段模型结合；

4.2020.02-2020.04 修正两阶段模型，形成完整成熟的算法步骤，在多个大数据集上运行实验，评估并比较聚类结果优劣性。

# **（二）研究工作进展及成果**

研究团队按原定计划认真开展了研究工作，取得了预期的研究成果，顺利完成了各项预定的考核指标，具体研究成果如下：

成果一：考虑到多视角数据存在的普遍性，将多视角学习引入到传统的聚类分析中，从较为创新的角度进行大数据聚类算法研究；

成果二：发现一种能在单视角上进行快速高效聚类的算法模型CFDP,原理简单明快，能够降低一定聚类时间复杂度；

成果三：归纳整理出多种传统网络聚类算法模型和新型多视角领域的聚类思路相关文献。

# **三、研究过程**

# **（一）文献研究与算法学习**

我们团队针对所选的网络聚类课题，研读著作《汪小帆复杂网络专著》，文献《Attack Robustness and Centrality of Complex Networks》《Finding Critical Links for Closeness Centrality》《Efficient algorithms for game-theoretic betweenness centrality》等，着重对复杂网络下的多种聚类算法进行了深入学习，探究其性质及相应的适用范围。经研究后发现，一些已成熟的用于寻找网络聚类结构的算法，例如谱平分法、分类算法、凝聚算法等在现有基础上改进的难度较大，故决定将之后的重心转为寻找目前较少有人开发实验的网络聚类应用领域。

明确上述研究方向后，我们团队主要通过文献调研获得了具有操作性和前景性的几大领域：①金融贸易：基于复杂网络聚类模型分析证券投资、股票指数、全球贸易的分类标签和动态演变等；②社交网络：选取论坛网页、在线交易数据等研究用户划分、评论有用性识别；③交通航线：城市公共交通复杂网络、航空货运系统层级结构发现；④充电桩：基于电动能源汽车大力发展的背景，利用复杂网络理论，选取一定的约束因素来研究充电服务网络规划布局；⑤多视角学习：在网页大数据中应用高维多视角进行智能聚类。经比较研究，我们认为相比其他研究已相对深入透彻的领域，多视角聚类课题较新，有较好的实验意义，故选此应用领域。

# **（二）数据收集与整理分析**

由于研究对象为基于网络大数据的聚类算法，应用领域设定为多视角数据，我们参考多篇权威文献以及搜索开源数据网站，寻找符合多视角特征的数据库。要求样本标签完整，样本特征值明确，样本量适度。

运用MATLAB 软件，先进行小型样本实验以确定数据集的可用性、规范性，后测试大数据样本聚类的代码实现可行性。

# **（三）多视角降维模型、单视角快速聚类算法建立**

深入研究如何将不同视角上的数据进行并归处理，选择采用降维思想，同时考虑视角间和视角内的变量关系，将原始空间多视角数据集投影到低维空间，获得子空间内新的多视角数据集。

区别于一些传统的、已广泛应用的单视角聚类算法，发现一种新型的子空间单视角快速聚类方法，实验测评其算法优越性。

# **（四）算法的案例实验及结果评估**

建立评价指标体系，测试算法在多个真实数据集上的运行效果，并分析其优缺点。

# **四、研究内容**

# **（一）多视角**

在科学的数据分析中，研究员们会通过不同的视角或方式收集一个事物的信息。例如，就图像数据而言，可以将颜色，纹理质地，形状，文本文字等不同的特征作为不同视角对其进行描述；对于网页数据，可以从文本、图片、超链接信息三个不同的视角来进行研究。多视角数据是一种拥有多个视角、多个来源、多个表达方式、多个方面属性特征的数据。多视角数据集里的数据可以划分成多个视角，换言之多视角数据是由多个视角共同构成。各个视角之间既有一致信息，也有不同信息。如图1，（a）web页面可以从它的URL和页面上文字来描述；（b）web图像可以从其可视的图像以及周边文字来表达；（c）3D图像可以从不同视角来观察；（d）视频可以从音频和视觉框架来分析；（e）多语言文本中，每种语言可作为一个视角。

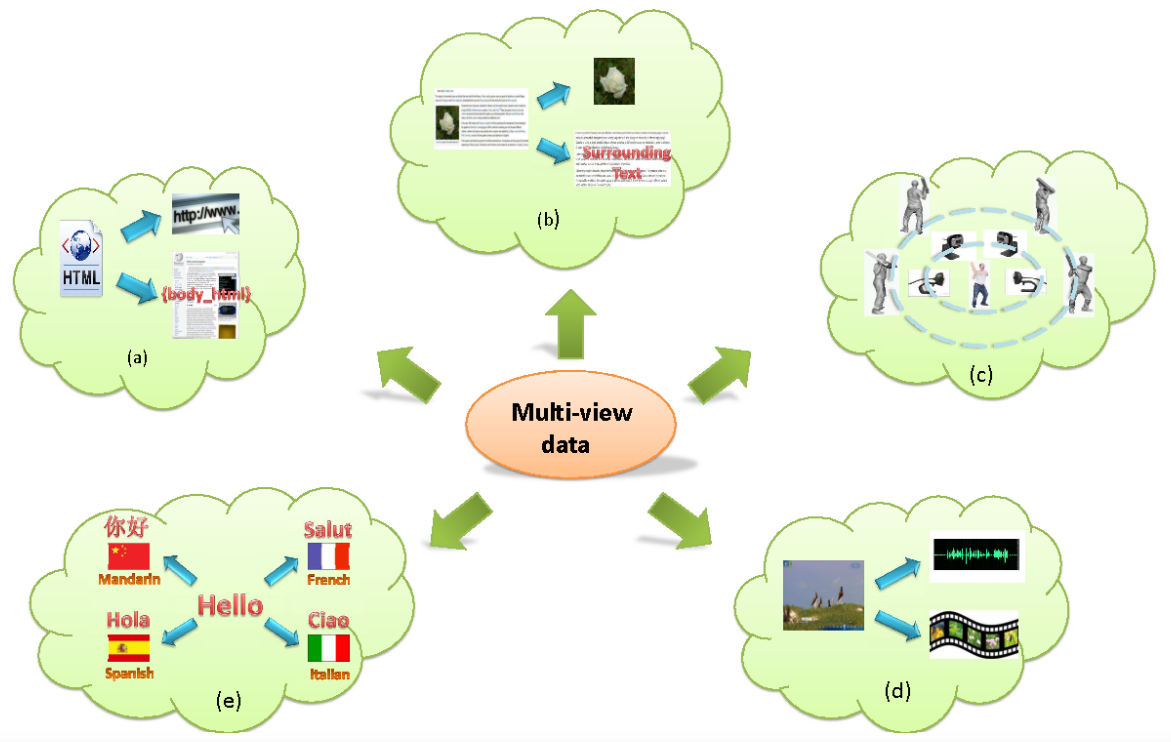


图1[[1]](#footnote-0)：多视角数据

在有了多视角的概念以后，人们主要研究多视角数据的降维和多视角数据的聚类这两个多视角数据的研究方面。

多视角降维是大数据时代针对多视角数据的一项非常有意义的研究。由于实际应用中的数据所包含的内容越来越多，数据维度越来越高，数据结构越来越复杂，人们更加注重数据分析结果的可靠性，这导致数据分析越来越难。数据量大，数据冗余，可视化困难，在低维数据上表现良好的算法在高维数据上失效，来自不同视角的样本也许会由于处在不同空间而无法比较等问题相继出现。多视角降维能够为多视角高维数据寻找一个公共空间并进行降维，这不仅有利于提高模型精确度和效率，还能帮助理解复杂的视角关系。

多视角聚类是面向多视角数据的一种数据分析手段，通过分析多视角数据的分布特点和利用多视角间的相容互补性，从而得到数据的潜在分组结构。在多视角数据的聚类问题上，多视角聚类可以比单视角得到更好的聚类结果。多视角聚类算法依据使用多视角信息的方式不同可分为以下四个类别：

1.基于协同训练的方法。协同训练（Co-training）是一类半监督学习。输入的有标记数据集L可以分为两个不相关的视角，在这两个视角上对L进行训练得到两个分类器，然后两个分类器对未标记数据集U进行标记，选出p个正标记n个负标记。将p+n个标记加入L。重复上述过程k次。

2.基于子空间学习的方法。多视角子空间学习要求找到一个多视角数据的低维嵌入（low-dimensional embedding）来代替原始的多视角数据，然后将其应用到聚类中。即用低维表示多视角数据，然后在这个新的表示（处理后的数据）上，使用传统聚类算法进行聚类。

3.基于多核学习的方法。多核学习（Multiple Kernel Learning, MKL）将多视角数据的每个视角看作一个核空间，对于每个视角通过核映射将样本从原始空间映射到一个高维特征空间。对每个核空间设置不同的权重从而将他们的信息结合起来，即采用多核形式将所有视角特征空间结合起来。基于多核学习的聚类方法使数据在新的特征空间得到更加准确、合理的表达，进而显著提高聚类性能。

4.基于集成学习的方法。聚类集成算法是将多个聚类分量进行融合，然后得到更优的聚类划分。首先分别对每个视角进行单视角聚类得到M个聚类划分，然后使用聚类集成算法对M个聚类划分进行集成，得到最终的聚类结果。

目前，多视角聚类己经成功应用于基因识别、文本分析、面部表情识别和医疗图像自动检测等领域。例如在光学影像序列的群组行为分析中，利用交互关系、运动方向分布、空间距离、运动规律等四个视角的相似图对影像块进行多视角聚类。在蛋白质功能模块分析中，通过提取基因表达数据和基因本体注释信息这两个视角数据中聚合特征最为一致的功能模块，分析得到蛋白质相互作用网络中的功能模块关系。

## **（二）基础聚类方法：CFDP**

聚类是将数据集合划分为相似数据类的过程，使得同类中对象间相似性较高，不同类中对象相异性较高。本文多视角聚类是基于Alex Rodriguez、Alessandro Laio在Science期刊发表的《[Clustering by fast search and find of density peaks](http://science.sciencemag.org/content/344/6191/1492" \t "https://segmentfault.com/a/_blank)》所提出的基于密度的快速聚类算法。

**1.算法思想**

该算法的核心思想在于设定聚类簇中心所具有的两个特点：一是类簇中心被一些局部密度低于它的点环绕，即中心密度高。二是类簇中心和其他高局部密度的点间相距较大。

设存在待聚类数据集为相应指标集，表示点和间距离。对中数据点均定义两个值：局部密度以及到其他高局部密度点的距离，而这两个值仅与两点之间的距离有关。

（1）局部密度（离散值）:为与点距离小于的点的个数。

 （1）

其中函数



 （2）

是截断距离（需要提前给定的参数）。该算法仅对的相对值敏感，所以对的选择具有较强鲁棒性。

1. 距离:用于描述点i到其他较高密度点之间的最小距离。

 （3）

对于密度最大的点, 设

 （4）

综上所述，那些局部密度或全局密度最大的间距会远大于普通的相邻点间距，因此类簇中心是值异常最大的点。

**2.聚类过程**

根据上述可知，具有较大局部密度和异常大值的点被视为类簇中心，具有较小局部密度和较大值的点是异常点。该算法中，在类簇中心确定后，剩余点将会被分配给最近邻居相同的密度较高的类簇。

为描述方便起见，给出下述记号。

* 待聚类数据集，设包含个聚类。
* :各个聚类中心对应的数据点的编号，即为第个聚类中心。
* :数据点归属聚类标记，即表示中第个数据点属于第个聚类。
* :中距离最远的两个数据点之间的距离。
* 表示中多有局部密度比大的数据点与距离最近的数据点编号。
* 若，表示局部密度较小,对应聚类边缘部分，即离群点；若，则表示局部密度较大，属于某聚类，称为聚类核心部分。

具体算法流程如下：

|  |
| --- |
| CFDP算法步骤 |
| 输入：待聚类数据集 |
| 1：确定截断距离 |
| 2：计算及 |
| 3：确定类簇中心，并对数据点进行类簇归属标记 |
| 4：对非聚类中心点进行归类 |
| 5：若，则将每个类簇中的数据点细分为类簇核心部分和类簇边缘部分 |
| 输出：聚类中心及数据点对应归属标记 |

**3.聚类效果**

CFDP方法思路明确清晰，在实际运用中的精度和鲁棒性均有不错的表现。以下将举例说明该方法聚类效果，图A表示数据点分布，包含非球形点集及双峰点集。B图表示4000个点按照A图所示方式分布，C图为1000个点时的分布情况，其中不同类簇点着色不同，黑点为类簇光晕（异常点）。D和E是分别对应图B、C的聚类中心决策图，而F表示不同数据点数量下错误聚类的点的比率，误差线为平均值的标准差。由图2可知，该方法可以较好的分辨出图形，在不同的数据量下，聚类表现稳定，且可以有效筛选离群点。

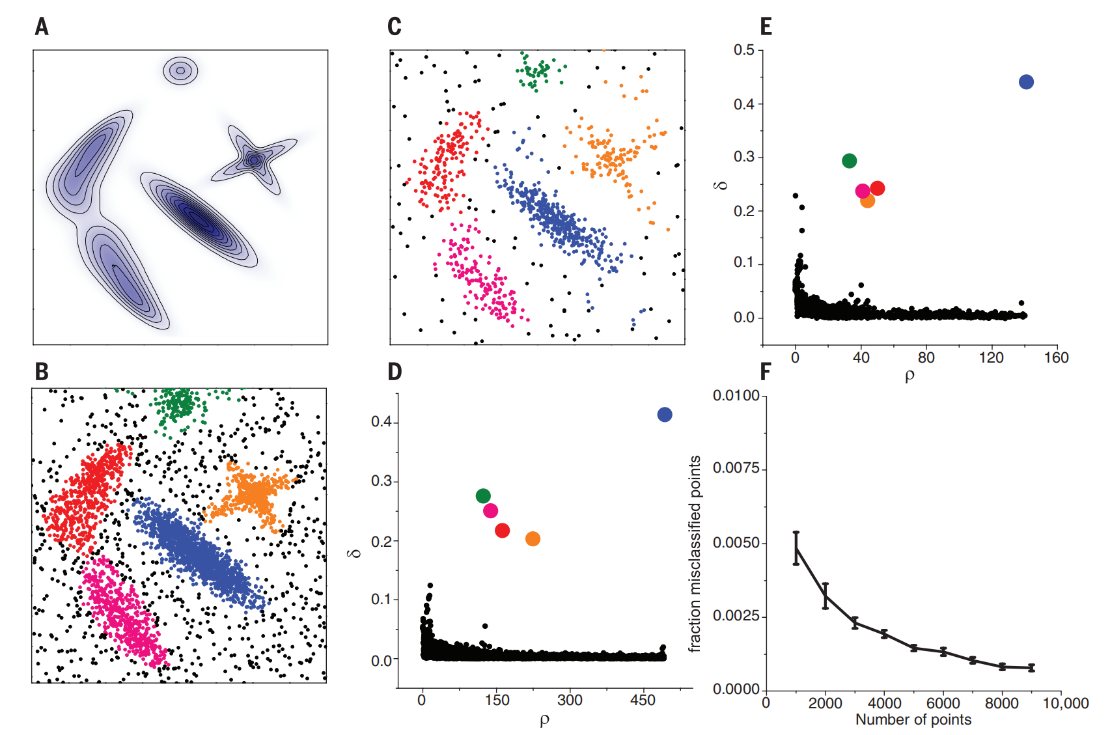


图2[[2]](#footnote-1)：多视角数据

CFDP算法有如下特点：(1)CFPD算法是基于密度的快速聚类算法，其方法的核心思想是类簇中心被一些局部密度低于它的点环绕，即中心密度高。二是类簇中心和其他高局部密度的点间相距较大。(2)CFDP算法将非类簇中心点的聚类过程分离出来，变为一个独立过程。将类簇中心的选择和非聚类点归类两部分分离开来，使得聚类精度增大。(3)此算法广泛适用于图片和非球形点集聚类。(4)该算法思想巧妙简洁，创新性较高，可研究性强。

## **（三）多视图降维及MVDA算法**

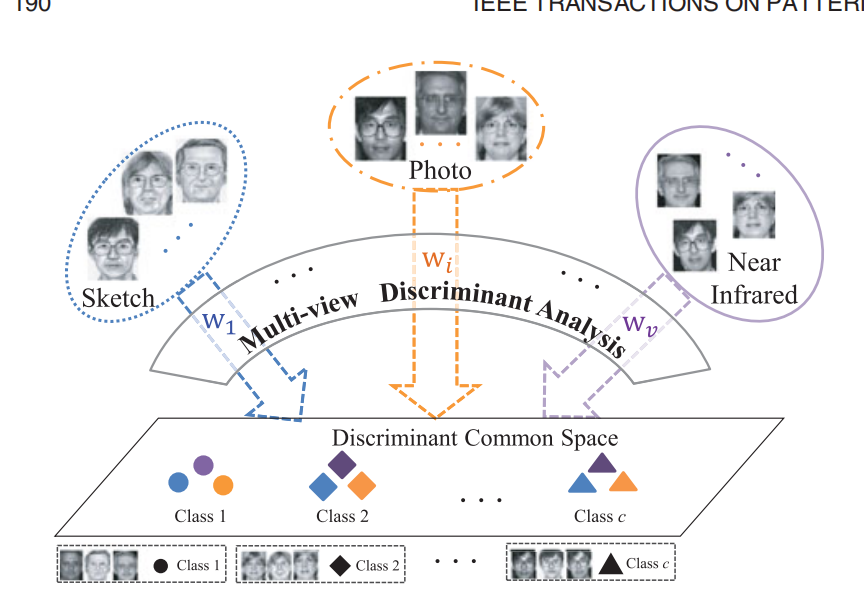
在众多的多视图降维方法中，MVDA算法是有监督的一种，MVDA算法的基本原理及思想是找到个视图相关的线性转换矩阵，使得有一个联合的判别公共子空间，可以让来自个视图的样本都映射到这里。在这个空间内，既要考虑数据在视图内的分布特点，也要考虑数据在视图间的分布特点。这样使得一个类的样本在一起，不同类的样本分离。MVDA联合学习多个视图相关的线性投影，形成一个公共判别子空间。原理示意如图3：

图3[[3]](#footnote-2)：MVDA算法示意图

在该算法中，一组数据集 为第个视图上的样本集，为样本总数， ∈表示维空间内,第个视图上的属于第类样本集的第个样本。其中，是类别数，是第个视图上类别为的样本个数。将所有样本分别通过*V*个线性转换投影到一个子空间，得到数据集，联合投影。在这个公共子空间内，所有视图上的类间离散度矩阵最大化，类内离散度矩阵最小化。定义目标函数：

（5）

其中，为类间离散度矩阵，为类内离散度矩阵。低维公共子空间中类内离散度矩阵表达式为：

（6）

（7）

低维公共子空间中类间离散度矩阵表达式为：

（8）

（9）

将类内离散度矩阵因式乘开，又可以得到以下形式：

= （10）

其中，

（11）

类似的，类间离散度矩阵又可以写成以下形式：

（12）

其中，

（13）

F=,G=

综上，可以得到多视图下的联合的投影矩阵，利用该投影矩阵可以获得一个联合的公共判别子空间。

多视图判别分析算法如下：

|  |
| --- |
| MVDA算法步骤 |
| 输入：多视图数据集={},样本类标{},聚类个数为，视图数。 |
| 1：求得子空间类内离散度矩阵 |
| 2：求得子空间类间离散度矩阵 |
| 3：求得投影矩阵 |
| 输出：投影矩阵。 |

## **（四）算法总流程**

综合上述内容，本文所研究的多视角聚类主要算法思路为：

1.在所收集多视角数据集中选取一单视角进行聚类分析，得到初始标签信息，作为降维输入标签。

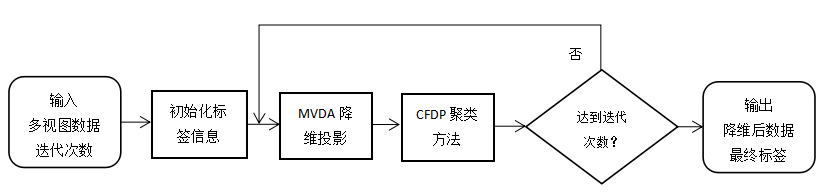
2.将多视角数据集和初始化标签信息作为输入，同时考虑视角间和视角内的变量关系，由多视角判别分析方法（MVDA）对数据进行降维，得到投影矩阵，并且利用投影矩阵将原始空间多视角数据集投影到低维空间，获得子空间内新的多视角数据集。

3.利用新数据矩阵计算各样本间距离，得到样本距离矩阵。

4.输入距离矩阵,通过CFDP聚类方法，获取新的对应标签信息。

5.将多视角所得标签信息作为降维输入标签，再次进行降维并聚类，进行多次迭代。

6.输出最终子空间多视角数据集和最终的标签信息作为结果。

图4：算法流程图

## **五、测试结果分析**

## **（一）数据集**

本文采用UCI上的两个多视角数据集进行实验，由于技术和运行环境的限制，选取的数据集规模较小。数据集的详细信息如下：

1. **Image segmentation data set图像数据集**

从一个户外图像的数据库中随机抽取了2310个样本，对图像进行分割，抽取2个特征（即2个视角）包括shape view（包含9个形状信息属性），RGB view（包含10个颜色信息属性）。选择其中210个用于训练，2100个用于测试。有shape view，RGB view 2个视角，brickface, sky, foliage, cement, window, path, grass 7个类别。该数据集在多篇关于聚类、分类和机器学习的文章中被使用，是一个比较常用和权威的数据集。

1. **Mfeat手写数字数据集**

包含0-9共10个数字的6个特征数据集，每类20个样本，共200个样本。从二值化手写体数字图像中抽取的6个特征（即6个视角）包括侧里叶系数、轮相关特征、 Kahuna-Level展开系数、像素平均、 Zemke矩和形态学特征，对应的数据集合名称和样本维数分别为(mfeat\_fou，76），(mfeat\_fac，216)、(mfeat\_kar，64)、(mfeat\_pix，240)、(mfeat\_zer，47)和(mfeat\_mor，6)。可分别将其简记为Fac、Fou、K、Pix、Zer和Mor。有6个角度，数字0-9 共10个类别，是多视角学习中较为权威的一个数据集。

## **（二）聚类评价指标**

常用的聚类指标分为内部指标和外部指标，我们在本文中主要使用标准化互信息NMI、F-measure和RI对实验结果进行评价，现对此三种评价指标进行简要的介绍：

**1.NMI评价指标**

NMI计算公式：

（14）

为两个聚类结果的标签信息，表示和的互信息，表示的熵。具体计算方法如下：

（15）

（16）

（17）

如果表示聚类结果标签，表示样本的真是标签，通过计算NMI，得到一个介于0和1之间的值，值越大表示聚类效果越好。

**2.F-measure评价指标**

在一个聚类中，对任意簇,相对于真实划分的准确率计算公式如下：

（18）

聚类结果第簇相对于真实划分的召回率计算公式如下：

（19）

其中，表示第簇与划分中相同样本对象的数目，表示簇的样本对象的数目，表示真实划分的样本对象的数目。则，第簇相对与划分的F-measure计算公式如下：

（20）

计算聚类结果中的每一个簇相对于全部真实划分的F-measure值。则对于簇，其F-measure值定义如下：

（21）

对于整个聚类结果的F-measure值由每个聚类划分的加权平均得到：

（22）

F-measure值是一个0到1之间的值，对于相同的数据集，F-measure值越大表示聚类结果越好。

1. **RI评价指标**

假设数据集的真实划分为,对于聚类划分,有以下定义：

（1）一对样本对象在聚类划分和真实划分中都在同一个簇中，全部的这样的样本对的数目用表示。

（2）一对样本对象在聚类划分和真实划分中都不在同一个簇中，全部的这样的样本对的数目用表示。

（3）一对样本对象在聚类划分中在同一个簇中，但在真实划分中不在同一个簇中，全部的这样的样本对的数目用表示。

（4）一对样本对象在聚类划分中不在同一个簇中，但在真实划分中在同一个簇中，全部的这样的样本对的数目用表示。

则，RI的计算公式如下：

（23）

RI值是一个0到1之间的值，对于相同的数据集，RI值越大表示聚类结果越好。

## **（三）算法实验结果及分析**

在本文实验中，选择k-means算法得到每个视图的初始类标签，输入到MVDA算法中进行降维得到新的数据集，再利用欧式距离公式得到距离矩阵，输入到CFDP算法中得到聚类结果，将得到的聚类标签再次输入MVDA算法进行迭代，直至类标签不变为止。本文用Matlab代码实现算法，其中CFDP算法的运行情况如下图5和图6所示：

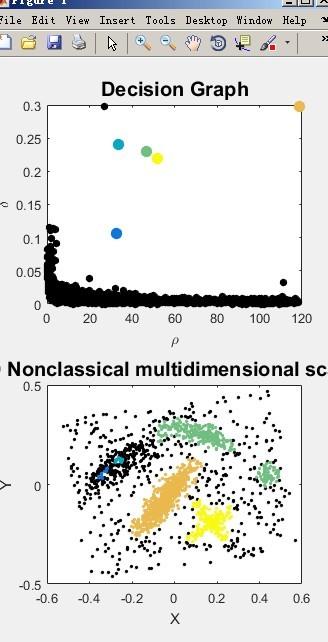


图5：CFDP算法--决策图

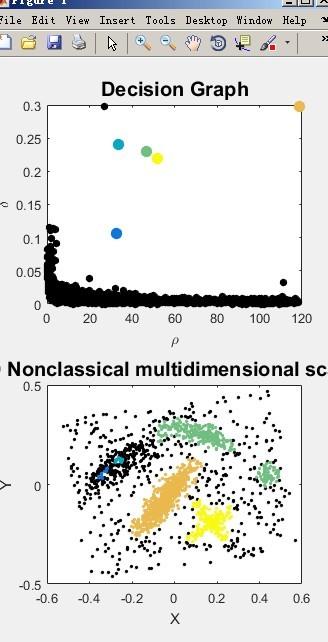


图6：CFDP算法--聚类情况

在实验过程中，由于受个别异常数据的影响和计算样本间距离的需要，我们对数据进行预处理，包括剔除异常数据，归一化等操作。初始的算法参数也有不太合理之处，通过划分training set和testing set进行调整。为了获得稳定的结果，每个数据集用算法运行20次，实验结果取平均值，实验结果汇总信息如表1和表2所示。

表1：Image图像数据集实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | NMI | RI | F-measure |
| 1 | 初始化视角1 | 0.317491 | 0.761267 | 0.41767 |
| 2 | 初始化视角2 | 0.452826 | 0.757029 | 0.452477 |
| 3 | 迭代1次 | 0.49167 | 0.842789 | 0.428764 |
| 4 | 迭代2次 | 0.496774 | 0.853315 | 0.44744 |
| 5 | 迭代3次 | 0.53543 | 0.866483 | 0.457024 |
| … | … | … | … | … |
| n | 最终结果 | 0.55419 | 0.876008 | 0.47799 |

表2 ：Mfeat手写数字数据集实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | NMI | RI | F-measure |
| 1 | 初始化视角1 | 0.629142 | 0.906205 | 0.6559 |
| 2 | 初始化视角2 | 0.653737 | 0.905207 | 0.699295 |
| 3 | 初始化视角3 | 0.700358 | 0.912131 | 0.779376 |
| 4 | 初始化视角4 | 0.670017 | 0.9111934 | 0.682897 |
| 5 | 初始化视角5 | 0.503955 | 0.879957 | 0.556064 |
| 6 | 初始化视角6 | 0.472442 | 0.860423 | 0.553083 |
| 7 | 迭代1次 | 0.549053 | 0.91235 | 0.512635 |
| 8 | 迭代2次 | 0.6177331 | 0.91627 | 0.532155 |
| … | … | … | … | … |
| n | 最终结果 | 0.709816 | 0.928327 | 0.788431 |

由表1可以看出，使用k-means算法在Image数据集单试角上聚类的NMI值分别为0.317491和0.452826，而用本文所提出的多视角聚类算法得到最终结果的NMI值为0.55419，相较之前有了很大的提升，可以看出考虑多视角因素的结果相比在单视角上聚类的准确性更高。由表1还可以发现，NMI值随着迭代次数的增加逐渐增大，由0.49167到0.53543再到最后的0.55419，说明由于算法采用修正后的标签指导联合判别子空间的形成，使得下一轮算法的运行可以融合上一轮运行的优化结果，迭代过程具有重要的意义。类似地，用RI和F-measure两个指标也能得到上述的指标值提升，数据集的RI值从0.76左右增大到0.88，F-measure值由0.44左右增大到0.48，可以得出本文提出的算法相比传统的聚类算法得到的结果更好。

表2是在Mfeat数据集上进行实验得到的结果，使用k-means算法在单视角上聚类的NMI平均值约为0.60，用本文算法得到最终结果的NMI值为0.70，聚类效果提升明显。同时，NMI值随着迭代次数的增加逐渐增大，由迭代一次的0.549053到最后的0.709816，迭代提升效果较好。类似地，用RI和F-measure两个指标也能得到上述的指标值提升，数据集的RI值从0.89左右增大到0.92，F-measure值由0.65左右增大到0.78。

不同指标的计算方法会影响指标值的大小，一般来说，RI指标的结果相较其他两个要更大。综合来看，不管使用哪一个指标，用多视角迭代聚类算法得到结果的指标值都是最大的，聚类正确率随迭代次数的增加逐渐提高。通过与k-means算法得到的初始化视角比较，我们可以看出考虑多视角因素的聚类效果相比在单视角上的要更好，说明了结合各个视角特征考虑聚类问题的重要性，这也是我们算法的优势所在。

为了研究基础聚类算法对算法迭代结果的影响，我们在上述基础上将所迭代的聚类算法分别调整为k-means聚类（典型的基于距离的聚类算法）和层次聚类（通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树算法），在Image图像数据集上，进行对比实验。实验结果汇总信息如下表3、表4、表5所示。

表3：Image图像数据集NMI对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | CFDP | K-means | 层次聚类 |
| 1 | 迭代1次 | 0.49167 | 0.244998 | 0.091519 |
| 2 | 迭代2次 | 0.496774 | 0.420391 | 0.462189 |
| … | … | … | … | … |
| n | 最终结果 | 0.55419 | 0.444875 | 0.508133 |

表4：Image图像数据集RI对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | CFDP | K-means | 层次聚类 |
| 1 | 迭代1次 | 0.842789 | 0.656869 | 0.239143 |
| 2 | 迭代2次 | 0.853315 | 0.767236 | 0.599316 |
| … | … | … | … | … |
| n | 最终结果 | 0.876008 | 0.805013 | :0.483390 |

表5：Image图像数据集F-measure对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | CFDP | K-means | 层次聚类 |
| 1 | 迭代1次 | 0.428764 | 0.373923 | 0.254483 |
| 2 | 迭代2次 | 0.44744 | 0.484554 | 0.468851 |
| … | … | … | … | … |
| n | 最终结果 | 0.47799 | 0.519653 | 0.505539 |

由表3可以看出，使用CFDP算法，k-means算法和层次聚类算法在Image数据集单视角上聚类的NMI值分别为0.55419,0.444875,0.50813，其中CFDP算法的NMI最高，聚类效果最好。同样的在表4的实验结果里，CFDP算法的RI值最高，表明CFDP算法优于k-means算法和层次聚类算法。RI和NMI值体现出本文提出的算法相较传统的聚类算法结果更优。但在表5中，CFDP算法的F-measure值偏低，可能是由于算法存在未知缺陷或者数据集数量过少存在实验偶然性。

为了进一步分析距离公式对算法性能的影响，我们将以CFDP聚类为基础的算法在Mfeat数据集上采用不同距离公式进行实验，结果如表6所示。

表6：Mfeat数据集 采用不同距离公式结果对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | NMI | RI | F-measure |
| 欧式  距离 | 迭代1次 | 0.549053 | 0.91235 | 0.512635 |
| 迭代2次 | 0.617733 | 0.91627 | 0.532155 |
| … | … | … | … |
| 最终结果 | 0.709816 | 0.928327 | 0.788431 |
| 标准  欧式  距离 | 迭代1次 | 0.559053 | 0.909001 | 0.557141 |
| 迭代2次 | 0.625436 | 0.923794 | 0.673518 |
| … | … | … | … |
| 最终结果 | 0.704383 | 0.937898 | 0.742251 |
|  | 迭代1次 | 0.559053 | 0.92235 | 0.502635 |
| 曼哈顿 | 迭代2次 | 0.646669 | 0.923794 | 0.504747 |
| 距离 | … | … | … | … |
|  | 最终结果 | 0.70282 | 0.930335 | 0.776338 |
| 闵氏  距离 | 迭代1次 | 0.647282 | 0.889503 | 0.623943 |
| 迭代2次 | 0.680215 | 0.912893 | 0.638869 |
| … | … | … | … |
| 最终结果 | 0.709051 | 0.920651 | 0.697705 |

由表6可以得出，在NMI和RI两个指标上，不同距离公式数值差距不明显。主要差距体现在F-Measure结果上，其中欧氏距离表现最好达到0.788，闵式距离表现最差为0.697。综合来看，算法采用不同距离公式结果虽然存在一定差异，但整体较为稳定，鲁棒性较好。

## **（四）优缺点总结**

由以上实验结果可以看出，本文所用的多视角迭代算法正确率较高，相比单视角下的聚类有显著的优势，聚类效果随迭代次数不断提升，也证明了迭代的重要性。但是该算法存在一定的缺陷，利用MVDA算法将多个视角投影到一个公共空间时，数据的维度会变大，对于高维数据来说，后续计算距离矩阵会比较麻烦。关于迭代过程，它在提高准确性的同时也增大了运行时间复杂度，当样本数量很多时，要满足输入输出类标签一致需要迭代非常多次，所以在实际应用时，可能需要设置迭代次数。

# 结论

聚类分析是数据挖掘、机器学习、模式识别和计算机视觉等领域的基础研究课题之一，在大数据背景下对其进行研究具有十分重要的现实意义。虽然聚类算法得到了极大的发展，但是并不存在一个能完美适用于多种数据结构的聚类算法。在现实情况中，许多数据都是从多个角度描绘的，这就需要我们引入多视角数据进行聚类分析。本文就是在多视角降维的基础之上采用一种新型的快速聚类算法进行聚类，不仅考虑到各个视角的特征，还提高了聚类的准确性。

本文算法的基本思路：首先输入数据集，利用单一视角聚类方法得到初始化标签信息。将多视角数据集和初始化标签信息作为输入，利用MVDA算法对数据进行降维，将原始的多视角数据集投影到低维空间，获得子空间内新的数据集。再在新的特征空间里计算每个样本的距离，得到距离矩阵，运用CFDP算法进行聚类，得到对应新的标签信息。最后将新的标签信息和数据集作为降维过程中的输入，多次迭代使类标签一致，得到聚类结果。

本文采用了UCI的两个数据集对算法进行实验，聚类效果随迭代次数增加逐渐提高，最终的聚类结果准确性也比单试图上的要好得多，体现了算法的优势。同时，在改变初始聚类方法和距离公式的情况下，聚类效果依旧较好，说明了算法具有一定的鲁棒性。但是由于技术和运行环境的限制，我们没有用大规模数据集进行检验，算法关于时间复杂度和维度方面的缺陷需要进行更深入的研究，与其他多视角聚类方法的比较也需要进一步探讨。

# 参考文献

[1]邓强. 多视角子空间聚类集成方法研究及分布式实现[D].西南交通大学,2016.

[2]何云. 面向多视角数据的降维与聚类算法研究[D].南京师范大学,2019.

[3]李学龙,陈穆林,王琦.光学影像序列中基于多视角聚类的群组行为分析[J].中国科学:信息科学,2018,48(09):1227-1241.

[4]刘正,张国印,陈志远.基于特征加权和非负矩阵分解的多视角聚类算法[J].电子学报,2016,44(03):535-540.

[5]张恩豪,陈晓红,刘鸿,朱玉莲.基于典型相关分析的多视角降维算法综述[J].计算机工程,2020,46(02):1-10.

[6]张媛,贾克斌,ZHANG Aidong.基于多视角融合的蛋白质功能模块检测方法[J].电子学报,2014,(12):2337-2344. DOI:10.3969/j.issn.0372-2112.2014.12.001.

[7] Issam Falih,Nistor Grozavu,Rushed Kanawati,Younès Bennani. Topological multi-view clustering for collaborative filtering[J]. Procedia Computer Science,2018,144.

[8] Kan M，Shan S，Zhang H，et al. Multi-View Discriminant Analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 38(1):188-194.

[9] Nogueira M,Piella Fenoy G,Sanchez Martinez S,et al. Characterizing patterns of response during mild stress-testing in continuous echocardiography recordings using a multiview dimensionality reduction technique[J].2017.

[10] Stefan Rüping, Scheffer T， Learning With Multiple Views[J]. Proc Icml Workshop on Learning with Multiple Views, 2005, 27(1):203-233.

[11] SU Shuzhi. Theoretical research and application of multi-view canonical correlation analysis［D］． Wuxi: Jiangnan University，2017．

[12] Tao Qian,Gu Chunqin,Wang Zhenyu,Jiang Daoning. An intelligent clustering algorithm for high-dimensional multiview data in big data applications[J]. Neurocomputing,2019(prepublish).

[13] Weizhong Yu,Rong Wang,Feiping Nie,Fei Wang. Multi-view embedded clustering with unsupervised trace ratio LDA[J]. Neurocomputing,2018,315.

[14] Yan Fei, Wang Xiao-dong,Zeng Zhi-qiang,Hong Chao-qun. Adaptive multi-view subspace clustering for high-dimensional data[J].Pattern Recognition Letters, 2020,130(C).

1. 图源自文献《A survey on multi-view learning》 [↑](#footnote-ref-0)
2. 图源自文献《Clustering by fast search and find of density peaks》 [↑](#footnote-ref-1)
3. 图源自文献《Multi-View Discriminant Analysis》 [↑](#footnote-ref-2)