学校代号 10532　　学　　号 S1710W0868

分 类 号 TP312　　密　　级 普 通

工程硕士学位论文

LINCS 数据相似性度量研究

学位申请人姓名 刘 伟

培 养 单 位 信息科学与工程学院

导师姓名及职称 彭绍亮 教授　　邓子云 高工

学 科 专 业 软件工程

研 究 方 向 生物信息学

论文提交日期 2019年04月19日

学校代号：10532

学　　号：S1710W0868

密　　级：普通

湖南大学工程硕士学位论文

LINCS 数据相似性度量研究

学位申请人姓名　　刘 伟

导师姓名及职称　　彭绍亮 教授　　邓子云 高工

培 养 单 位　　信息科学与工程学院

专 业 名 称　　软件工程

论文提交日期　　2019年4月19日

论文答辩日期　　2019年4月26日

答辩委员会主席　　王 东 教授

The Research of Similarity for LINCS Biological Data via Metric Learning

by

WEI LIU

B.E.(University of Shanghai for Science and Technology) 2013

A thesis submitted in partial satisfaction of

the Requirements for the degree of

Master of Engineering

in

Software Engineering

in the

Graduate School

of

Hunan University

Supervisor

Professor Shaoliang Peng

Associate Researcher Ziyun Deng

April, 2019

湖 南 大 学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容以外，本论文不包含任何其他个人或集体已经发表或公开的作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果并由本人承担。

作者签名： 日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1．保密□,在_____年解密后适用本授权书。

2．不保密□。

(请在以上相应括号内打”√”)

作者签名： 日期： 年 月 日

导师签名： 日期： 年 月 日

摘　要

LINCS 是近期公开的一项大数据计划，它基于典型人体细胞在小分子化合物刺激下的反应而测定，数据丰富而规整，配套处理工具完善。由于基因表达具有高度相关性，探究 LINCS 基因表达的相似性对于基因推断、药物发现、多组数据融合分析、通路发现等具有重要的意义和参考价值。GSEA 算法是目前研究 LINCS 数据相似性的主流算法，需要先预测实验结果然后再进行计算比对，由于其计算过程的复杂性，GSEA 算法在相似性判定和时间开销上难以满足海量表达谱数据的分析需求。度量学习算法立足点在于学习，通过学习训练数据获得适合的度量空间从而进行相似度的计算，是比较理想的表达谱相似性判定方法，目前针对表达谱数据尤其是 LINCS 数据相似度分析的度量学习模型很少。基于此，本文以 LINCS 数据之间的相似度为出发点搭建了两种不同的的度量学习模型，另外，本文还提出了新的分类方法以拓展 LINCS 数据相似性的应用。课题主要工作包括：

1．基于改进余弦距离的基因表达谱距离度量算法。本文首先提出基于 H5py 的数据提取优化方法用于 LINCS 数据提取，然后通过实验得出余弦距离是较为适合的相似度计算函数，接着对余弦距离进行改进，通过中心化和归一化使得算法对于基因表达谱各维度上的值更加敏感，结合 NCA 算法，组成基于改进余弦距离的近邻成分分析度量算法。在多个数据集上验证得出，该算法是一种对于基因表达谱相似度分析较为适用的度量算法。

2．基于深度学习的基因表达谱距离度量算法。本文基于 Siamese 框架,构建 DenseNet 网络和 Cosine 距离相结合的深度学习模型，拓展了隐式的度量学习，采用 Center loss 和 Cross-entropy loss 结合的损失函数计算损失，在减少人工干预的同时提高了模型学习到的高层次特征表达的判别性。该方法的一个关键点在于数据的转换处理，需要预先将基因表达谱转换成基因方阵。在多组细胞系数据验证得出，该算法度量效果效果远远好于常用的度量学习方法和 GSEA 算法。

3．基于共享字典学习的 LINCS 数据分类算法。本文设计了一种基于判别投影的共享字典学习模型，在训练字典时，还训练投影矩阵，并且投影矩阵对测试样本的投影可以加宽不同类型样本之间的距离间隔。另外，通过共享性能获取所有类别的样本，提高分类的判别性。最后，利用重构误差和均值向量之间的距离来判定样本的类别。在多组实验数据验证得出，该方法的分类准确率要高于当前主流分类方法。

关键词：LINCS；相似度分析；GSEA；度量学习；基因表达谱；深度学习

Abstract

LINCS is a recently announced big data plan based on the response of typical human cells stimulated by small molecule compounds. The data is rich and regular, and the processing tools are perfect. Because gene expression is highly correlated, exploring the similarity of LINCS gene expression has important significance and reference value for gene inference, drug discovery, multi-group data fusion analysis, pathway discovery and so on. The GSEA algorithm is currently the mainstream algorithm for studying the similarity of LINCS data. It is necessary to predict the experimental results first and then perform the calculation comparison. Due to the complexity of the calculation process, the GSEA algorithm is difficult to satisfy the massive expression spectrum data in the similarity judgment and time overhead. Analyze the demand. The metric learning algorithm is based on learning. It is an ideal method for judging the similarity of expressions by learning the training data to obtain the appropriate metric space. It is currently a measure learning method for expressing spectral data, especially LINCS data similarity analysis. There are very few models. Based on this, this paper builds two different metric learning models based on the similarity between LINCS data. Besides, this paper also proposes a new LINCS data classification method to extend the application of similarity judgment. The main work includes:

1. Gene expression profiling distance metric algorithm based on improved cosine distance. This paper first proposes a data extraction optimization method based on H5py for LINCS data extraction, and then finds that the cosine distance is a suitable similarity calculation function. The cosine distance is improved in the next step, which makes the algorithm is more sensitive to values in each dimension by centering and normalization. A near-neighbor component analysis metric algorithm based on improved cosine distance constructed by combining with the NCA algorithm. It is verified on multiple datasets that the algorithm is a metric algorithm that is suitable for similarity analysis of the gene expression profile.

2. Gene expression profiling distance metric algorithm based on deep learning. Based on the Siamese framework, this paper constructs a deep learning model combining DenseNet network and Cosine distance and expands the implicit metric learning. The loss function combined with Center loss and Cross-entropy loss is used to calculate the loss, and the model is improved while reducing manual intervention.

The discriminative of the high-level feature expressions learned. A key point of this method is the data conversion process, which requires the gene expression profile to be converted into a gene matrix in advance. It is verified in the data of multiple groups of cell lines that the algorithm measures far better than the commonly used metric learning method and GSEA algorithm.

3. LINCS data classification algorithm based on shared dictionary learning. In this paper, a shared dictionary learning model based on discriminant projection is designed. When training the dictionary, the projection matrix is also trained, and the projection of the projection matrix to the test sample can widen the distance between different types of samples. In addition, the classification of all categories is obtained by sharing performance, and the discriminability of classification is improved. Finally, the distance between the reconstruction error and the mean vector is used to determine the class of the sample. It is verified by multiple sets of experimental data that the classification accuracy of this method is higher than the current mainstream classification method.

Key Words：LINCS; Similarity analysis; GSEA; Deep Learning; Metric Learning; Gene Expression Profile

目　录

插图索引

附表索引

第一章 绪 论

全基因组基因表达谱（GWGEP）对于生物信息学及相关领域的研究而言，可以算是其中的核心内容，GWGEP 能够将生物形态表征和 DNA 中上的重要编码信息结合起来，GWGEP 的相关分析可以让研究者人员在特定实验条件在不同时间点测量细胞或组织的转录水平。

其研究价值是：（1）运用差异分析等方法，来探究差异表达在生物过程中或者在疾病发病机制的作用；（2）挖掘特定表型和基因表达模式与内外部扰动之间的关联；（3）判别可能的分子标志物，并将其用于疾病临床诊断；（4）发掘药物开发中新的可能目标或者有用的治疗方案等。

随着转录组测序技术的蓬勃发展，不仅能够通过相对简单的实验来检测所需的 GWGEP 数据，而且还能检测出各种各样癌症和病毒感染。GWGEP 数据的探究和处理，可以发现细胞中各种基因的异常表达并确定导致细胞状态模式的关键因子，以此来获得转录表达水平及细胞特定的状态。对于各种小分子相关作用机制的深入探究，还能够推导出恢复细胞正常状态的药物或者特定的治疗方案。

1.1 课题背景和意义

1.1.1 课题来源

本课题源自于 2018 年至 2020 年国家重点研发计划项目《精准医学大数据的有效挖掘与关键信息技术研发》（2018YFC0910405），和深圳市科技计划项目《面向生物大数据药物重定位的小样本机器学习方法》（JCYJ20170818110101726）。在这些项目中，北京军事医学科学院针对海量生物数据和药物重定位问题提出了一些生物医药大数据和机器学习相结合的处理方法。实验数据来自于部分医院以及大型开发项目的数据库，其中 NIH LINCS 项目是重要的数据来源。本论文主要针对 LINCS 生物数据的分析和挖掘而开展。

1.1.2 研究背景

全基因组表达谱的数据分析和研究一般针对小分子在细胞中的功效而开展，研究的主要目标是小分子下的调节途径[1]。数据分析能够让研究者了解到小分子的作用机理、调节作用、作用方式等，但是这种分析很可能因为其来源和检测标准等存在的问题而导致不准确性，或者称为分析的片面性。CMAP [2]数据库的公布为解决这个问题提供了资源和途径，美国的 BROAD 研究所 2006 年就公开了这

个数据库。CMAP 包含的化合物有 1,309 个，全基因组基因超过 6,100 个，这些数据基于 5 个细胞系的检测而产生。CMAP 项目分析过程如图 1.1[3]所示。

表达谱数据的相似度水平，可以经过有差异的表达谱的印记基因集（SGS）的富集分析而获得。表达谱的差异性可以这样来描述，存在两种可能，一是如果两个表达谱相似，则能够认为这两个表达谱之间存在诱导因子的叠加性，二者的表征可能得到加强；二是表达谱不相似，则和前面的规则相反，可以认为二者之间的诱导因子会产生相斥作用，二者之间的相互作用表征会相互排斥。所以，相似性表述可以用来实现细胞状态的恢复或者逆转。

基于以上理论，CMAP（连通图）数据比较普及而且被运用于多个领域，例如药物重新定位（DRP），关键因素发现（KFD）和作用模式（MoA）。然而，由于数据量的制约存在，比如实验药物类型、细胞系类型、试验时间和试验剂量等，CMAP 数据的探究被限定在一定的范围。LINCS[3][4]（基于细胞签名的综合网络数据库）也由 BROAD 研究所开发并且广泛应用，该计划已经公布 130 万个全基因组表达谱，这些表达谱来源于 77 种典型人体细胞系，它们是在 4,000 多种基因沉默试剂和 7,000 多个化学物小分子刺激下检测得出。

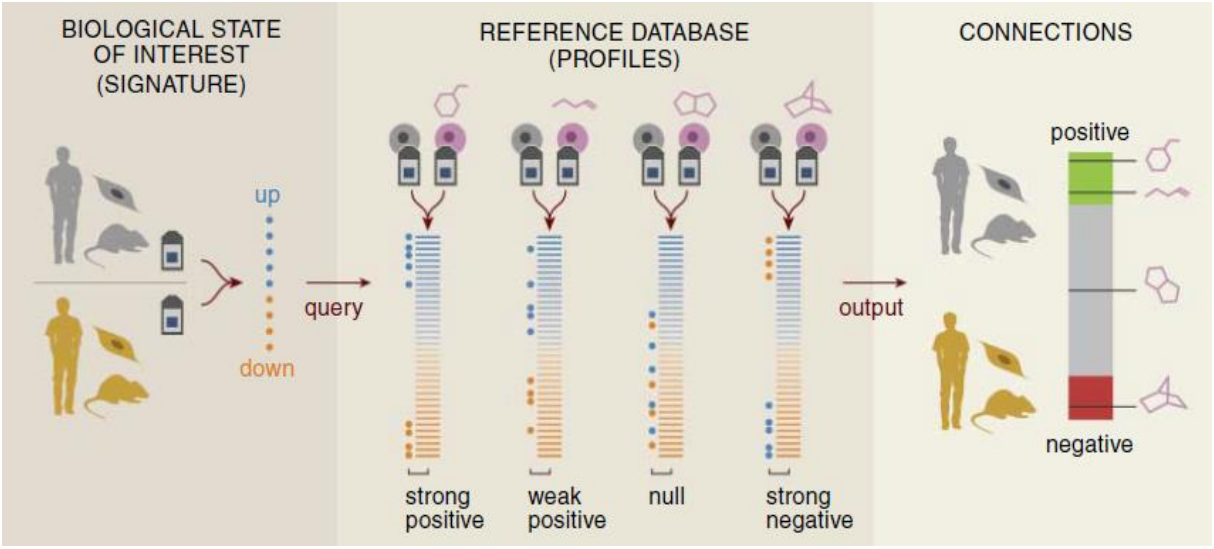

图 1.1 Cmap 项目分析过程

人类基因组测序计划自 2001 年开展开来，海量的数据不断积累。然而，分析这些庞大而多样化的数据非常困难。由于生物信息学具有综合性，需要用到多方面的知识如数学、生物、计算机等。它可以捕获、处理、解释规模的生物数据，是生命科学研究领域的前沿。为了更深入地了解人类生命机制，由结构基因组学相关研究转变成为功能基因组学相关研究已经是当今生物信息学研究的重点和热点。功能基因组学，用于探究个体基因的功能作用，以解释基因与疾病之间的关联。在后基因组学研究中，使用基因芯片技术制备的大量基因表达谱成为研究人员重要的数据来源。如何系统地处理、分析和解释这些表达谱，挖掘这些数据之

间的内在联系已成为当前生物信息学中的热点问题。

1.1.3 问题描述

不难看出，在基因表达谱应用分析方面，计算其相似度是一个基础却非常重要但是容易被忽略的难点。LINCS 数据的相似度探究，就是研究转录组表达谱之间的差异性表达。换言之，就是通过表达谱两两之间的比对来衡量表达谱的差异性。一般是通过计算二者之间的距离的方式来衡量差异性，而获得的差异性数据能够为下一步的研究提供相关的试验数据和基础。设想这样一个问题，因为表达谱数据的逐年递增，寻找表达谱之间相似的基因变得越来越困难。而当表达谱维度高达成千上万的时候，简单的线性欧氏距离虽然能算出相似度，但是效果却不好，因此急需新的计算工具和方法应用于此问题，为后续的 LINCS 数据持续性研究提供可靠的数据资源，并且这些工具或者方法能广泛适用于表达谱相似度计算。

GSEA[5](Gene Set Enrichment Analysis，基因集富集分析)算法是目前计算 LINCS 表达谱相似性的主流方法。这个算法提出了富集积分这样一种概念，是一种基于 kologorov-smirnov 的排序统计计算方法，需要结合重点探究和多种假定测验办法对取得的积分和测量结论的可靠性开展统计分析。但 GSEA 是一种先验算法，需要先对实验结果进行预测然后再进行生物比对，而且受限于其本身复杂的计算过程，GSEA 算法在相似度性能判定和时间开销上不能让人满意，也满足不了大规模的表达谱分析和运算。具体的基因印记和 GSEA 相似度计算过程如图 1.2[6]所示：

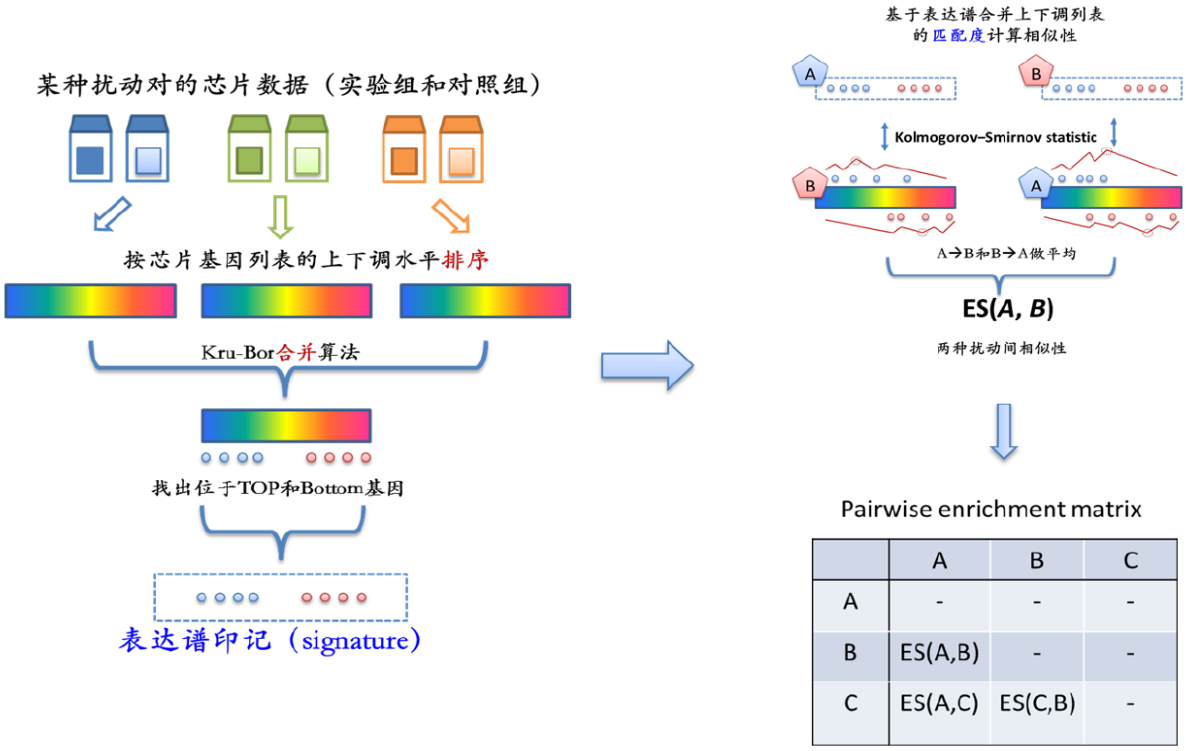

图 1.2 表达谱印记和 GSEA 相似度计算过程

而度量学习（Metric learning）[19]能够依据具体的工作来主动学习获得特定工作下的度量距离函数，度量学习原始的目标是自动学习出相应度量空间，基于这个空间，能够使同类样本的间隔尽可能减小，而使有差异的样本之间的间隔尽可能增大。和 GSEA 相比，度量学习通过学习去判断其相似性，通过搭建模型、调整参数积累数据量达到优化，二者有着本质区别。用于 LINCS 方向的度量学习模型很少，不过度量学习已经在比较大的范围内有着非常多的运用案例，已成熟应用于计算机视觉、图像检索和分类、姿势预估等，有较高的可迁移性。综上所述，将度量学习应用于 LINCS 数据分类是一个很好的创新和尝试。

LINCS 相似度研究需要遵循这样一个前提：需要有可靠的评价标准来衡量度量的好坏。目前广泛使用的 GSEA[5]富集方法得到了业界认可，是对比衡量的重要指标之一。如果能在度量学习的研究上取得进展，在度量效果和时间开销上面对比 GSEA 有所提升，那么这大大降低了数据研究的成本和研究，丰富了表达谱数据，为数据之间的关联性发掘和具体应用夯实了基础。同时，相似性表达的研究数据是 LINCS 数据集中的对照组，基因沉默下的表达谱，因为在受到外界刺激的时候，差异的刺激会产生差异表达的基因值。此时，需要找寻新的方法比对前后的突出基因因子，这又是另外的一个新的研究内容，这里不做展开。

LINCS 生物大数据一个重要的应用就是针对基因表达谱研究其分类性能。由于 LINCS 数据各种细胞系大部分都是各种人体肿瘤细胞系的表达谱，因此研究其分类具有十分重大的意义。经过 LINCS 表达谱的分类探究，我们可以从分子层面认识肿瘤的致病机理，从本质上认识肿瘤，并为彻底治疗肿瘤提供基因层次的解决方案。同时，借助基因表达谱数据利用计算机技术对肿瘤进行鉴定和分类还有许多常规医学分类方法不可比拟的优点。另外，由于 LINCS 数据相比较 GEO 和 NCI 等规模较大适用性较强的肿瘤基因表达谱数据库而言，在来源、标准、处理上的严格控制和统一，用 LINCS 数据进行分类也相比较更加可靠。

1.1.4 研究意义

目前 LINCS 数据的分析一般是依据印迹基因集比对的结果来判定。印迹基因集表示的是一种基因的集合，这种集合基于这样一种状态，即细胞在某种状态中表征到的表达情况在全基因数据中下调、上调较为明显的状态。经过计算基因的差异表达情况，挑选明显上调的基因（差异表达中的表达倍数显著增加）和明显下调基因（差异表达中的倍数表达显著降低）而组成的因子称为印迹基因集。印迹基因组的大小正常情况下比全基因组数据的规模要小的多，往往只是包含几十或者几百个基因。故而，我们能够获得各类细胞系的表达谱数据和印记基因集，这些数据产生于不同的药物分子的扰动。经过查询和对比不同药物扰动之后获得的表达谱数据，进而分析细胞的状态，这很有可能会产生相同或不同的印迹基因

集，然后再探究诱导原因以及对应的治疗药物。

印迹基因是能够通过相关样本的相似度比对来获取的，以此为依据，能够对符合要求的分子化合物进行分析和选择。通过将已知药物的和未直接实验检测的印记基因集进行比对，可以以此来推导相关的化合物的用途和可能存在的效果[6]；另外，作用机理以及药物靶点的潜在性是可以经过相关的试剂的比对结论来进行判定的。

细胞所处的具体形态能够从表达谱的分析比对状况进行推断。比如，肿瘤组织的形态是目前研究者在肿瘤分类问题上的客观的判别标准，这些标准往往基于微阵列技术[7]，研究者非常依赖其状态表现。关于肿瘤的发生机制以及医学治疗方案参考的一个重要标准就是相关细胞表达谱的分类和应用研究，相关人员能够依据基因表达谱的表达差异来辨别状态上类似的肿瘤，基于此项判别的精确诊断对于佳治疗方案的制定能提供科学指导。随着 RNA-seq [8]的逐渐进步，可以精确定量表达的程度，还能够让人精准判定碱基点，进而为疾病预防、诊断带来数据精准的优势。基于相似度的药物推断和诱导因素之间的关系如图 1.3[6]所示。

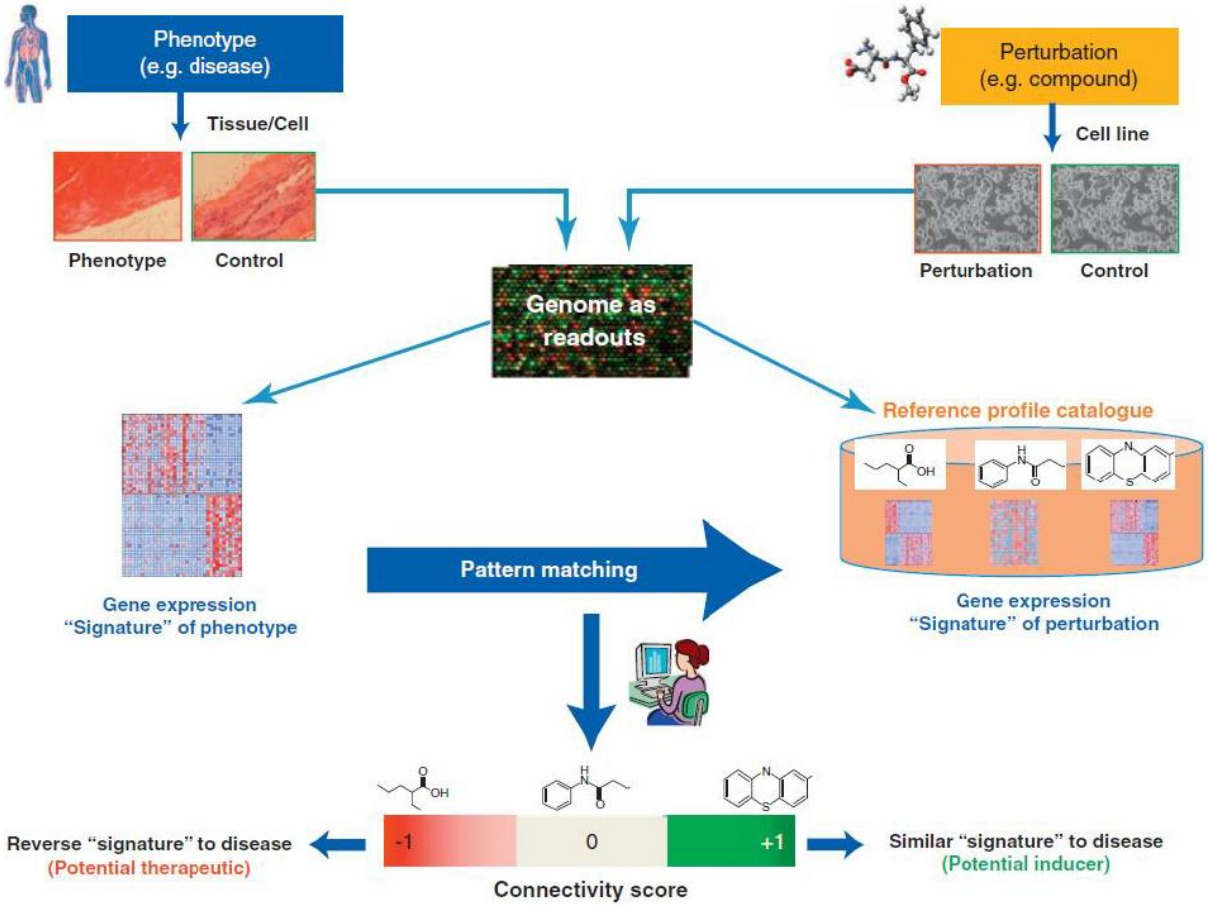

图 1.3 相似度推断药物和诱导因素之间的关系

通过相似性比对可以研究细胞的差异性表达（DEG），DEGB 表示的是当细胞处于差异的实验环境和条件下的时候，表征程度有区别的基因。或者可以这样认

为，DEG 代表的是和正常表达有明显差别的基因。倍数变换（fold change, FC）是差异衡量的基础，它可以用所有实验数据和对照数据的商来表示[9]。通过研究表达基因差异，尤其是表达程度有明显差异的数据，能够精准定位表达明显改变的蛋白或者基因，以此为基础，能够判定潜在的由于基因改变所致使的疾病和细胞形态，更近一步，相关疾病的诊疗也可以调控这些表达来完成。

经过 LINCS 表达谱的分类探究，我们可以从分子层面认识肿瘤的致病机理，从本质上认识肿瘤，并为彻底治疗肿瘤提供基因层次的解决方案。同时，借助基因表达谱数据利用计算机技术对肿瘤进行鉴定和分类还有许多常规医学分类方法不可比拟的优点。

1.2 国内外研究现状

LINCS 项目尽管刚开始发展，就目前已经公布的实验数据上来看，有关的研究普遍已经取得了不错的成果，尤其在基因网络推断、药物发现关键位点分析、疾病和药物的相关联等方面。虽然 LINCS 数据的发掘和应用非常广泛而且越来越受到重视，而度量学习由于图像识别、行人再识别问题的研究变得比较火热，但是将度量学习应用于 LINCS 表达谱的模型和算法是很少的，二者结合的成熟应用更是难以发现，双方的最新发展成果主要包括以下内容。

LINCS 数据的相关研究进展：在基因调控关系方面的应用，如宋欣雨等[10]，通过先验概率联合后验概率并且结合回归模型在基因沉默的数据实验下有了较大进展，他们的研究也在相关转录基因库的关系识别上得到检验，数据库是 JASPAR。在 L1000 数据应用方面，Chen 等[11]的研究团队通过机器学习的分类器来探究药物不良反应，该分类器是基因表达特征结合化学结构而最终形成的。他们还利用分类器试验了差不多 20000 个小分子化合物。最终开发了浏览、 搜索、预测小分子药物不良反应的工具。Otvos 等[12]采用基于机器学习算法 Softmax 的 L1000 药物干扰转录水平数据，适用于多分类领域。挖掘并预测了 480 种已上市药物重定位于其余医疗属性的潜在能力。熊志勇等基于药物与疾病扰动[13]通过 L1000 数据探究了基因表达特征相互逆转关系，并且预估出 4 种化合物的基因表达，最后通过 5 种肝肿瘤细胞系得以检验。

度量学习分类及其研究概述：距离度量学习（DML）[42]的目标是为给定问题学习适当的距离函数。DML 对于许多学习模型非常重要，例如 KNN 和 SVM 分类。根据潜在的学习范例， DML 方法的一种流行分类是：监督 DML[32-36]和无监督 DML [34-41]，还有一些半监督的论文结合了这两种范式[43-44]。运用降低维度的手段将数据映射入低维子空间，从而获取到低维而紧凑的样本关系，是无监督度量算法核心思维，而无监督的核心思维则是通过训练优化目标函数从而获得相应

的距离矩阵，而这个距离矩阵能有效地反映样本在空间上的相互关系。我们的研究建立在有监督的度量学习基础上,因此我们只参看此类别中的一些代表性作品。Xing 等[45]提出了一种用于监督 DML 的经典算法，其中作者为度量学习提出了约束凸优化问题。主成分分析（RCA）[46]利用定义的小块来通过减少不相关维度的权重和放大相关维度的权重来学习度量。Schultz 等[47]为 DML 引入了可以使用查询反馈轻松获得的相对比较约束。该公式是二次规划问题，通过调整标准 SVM 求解器得到解决。邻域分量分析（NCA）[48]学习直接最大化最近邻分类性能的度量。这是通过随机邻域选择优化训练集上的留一法分类错误来实现的。大间隔最近邻居（LMNN）[49]也基于近邻分类，但使用大边缘策略。从信息理论的角度来看，Davis 等 [50]提出在差分相对熵的意义上学习接近给定先前距离度量的 Mahanobis 矩阵，并同时满足距离约束。Jin 等[51]针对规范化 DML 提出了一种有效的在线算法,其中证明了如果使用适当的约束,泛化误差可以独立于特征维数。Yang 等 [52-53]强化了相对距离这样一种概念指标，从而引申出了维持几何形态的多任务 DML 算法。

度量学习在边信息聚类中的具体应用：朱明敏等[14]提出了一种算法，该算法在多维空间中给出相似点对的示例的情况下，学习相对于这些关系的距离度量。该方法的基础是凸优化的转变，即将度量学习推导为凸优化探究这使其具有局部最优解和有效性双重特性，该算法验证得出其在聚类性能上的优越性，学习的度量能够显着提高聚类性能。Goudarzi 等[15]展示了如何通过半正定规划来学习 KNN 分类的 Mahalanobis 距离度量。框架不对数据的结构或分布做出任何设想，并自然地扩展到大量类。他们将 LMNN 分类应用于数百或数千个类的问题，其优势最明显。其次,他们也探究了在非线性特征空间中执行 LMNN 分类的核心技巧。然而，验证得到由于 LMMN 已经在原始输入空间中产生高度非线性决策边界，因此对该算法进行核化将导致进一步的改进变化并不明显。

通过折叠类进行度量学习具体应用：詹增荣[16]提出了一种用于学习二次高斯度量（马哈拉诺比斯距离）的算法，用于分类任务。该方法依赖于简单的几何直觉，即一个好的度量是指同一类中的点同时彼此靠近并且远离其他类中的点的度量。 同时构造一个凸优化问题，其解决方案通过尝试将同一个类中的所有示例折叠到单个点并在无限远的其他类中推送示例来生成此类度量。学习的度量尺度用于不复杂的分类器时,它会在各种问题上产生相对于标准替代方案的实质性改进。他们还讨论了如何使用学习度量来获得原始输入空间的紧凑低维特征表示，从而允许更有效的分类，同时几乎不降低性能。

深度度量学习具体应用：Ngugen 等[17]描述了一种算法，通过将批次内成对距离的矢量提升到成对距离矩阵，充分利用神经网络训练中的训练批次。该步骤使算法能够通过在提升的问题上优化新的结构化预测目标来学习现有技术特征嵌

入。

度量学习的相关综述：Yang 等[18]调查对距离度量学习中的问题和算法进行了全面的回顾，将距离度量学习中的不同问题分类。在每个类别中，总结现有工作，披露其基本联系，优势和劣势。它主要包括全局监督和测量学习，局部监督和测量学习，以及无监督学习。详细阐述了度量学习的核心思想和核心方法。

对度量学习进行研究的机构：海内外的众多高校和研究小组针对距离度量算法也展开了相应而深入的探究，比如中国科学院[58-59]，澳大利亚国家信息与通信技术研究所[56-57]，美国密歇根州立大学[41,54]，香港中文大学[60]，新加坡南洋理工大学[55]。

基于距离度量的表达谱分类研究：在当前研究中，对肿瘤基因表达谱进行分类利用的更多的是监督分类方法，即首先通过大量的训练数据去训练一个分类器，使该分类器能不断调整内部参数，从而学习到数据集中对分类有帮助的关键特征，然后在输入待分类样本后就能对该样本进行正确分类。目前比较常用的分类器包括 SVM，KNN，贝叶斯，人工神经网络等，都取得了较好的研究成果。而由于基因表达谱数据的高维特征冗余和噪声，数据往往难以处理，从而导致常规分类方法效果不理想。字典学习是一类比较适用于基因表达谱数据的方法，已经有不少研究证明了利用训练好的字典对测试样本稀疏编码并分类的准确率要高于原始的 SRC 方法[102-104]。2011 年，Meng 等使用了一种称为 Fisher 判定的字典学习方法[105]，通过训练字典，使其具有结构化的特征，同时，对编码系数适用 Fisher 准则，然后同时利用字典的重构误差和编码系数之间的距离共同判断一个测试样本的类别，相关几个公共数据库上的实验结果显示，其对人脸的识别准确率要超过与之相关的其他算法。2013 年， Feng 等提出了一种联合判别降维的字典学习方法[106]，通过训练样本训练出的降维矩阵可以将高维度的测试样本投影到更容易区分的低维度空间，进一步提升了识别的准确率。

1.3 研究困难和挑战

LINCS 相似性研究虽然是一项基础的研究工作，但是在 LINCS 数据使用上往往被忽视，而且基于 GSEA 的富集方法成本高、比对效果和时间开销满足不了海量表达谱数据。度量学习的目标是学习距离来判定样本之间的相似度用于分类或者应用，属于比较基础的研究。而将二者结合的研究模型几乎没有，对于 LINCS 数据是否适用于度量的效果还值得探究。如果拓展到实际应用，则可能面临如下问题和挑战：

一、LINCS 相应的细胞系基因表达谱提取问题。目前在数据预处理过程中，算法始终需要依托于 1ktools[20]解析工具，数据提取过程复杂而且解析慢。

二、基因表达谱数据和人脸图像数据有不少共同点，比如高维度，大量冗余噪声等。单样本数据基因维度很高而与度量学习的结合性未知。

三、数据维度高而导致大量的冗余和噪声问题。深度学习用于提取特征理论效果好，但是如何应用到表达谱数据是个问题。

四、相关的衡量标准，如何度量学习的效果好坏。如何判定度量学习效果的好坏是一个基准选择问题。

五、工程实际问题的挑战。在实际应用过程中，还要考虑度量问题在表达谱研究中的应用拓展，如聚类、基因表达谱网络构建等。表达谱的两两比对是快速药物发现的主要手段，同时也是表达谱聚类分析的基础。

除了以上这些困难，LINCS 度量学习问题的研究还面临其他一些挑战，总之，这是个值得深入研究的问题。

1.4 论文的主要内容

1.4.1 论文的主要工作及贡献

针对 LINCS 数据相似性问题的研究，本文提出了两种不同的度量学习模型，同时还提出了判别投影的共享字典学习方法，在 LINCS 数据提取的 A375、MCF7、PC3 三个细胞系的不同样本、不同基因数、不同刺激条件下数据集进行验证，提出的算法相比较 GSEA 有明显提高，而且提出的字典学习分类算法实验取得了较好的分类准确率，本文主要有以下几个方面：

1. 基于改进余弦距离的的基因表达谱距离度量算法。鉴于基因表达谱维数高、数据量大，本文首先在数据集上验证常规的度量学习算法和各种相似度函数对于 LINCS 数据的适用性，验证得出余弦距离是较为适合的相似度计算函数。对余弦距离进行改进，通过中心化和归一化使得算法对于基因表达谱各维度上的值也更加敏感。结合 NCA 算法，组成基于改进的余弦距离的近邻成分分析度量算法。在多个数据集上进行验证，和 GSEA 相比较，该算法度量的矩阵不仅显著提高了分类性能，还在时间消耗上大大减少，是一种对于基因表达谱相似度分析较为适用的度量算法。

2. 基于深度学习的基因表达谱距离度量算法。因为常见的度量学习算法测距用的是欧氏距离或是马氏距离，较难解释数据的非线性关系，在特征提取方面人工干预较多。 本文基于 Siamese 框架，构建 DenseNet 网络和 Cosine 距离相结合的深度学习模型，拓展了隐式的度量学习，采用 Center loss 和 Cross-entropy loss 结合计算损失，提高了模型学习到的高层次特征表达的判定性能。提取 LINCS 数据中典型的三种肿瘤细胞系基因表达谱作为数据集，进行验证，达到 97%以上的分类准确率，并且与 GSEA 和典型的度量学习算法做对比，算法的度量性能在

KNN 上的分类准确率比 GSEA 有 8.4%以上的性能提升，效果也远远好于典型常用的度量学习方法。

3．提出一种基于字典学习的 LINCS 数据分类算法。字典学习分类算法是一种适用于处理基因表达谱数据的算法，而一般字典学习模型仅侧重于提高训练字典重建样本的能力，但忽略了其区分样本的能力。 针对这一情况，本文设计了一种基于判别投影的共享字典学习模型。 在训练字典时，还训练投影矩阵，并且投影矩阵对测试样本的投影可以加宽不同类型样本之间的间隔。另外，通过共享性能获取所有类别的样本，提高分类的判别性。最后，使用字典重建测试样本的错误来确定样本的类别。在皮肤肿瘤、前列腺肿瘤、乳腺肿瘤细胞系数据上的实验结果可以看出，该方法的分类准确率要高于当前主流分类方法。

1.4.2 论文的组织构架

本文主要是对 LINCS 生物表达谱数据之间的度量和分类进行研究，并提出了基于改进余弦函数的度量学习方法、基于深度学习特征提取的度量学习方法和基于共享字典学习的分类算法。

第一章是引言，阐述了本课题的背景和研究价值，指出了本课题的创新之处。接着对相关研究情况展开剖析，对该研究面临的问题和可能存在的困难进行了预估。

第二章介绍了两个内容，一是 LINCS 数据，并对 LINCS 数据的提出做了优化。二是 GSEA 富集分析方法，他们分别是实验的数据来源和度量学习研究的比对标准。

第三章提出了基于改进适应余弦函数的近邻成分分析度量算法。介绍了传统的度量学习模型，提出算法的设计过程、原理和优化方法。

第四章是结合深度学习的度量学习方法介绍。针对目前火热的神经网络，将深度学习用于数据的特征的提取和降维过程。

第五章提出了共享字典的的 LINCS 分类算法。它属于 LINCS 数据的一个最基本的应用。主要是字典学习如何应用于基因表达谱并取得良好效果。

第六章是总结章节，归纳了本文已经完成的工作，对创新点和不足支出展开讨论，并展望了下一步研究的方向及内容。

第二章 LINCS 数据分析和 GSEA 方法

根据已经存在的成果开展对 LINCS 的分析，包括 LINCS 的产生、特点来源及核心组成、数据提取、处理方法，LINCS 是本论文实验部分的数据来源。接着，对差异表达及其功能展开详细说明，引入了基因富集分析这个概念，并对目前性能表现较好且接受度和实用度较好的 GSEA 算法展开介绍，阐述了 GSEA 的原理和结构，GSEA 算法是目前研究 LINCS 数据相似度的主要方法，也是衡量本文提出的度量算法优劣的对比标准。

2.1 LINCS 来源和基本组成

美国国家卫研所（NIH）于 2010 年开启 LINCS 计划[3]。第一阶段（测试实验阶段）从 2010 年进行到 2013 年，重点是开发技术和方法。测试实验阶段数据的产生及分析由哈佛医学院和 Broad 研究所共同承担，并通过 Broad-HMS 合作项目来确保产生数据的一致性，该阶段主要研究高通量实验[4]。目前该计划第一期已经公布 130 万的全基因组表达谱，这些表达谱来源于 77 种典型人体细胞系，它们在 4,000 多种基因沉默试剂和 7,000 多个化学物小分子刺激下检测得出。LINCS 数据分阶段给出，具体的名称及其解释如表 2.1[6]所示。目前的数据是二进制，它的存储的格式是 HDF5。

表 2.1 LINCS 四个阶段数据类型及描述

Level	Type	Description
1	LXB	Raw, unprocessed flow cytometry data from Luminex scanners. One LXB file is generated for each well of a 384-well plate, and each file contains a fluorescence intensity value for every observed analyte in the well
2	GEX	Gene expression values per 1,000 genes after de-convolution from Luminex beads
3	Q2NORM	Gene expression profiles of both directly measured landmark transcripts plus imputed genes.Normalized using invariant set scaling followed by quantile normalization
4	Z-SCORES	Signatures with differentially expressed genes computed by robust z-scores for each profile relative to population control

LINCS 数据具有四层结构：第一层是经过扫描的原始数据；第二层的近 1000 个数据（具体是 978 个基因数据）是由第一层数据反卷积操作后所获取；第三层数据是经过标准化测量数据加外推和估计所得的表达谱数据；第四层为 Z 归一化后所得基因差异表达的印记数据。在 LINCS 里面具体包括的数据类型主要有以下几种：一是扰动和配体扰动，二是各种试验阶段或者已经进入临床批准的化合物小分子，三是基因的沉默、敲除、过表达，四是各种类型的空白对照实验如 untrt 等[4]，如表 2.2[6]所示。

表 2.2 LINCS 数据的构成类型

Class	perturbation	Pert_type
Treatment	Chemical compound	Trt_cp
Treatment	Gene Knockdown	Trt_sh
Treatment	Gene over expression	Trt_oe
Treatment	Mutant gene over expression	Trt_oe.mut
Treatment	Ligand treatment	Trt_lig
Control	Untreated	Ctl_untrt
Control	Control vector	Ctl_vector
Control	Vehicle control	Ctl_vehicle

另外，LINCS 的数据规模较大，每个阶段的数据量都超过了 110G，目前还在逐步增加中，如果直接使用计算机读取的话，会有内存错误的情况发生。

针对这个问题，Iktools[20]被当做数据的读取工具由研究团队专门开发出来，开放于 GitHub[21] 上，它能够支持多种编程语言如 JAVA，R 等，并且一些常用软件的数据包包含其中，能够以此根据具体的 meta 数据获得所需要的信息。

Distill_id 是其中的核心属性，能够用来提取各种各样的信息比如干扰小分子种类、具体的细胞类型、干扰的时长等[4]; 同时，lincscloud 作为一个在线的云平台，是该项目面向线上操作和信息交流的工具以及通道，这个平台提供了便于操作和调用的 API——L1000[22][24][25] 以及各类 APP 应用[23]。LINCS 数据的官方网站可参考图 2.1。

另外，在 NIH LINCS 项目官网还提供了 50 多种分析数据和使用数据的视频教程和工具，而且操作过程简单，基本能够实现操作步骤和数据结果展示的可视化。这些配套的使用工具包含药物通路、各种子功能浏览器、移动端接口等。

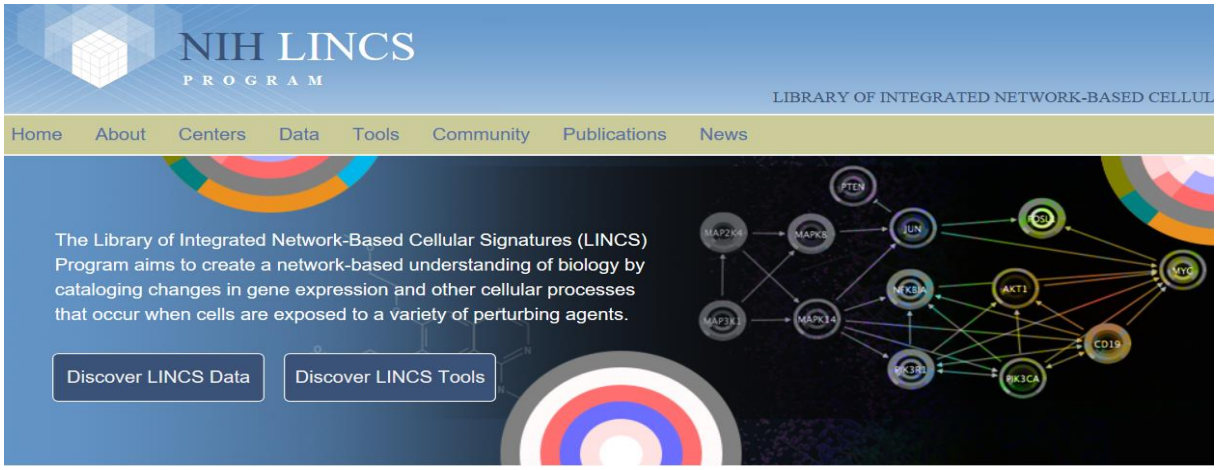

图 2.1 LINCS 项目官网

数据产生和处理流程：L1000 技术是数据获取的核心技术，具体操作是通过 L1000 实验平台来检测不同类别的癌细胞在不同的外界刺激和实验条件下的表达状况。目前 LINCS 项目进行到第五个阶段并且公布了对应的数据，这个大数据计划一共有七个阶段。

LINCS 数据包含 978 个界标（或者称为基础）基因，这些基因表达来自于 L1000[24][25]平台获得的试验图像的峰值分解和反卷积，也就是 Level2 的数据；与此同时，因为具备 80 个处于各类模式以及外部环境中维持转录表征水平的基因存在，所以这些基因可以当成试验的数据标准化的对比值。这些数据的对比操作是需要数据处于同一个范围内，因此需要进行各类数据的放缩操作。因为具体数据分布情况有区别，而所有的数据需具备统一的分布情况，因此需进行数据标准化的操作，此操作是基于分位数来完成的，详细的步骤参考图 2.2；基于在这个阶段获得的数据，利用表达谱之间的相互关联性可以推导出其他 20000 多个人类基因的表达程度，由此就得到了 Level3 的数据；基因的印记数据 Level4 是通过空白实验组以及具体比较组数据比对得到的差异性而获得[4]，具体流程如图 2.2[3] 所示。

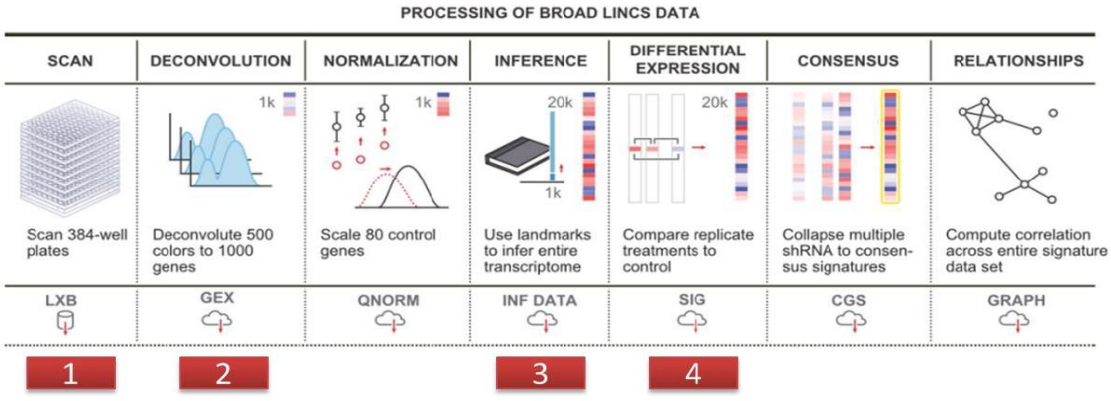

图 2.2 LINCS 数据处理流程

2.2 基于 H5py 的数据提取方法

传统的 Itools 方法需要对数据进行解析，只能逐条读取数据，如果能够自行实现原始数据的解析工作并将之整合到完整的算法流程中，无疑会使本课题工作更具实用价值。HDF（Hierarchical Data Format）是一种文件格式，这种格式文件带有特定的某些数据库，它是针对大规模数据存储和操作而设计出来的。HDF 是被 NCSA（美国国家超算计算研究中心）研发而来的，维护任务由 HDF 小组承担，该小组是非盈利性质的。HDF5 拥有许多的优点，比如它适用很多类型的数据，尤其适用于成规模数据处理。同时具有跨平台性、具有拓展性等高效的 I/O 性能，理论上支持无限量文件存储。

LINCS 的数据集包括大规模的表达谱数据，这些数据是多种试验条件下检测得出，属于 HDF5 格式的文件，文件以.gctx 亦或.gct 结尾。目前 LINCS 官方网站给出的数据处理工具是 Itools，需要用 Itools 对原始矩阵信息进行解析和提取，但在数据预处理过程中，算法始终需要依托于 Iktools 解析工具，一旦解析过程结束便无法返回源表达谱数据进行处理，这使得查询工具在使用的便利程度上略显不足，原始数据格式如图 2.3。

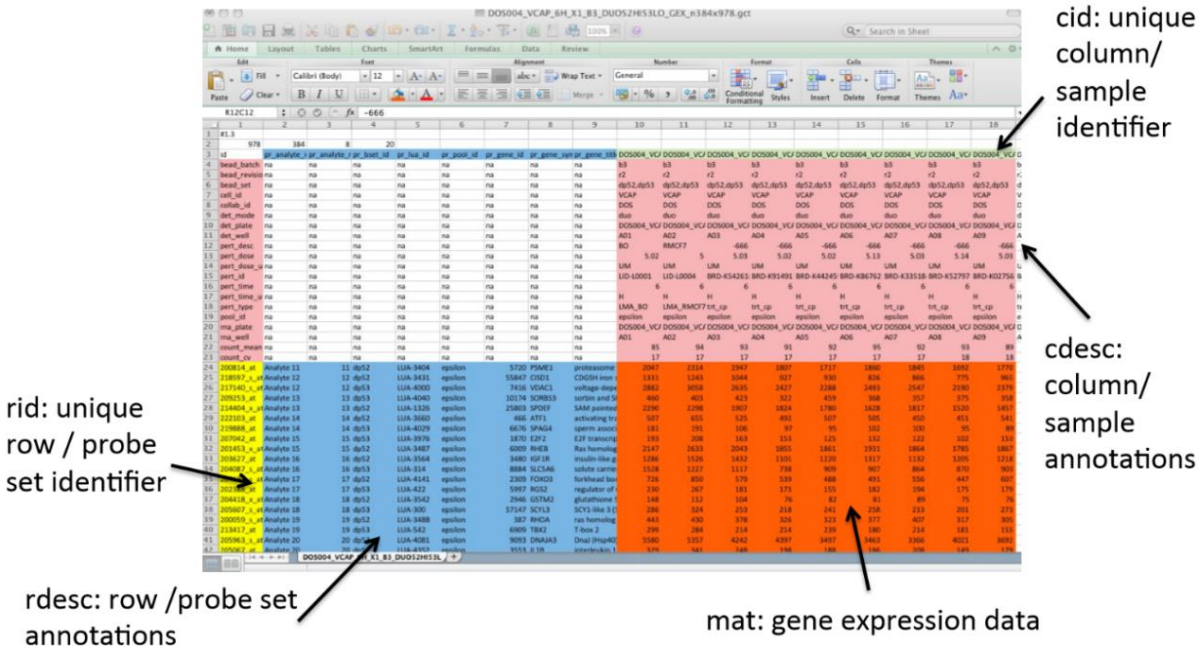

图 2.3 原始基因表达谱数据格式

在图 2.3 中，位于左上角的白色区域没有标识符，其余列表示具有列唯一标识符 cid 的表达谱；每行代表一个具有行唯一标识符 rid 的基因。左下方的蓝色区域是线描述信息 rdesc，它表示基因的一些相关特性，但是在本文的方法中不需要这些基因属性，因此没有必要注意；右上方的粉红色区域是列描述信息 cdesc，表明获取相关控制信息和样本表达谱的处理条件是本文需要注意的主要属性，因为

分析提取和描述表达谱由这些属性完成。通过这些信息，我们可以在分析结果中了解特定生物学特性的成因和抑制方法，从而达到寻找药物和治疗疾病的目的。

通常，对于药物发现，主要的五个属性支持表达谱的分析提取和最终结果的显示分析以帮助药物发现。这五个属性主要包括：细胞系，干扰素，实验组类型，治疗时间和治疗浓度。不难发现，使用这些属性，基本上可以确定出特定病状的特征，对于药物发现和重定位有很大的参考价值。由于本文的重点是研究特定病例中表达水平之间的相似性，细胞系，实验类型和干扰素是提取数据的主要依据。右下方的橙色区域是特定基因的表达水平，也是使用的主要计算数据。

H5py 是由 python 开展而来，依托于 python 功能强大的工具库和和各种包，在用 python 进行数据分析的时候能方便调用各种工具节省数据处理时间。具体的数据提取过程如下：

1. 安装 h5py，并且加载 numpy、pandas 工具应用库。

2. 导入 GCTX 文件，使用 map 映射文件，操作文件格式转换成 DataFrame。

3. 根据附件 gene_info 和 cell_info 以及 inst_info 按照要求筛查出所需要的细胞系、实验类型和扰动条件等信息。由于 inst 信息是唯一的，所以 inst 是我们选择信息的标准。

4. 根据 inst_id 返回主矩阵搜寻所需数据的样本。该样本的表达谱值就是所需数据。

5. 编写读取函数，运行 pandas 数据库的 DataFrame 语句，操作提取所需矩阵数据并保存为样本，储存为所需格式。详细的处理流程可查看图 2.4。

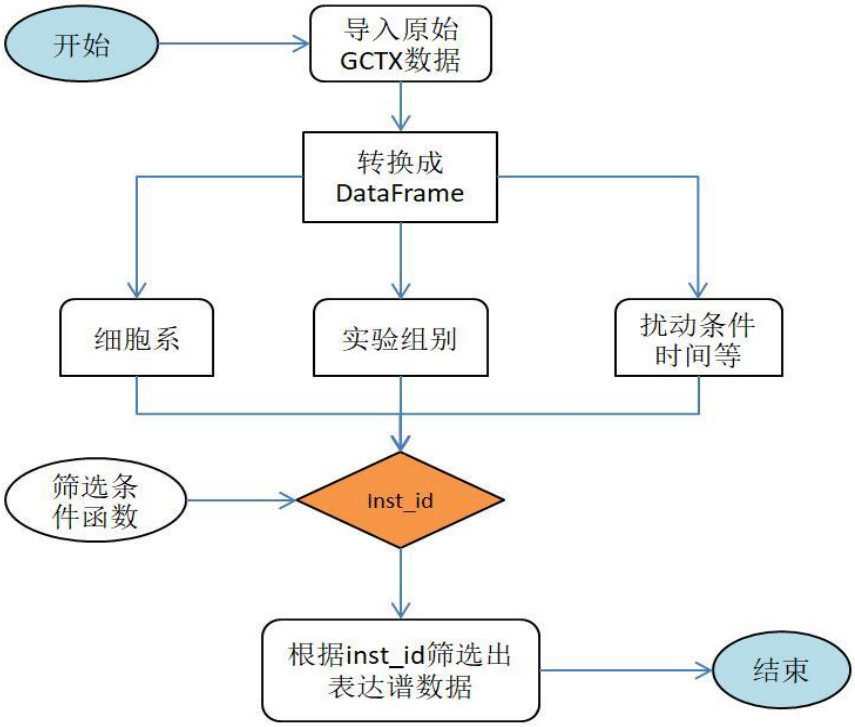

图 2.4 H5py 数据提取过程

2.3 GSEA 相似度算法分析

GSEA (Gene Collection Enrichment-Analysis，基因富集分析）在表达谱分析运算上属于很经典和普遍的方法，应用范围和场景也较为广泛。在国内外具有较大影响，其原始论文[5]引用数超过 9,128 次（Google Scholar），相关实现软件下载次数过十万次。

GSEA 的基本思想是比对和表征两个样本之间形态上的差异性，具体操作是检测预先定义的基因集 S 中的基因在 L 排序集合中的分布情况，查看其汇集在底端或者顶端。

GSEA 分为 3 个主要步骤：

1）富集积分 ES 的计算：富集积分运用 Kolmogorov-Smirnov 开展运算，其依据在于集中出现在指定集合的位置比如底部和顶部的概率，通俗的讲就是检测 L 序列的尾部和头部有多少 S 基因的表达基因量。计算过程见图 2.5[5]。

2）预估 ES 的显著性程度（Significance Level）：ES 的统计学显著性（标称 P 值）的估计是通过一个常规表型方法——置换检验，具体就是置换表型标签，以及为每个置换的 ES 重新运算结果。ES 的某个 0 分布都是成千个置换产生而来，所得到的检测数值就是经验性标称 P 值。因为本过程采用的是对样本的置换，因而保留了基因间的互相作用，与基因的替代相比，它能更真实地反映其生物学过程。

3) 调整多重的假定测验：评测整个基因库，通过调整标称 P 值来解释多次试验。包括创建归一化的富集积分 NES，通过对每个基因集进行归一化，以阐明使用 FDR（False Discovery Rate）对每个 NES 操控前的数据组大小，并产生基因列表的 FDR。GSEA 算法计算步骤参考图 2.5[5]：

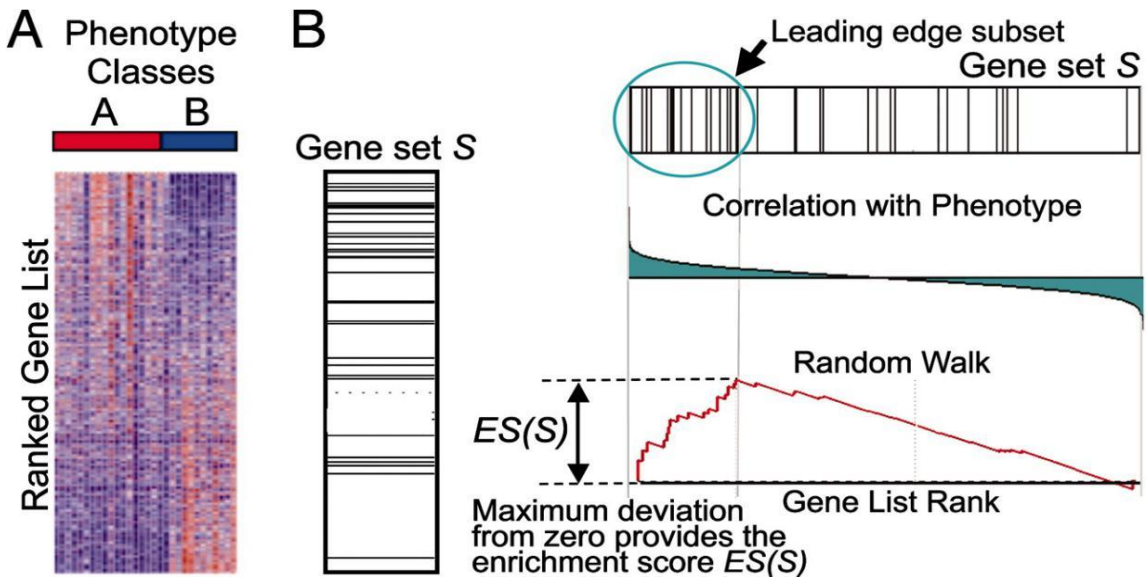

图 2.5 GSEA 算法过程

富集积分运算流程可以归纳为：

依据样本数据的表达相关性，对 N 个基因的表达谱进行分类，得到有序的基因谱$\mathrm{L}=\{g_1,g_2,\ldots,g_n\}$，让排序第 j 位的基因表达水平作为第 j 个基因g_j的表达值，推导得出式$\mathrm{r}(g_j)=r_j$成立。

得出基因探针（基因集）S 位于 L 上的 miss 值与 hit 值，公式（1.1）[5]（1.2）[5]展现了这两个值：

$$P_{hit}(S,i)=\sum_{g_j\in S}\frac{|r_j|^p}{N_R}\quad \text{其中}\ \ N_R=\sum_{g_j\in S}|r_j|^p \tag{2.1}$$

$$P_{\mathrm{miss}}(S,i)=\sum_{g_j\in S}\frac{1}{(N-N_H)} \tag{2.2}$$

S 是被随机挑选出来的，这个时候 ES(S)值不会很大，ES(S)的值很大的情况出现在聚集点位于 L 的底部或者顶端。在指数 p=0 的情况下，ES(S)转变成为 Kolmogorov-Smirnov 统计的分布（这属于一种非参数验测，用于查看两个样本是不是具有相似性）。

在实际工作中，根据式（2.1）和式（2.2）进行预排序和计算P_{hit}和P_{miss}，一般为了搜索到偏离 ES 比较大的值，首先需要对击中的 P 和未击中的 P 求和他们的前缀，然后按照前缀的迭代次数做差值，得到最大的差值，所需要的积分 ES 会通过扫描整个 L 得到。

在运算过程中的时间消耗复杂度可以轻松的从运算过程中获得。n 表示 L 序列长度，m 代表 S 基因集长度，下面的几个步骤可以完成运行的算法（其时间复杂度也分析如下）。

（1）排序原始的数据，按照表达量的规模进行；

（2）计算P_{hit}和P_{miss}：不管基因是否命中，探针总是会被扫描，故而时间开销为$O(mn)$；

（3）求解 ES：（2）中的向量前缀的和是需要遍历和计算的，因此这个步骤的复杂度$O(n)$。

总之，GSEA 作为实用的手段被应用于基因表达谱，但是面临的一个难题在于，GSEA-P-R.1.0[27]、GSEA2[29]、SAM-GS[28]等工具都存在的局限性，那就是脚本语言处理很慢，故而很难达到所需要的处理效率。另外一方面由于这些工具的实现方式比较直接，没有优化的操作，实际使用过程中也存在诸多问题，特别在处理大规模数据的进程上，所以 GSEA 迅速分析[30] 成为了一个可以探究的具有价值的问题。Peng 等[31]提出了基于超级计算的并行加速算法 ParaGSEA，计算效率上提升了近 50 倍的速度，但是硬件条件需求较高，普遍适用性能不高。

本文的对比实验所用的 GSEA 算法针对第二步和第三步，两个向量进行遍历改为双向同时遍历，简化了部分代码。具体的 GSEA 算法优化改进问题，就属于

另外一个研究领域，我们在这里不做深入讨论。

2.4 本章小结

在本章中，具体讲述了 LINCS 计划创立的背景，其次是 LINCS 数据的生成过程、处理过程和数据量。然后对 LINCS 数据的数据构成和结构进行了详细分析，阐述了选择相关特征的依据。基于官方给出的 Itools 工具缺陷 ，本文从 Python 出发，提出来基于 H5py 的数据提出过程，操作简单直接，而且提取的数据能保存成所需要的格式。本章第二部分对 GSEA 方法进行了详细说明，解释了将其作为后续算法比对标准的原因，针对 GSEA 算法的改进问题和前景也进行了探讨。

第三章 优化度量函数的近邻成分分析算法

3.1 引言

当前，度量函数是模式识别、分类算法等应用中的基础但是重要的部分，很多应用目前依赖于手动给定的度量，度量函数的选取针对具体的应用场景而言具有重要的意义，属于当下比较前沿的研究点。度量学习是这样一种方法，它依据给定的样本信息来获取度量函数，最大化度量空间以适应问题域，通过以往从理论和实验获得的度量函数获得样本之间的距离，确保准确率在聚类、分类、和检索上有显著提升。一般的基因表达谱研究普遍模式是：着重于先降维然后采用简单的距离函数计算相似度，这样就存在一个问题，在降维过程中会引入人工的判定标准，从而影响结构完整性和数据的内在关联性。而 LINCS 数据目前的聚类、分类算法计算相似度普遍采用简单的欧式距离或者富集分析，应用度量学习表达谱的算法比较少，虽然这是一项基础研究工作，但是确是一值得深入研究而且意义深刻的课题。

3.2 典型度量学习模型分析

Mahalanobis 距离是线性 DML 使用最多的度量函数，它的核心思想是寻找到最优的 Mahalanobis 距离让目标函数能够达到最优。Mahalanobis 距离所得到的距离矩阵可以看做半正定的，故而获取最佳的变换矩阵然后对数据展开线性转换是线性度量算法的根本所在。常用的距离度量函数如表 3.1 所示：

表 3.1 常用距离度量相似度

度量名称	适用场景	表达内涵	取值范围	表达意义
闵可夫斯基距离	连续性数据	度量空间的 Lp 范式	$[0,+\infty]$	差异度
马氏距离	多变量正态分布	无量纲的欧式距离	$[0,+\infty]$	差异度
余弦相似度	连续性及稀疏数据	向量夹角的余弦值	[-1,1]	相似度
皮尔森相似系数	多变量正态分布	两数据线性相关性	[-1,1]	相似度
杰卡德相似系数	二值数据	集合的覆盖比例	[0,1]	相似度
S-W 相似度	序列数据	序列的局部相似度	[0,1]	相似度

不同的相似度计算模型具有适用的使用环境，大多数生物信息学文献中，DNA 序列数据和蛋白质的相似度计算方式为 Smith-Waterman 相似度，而药物分

子相似度的计算模型为杰卡德相似系数。常用的度量模型参考表 3.2：

表 3.2 常用的度量模型

名称	监督形式	度量形式	拓展性	优化方式	降维	特点
LMNN	全监督	线性	特征长度敏感	全局	否	为 KNN 设计
ITML	弱监督	线性	样本量敏感	全局	否	可在线化
SDML	弱监督	线性	复杂度小	全局	否	n 远小于 d
LSML	弱监督	线性	特征长度敏感	全局	否	超参数少
LFDA	全监督	线性	复杂度小	全局	是	多类数据
RCA	弱监督	线性	复杂度偏小	-	否	最小化类内距离

3.2.1 LMNN

Weinberger KQ 等[61]针对监督 DML 问题，提出大间隔最近邻居（LMNN，large margin nearestneighbor）算法。数据集$\{(x_i, y_i)\}_{i=1}^{n}$用于训练，特征表示为 x，y 代表标记。这种算法用$y_{ij} \in \{0,1\}$判断样本 i 和 j 能否归为一个类别。上文中已经提到 Mahalanobis 距离和线性转换可以看做等价性，故而该方法是将 DML 问题转变成为线性的变换问题。即 $D(x_i, x_j) = ||L(x_i - x_j)||^2$。LMNN 所使用的优化目标公式（3.1）[61]所示：

$$\min_{A \geqslant 0} \sum_{(i,j) \in S} d_A(x_i, x_j) + \lambda \sum_{(i,j,k) \in R} 1 + d_A(x_i, x_j) - d_A(x_i, x_k)_+ \tag{3.1}$$

式子的第一部分代表单独的数据点和其同类相邻的数据点的间隔之和。第二部分的核心目标是用来惩罚每个数据点的距离间隔小于同一点的不同类点。LMNN 算法的思想比较简单：损失函数的目标是任何数据点应该与其邻域内的点共享相同的标签，而异构样本点应该相距更远并且与相似的样本点相比应该是间隔开的。故而，该目标函数更倾向于使相似的样本点更近，不同类样本点间隔较远，而且它们之间有距离函数。集合 S 表示的是一对由所有目标相邻点组成的点，一般是x_i，x_j对组成，其中 k 与点对x_i、x_j相邻。三元组（x_i，x_j，x_k）的集合可以用 R 来表示，x_i和x_j是目标的近邻点对，而x_k是异类点。参考公式（3.1），类同于 SVM 的优化目标，LMNN 能最大化异类样本对之间的距离，这种方法可以说是当下对线性 DML 处理比较优良的方法之一。

$$\begin{gathered} \min \textstyle\sum_{ij} \eta_{ij} (x_i - x_j)^t M (x_i - x_j) + c \sum_{ij} \eta_{ij} (1 - y_{il}) \xi_{ijl} \\ \text{s.t. } (x_i - x_l)^t M (x_i - x_l) - (x_i - x_j)^t M (x_i - x_j) \geq 1 - \xi_{ijl} \\ \xi_{ijl} \geq 0, M \geqslant 0 \end{gathered} \tag{3.2}$$

3.2.2 LFDA

作为数据挖掘中应用比较普遍的降维方法，Fisher 判别分析方法（FDA，Fisher Discriminant Analysis）[62]被运用的比较多。它的本质在于投影的选择上，需要选择出最佳的投影,最大化类与类之间的间,但是这种方法会造成局部特征的缺失。为了克服这个缺点，局部 Fisher 判别分析方法 LFDA（Local Fisher Discriminant Analysis）[63-64]被提出，这种算法的核心点是结合了 LPP[65]和 FDA 的优势，使用这种方法能让类和类在投影空间中有更好的区分度。假定某个训练数据集 $X=\{x_{1,}x_{2,}x_{3}...,x_{n}\}$且$x_i \in R^d\{(x_i,y_i)|1\le i\le m,m\le n\}$经过了类标注，则：

$$S^{(lb)}=\frac{1}{2}\sum_{i,j=1}^{m}W_{i,j}^{(lb)}\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}} \tag{3.3}$$

$$S^{(lw)}=\frac{1}{2}\sum_{i,j=1}^{m}W_{i,j}^{(lw)}\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}} \tag{3.4}$$

其中$W_{i,j}^{(lb)}=\begin{cases}A_{i,j}\left(1/m-1/m_y\right),y_i=y_j\\1/m,y_i\neq y_j\end{cases}$，$W_{i,j}^{(lw)}=\begin{cases}A_{i,j}/m_{y_i}y_i=y_j\\0,y_i\neq y_j\end{cases}$，而这当中的$A_{i,j}$具体可以表示为$A_{i,j}=\exp\left(-\frac{\|x_i-x_j\|^2}{\sigma_i\sigma_j}\right)$，$\sigma_i$表示$x_i$的局部衡量。LFDA 的最终投影目标函数[63]可表示为：

$$\max_{T}\frac{T^TS^{(l)}T}{T^TS^{(lW)}T} \tag{3.5}$$

3.2.3 ITML

信息论度量算法（ITML，Information-theoretic metric learning）是由 Davis 等[66]提出来的。对度量矩阵M而言，具有M^{-1}这样的协方差，它包括的多元高斯散布是$p(x;M)=(1/z)exp((-1/2)d_M(x,\mu))$，其中$\mu$代表平均值是归一化因子。两个高斯散布的评价通过相对熵来判定它们（先验矩阵M_0和需要学习到的矩阵M）的距离，即：

$$KL\left(p\left(x;M_0\right)\|p(x;M)\right) \tag{3.6}$$

其中，$KL(.)$代表相对熵，也被称作 Kullbacle-Leibler 散度，它的作用是衡量概率分布的差异量。

处于非等价约束集合 D 和等值约束集合 S 的作用下，距离度量目标最后会转变为求最优解：

$$\min_{M\succeq 0}KL\left(p\left(x;M_0\right)\|p(x;M)\right)$$

$s.t.$

$$\begin{aligned}&d_M\left(x_i,x_j\right)\le u,\quad\left(x_i,x_j\right)\in S\\&d_M\left(x_i,x_j\right)\ge l,\quad\left(x_i,x_j\right)\in D\end{aligned} \tag{3.7}$$

为了能够让距离矩阵满足在成对约束下条件下的阈值，上面的目标函数所展示的约束条件做了改进。另外，为了预防过拟合的发生，就得让M和M_0距离最小，这样就需要KL散度尽可能的小。

假定M和M的高斯分布相同，就可以使相对熵凸函数$\phi(X)=\log det(X)$所得到的布雷格曼散度展开运算：

$$\begin{aligned} &KL\left(p\left(x;M_0\right)\|p(x;M)\right)=\tfrac{1}{2}D_{ld}\left(M_0^{-1},M^{-1}\right)\\ &D_{ld}\left(M,M_0\right)=\operatorname{tr}\left(MM_0^{-1}\right)-\log\det\left(MM_0^{-1}\right)-d \end{aligned} \tag{3.8}$$

其中，$D_{ld}(.)$代表布雷格曼散度$tr(.)$代表矩阵A的迹，$logdet(.)$代表A矩阵的行列式上的对数。所以，式（3.7）[66]能够转变为:

$$\min_{M\succeq 0} D_{ld}\left(M,M_0\right)\ \ s.t.$$

$$\begin{aligned} &\operatorname{tr}\left(M\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}}\right)\le u,\left(x_i,x_j\right)\in S\\ &\operatorname{tr}\left(M\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}}\right)\ge l,\left(x_i,x_j\right)\in D \end{aligned} \tag{3.9}$$

ITML 通过引进松弛变量ξ并使ξ_0初始化（μ作为等值约束样本对，是样本取值方式，l表示的是非等值约束数据对）用于比较宽广的可行域内展开求解,式（3.9）经过改写为:

$$\min_{M\succeq 0,\xi}\left(D_{ld}\left(M,M_0\right)+\gamma D_{ld}\left(\operatorname{diag}\{\xi\},\operatorname{diag}\left\{\xi_0\right\}\right)\right)$$

$$s.t.$$

$$\begin{aligned} &\operatorname{tr}\left(M\left(x_i-x_j\right)\left(x_i-x_j\right)^{T}\right)\le \xi_{i,j},\left(x_i,x_j\right)\in S\\ &\operatorname{tr}\left(M\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}}\right)\ge \xi_{i,j},\left(x_i,x_j\right)\in D \end{aligned} \tag{3.10}$$

其中，γ是均衡参数。式（3.10）可以通过 Bregman 进行求解。Bregman 投影是通过迭代的方式计算，为了获得下个预估，需要在成对约束条件在本次解的基础上运算，即：

$$M_{t+1}=M_t+\beta M_t\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}}M_t \tag{3.11}$$

其中，M_t表示第t次迭代运算获得的距离矩阵，β表示映射参数x_i和x_j表约束对。

3.2.4 KISS

Kostinger 等[67]提出的保持简单直接的度量学习算法 KISS（Keep it simple andstraight）是在高斯分布假设的前提下开展的，这种方法使用似然比测验使得距离度量求解变成下式：

$$\delta\left(x_{ij}\right)=\log\left(\frac{p\left(x_{ij}|H_0\right)}{p\left(x_{ij}|H_1\right)}\right) \tag{3.12}$$

如公式，$x_{ij}=x_i-x_j$，H_0和H_1分别表示样本对$(x_i,x_j)\in D$和$(x_i,x_j)\in S$的假设。如果$(x_i,x_j)\in D$，则$\sigma(x_{ij})$的值会比较大，否则会与之相反。所以，x_i和x_j的间隔一

般用$\sigma(x_{ij})$来表示。高斯散布值等于零，以此对（3.12）的密度函数建模，会得到简化的式子:

$$\delta(x_{ij}) = (x_i - x_j)^{\mathrm{T}} \widehat{M} (x_i - x_j) \tag{3.13}$$

其中，$\widehat{M} = \sum_{(\boldsymbol{x}_t, \boldsymbol{x}_j) \in S}^{-1} - \sum_{(\boldsymbol{x}_t, \boldsymbol{x}_j) \in D}^{-1} \cdot \sum_{(\boldsymbol{x}_t, \boldsymbol{x}_j) \in S} = \sum_{(x_x, x_j) \in S} (x_i - x_j)(x_i - x_j)^{\mathrm{T}}$，表示的是等值样本对的外积之和。

其他常用的度量学习模型还有最小二阶乘度量学习 LSML（Least Squares Metric Learning）[68]和主成分分析 RCA（Relative Components Analysis）[69]，稀疏行列式度量算法 SDML（Sparse Determinant Metric Learning）[70]等。

3.3 PC-NCA 距离度量算法

基于改进余弦距离的近邻成分分析算法（PC-NCA）就是将改进的余弦距离度量（Promoted Cosine Distance）用于 NCA 算法中，替换掉算法本身的马氏测距度量，用以提升针对基因表达谱数据的距离度量效果的算法。

3.3.1 近邻成分分析算法（NCA）

机器学习中距离测度可以说是比较重要的一个步骤，其目标就是为了通过学习获取到距离测度$d(x_i, x_j)$来表示x_i和x_j二者的相似性[71]，它的核心是为了获得数据的更佳的空间特性，而采用的方式便是转变线性或者非线性以此获得类别区分度更好的表征。NCA 属于有效率的距离测量方法中的一种。这种方法经过对左一法（Leave-One-Out，LOO）交叉测试结果的优化，可以随机选择最近的点，得到马尔可夫距离中的变换矩阵。假设在R^D空间内传入n个样本对$x_1, x_2, ..., x_n$，它们的类标签表示为$c_1, c_2, ..., c_n$。限定马氏距离变换矩阵Q代表的是 Mahalanobis 距离之下的转换矩阵，它是半正定的，那么$Q = A^T A$，这时样本点的距离$d(x_i, x_j)$[72]就是：

$$\sqrt{(x_i - x_j)^{\mathrm{T}} Q (x_i - x_j)} = \sqrt{(Ax_i - Ax_j)^{\mathrm{T}} (Ax_i - Ax_j)} \tag{3.14}$$

其中，$i, j = 1, 2, ..., n$，留一法在运算误差过程中，由于误差函数 A 不是连续的，因此需要引进可微的函数 softmax：

$$p_{ij} = \frac{\exp\left(-\|Ax_i - Ax_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|Ax_i - Ax_k\|^2\right)} \tag{3.15}$$

p_{ij}表示的是概率，代表样本点x_i随机挑选某个近邻x_j而继承类标签x_j的概率[72]。这样，将样本点x_i正确分类的概率为$p_i = \sum_{j \in C_i} p_{ij}$，其中，$C_i = \{j | c_i = c_j\}$。

目标函数要使得正确分类点的数目最大，因此，定义为$f(A) = \sum_i \sum_{j \in C_i} p_{ij} = \sum_i p_i$。

这是一个连续可微的矩阵函数，算法就是要最大化该目标函数。这是一个无

约束优化问题，可以通过共轭梯度法或者随机梯度法求出 A。

利用这种算法降维和测距，过程比较简单，没有矩阵计算的复杂，也无需对样本空间分布进行假设。

3.3.2 改进的余弦度量距离

余弦相似性通过测量两个矢量之间的角度的余弦值来判定。值的表示范围是-1 到 1 之间，故而，余弦相似度确定的是两组向量的方向指向是否相同。余弦距离和每个方向上的矢量值大小没有关联，仅仅代表与矢量的方向一致性。余弦距离通一般都用于正空间求解，所以常见的值一般在 0 和 1 之间变动。

考虑到 LINCS 基因表达谱数据的高维度的数字特征，计算维度之间的夹角相似度是一个可行的方案。向量之间的余弦值能够使用欧几里得点积公式求得：$\mathbf{a}\cdot\mathbf{b}=\|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$，A 和 B 代表的是两个向量，它们的余弦相似性 θ 是经过向量长度和点积和$\frac{A\cdot B}{\|A\|\|B\|}$得到，如公式（3.16）所示：

$$similarity=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{3.16}$$

这里A_i,B_i分别代表向量 A 和 B 的各分量。表征的相似性范围从-1 到 1 变化。-1 代表 A 和 B 指向完全相反的方向，1 表示它们的方向完全相同，0 通常意味着 A 和 B 是独立的。改进的余弦相似度如公式（3.17）所示：

$$\mathrm{Sim}(A,B)=\frac{\sum_i A_i^*B_i^*}{\sqrt{\sum_i A_i^{*2}}\sqrt{\sum_i B_i^{*2}}} \tag{3.17}$$

其中，$A^*=\frac{A-mean(A)}{Max(A)-Min(A)}$,$B^*=\frac{B-mean(B)}{Max(B)-Min(B)}$，$min$代表所有样本中最小值，$max$代表样本的最大值，$mean$代表的是所有样本的均值。

改进余弦相似度（PC_NCA）的目的是，针对余弦相似度只参考向量维数方向的差异性但是却不参考各维的量纲的问题而开展的优化，因此在计算相似度时，对每个维的均值进行了改进运算。

3.4 实验评估

3.4.1 实验平台和数据集

本章实验在基于 win10 专业版系统的的电脑进行，硬件配置如下：CPU 为英特尔 i7 的型号是 6700HQ 的处理器，,显卡的配置是英伟达的 GTX970 显卡，电脑内存 32GB，固态硬盘 256GB。软件配置为 win10 专业版+Pycharm+tensorflow（GPU）版本。

本章实验的数据集 LINCS Phase 3 (GSE92742) 提取的三种细胞系：前列腺肿瘤（PC3），乳腺肿瘤（MCF7），皮肤肿瘤（A375），在 untreated 和 vehicle control

两种实验条件下的 2000 基因表达谱样本，表达谱数据的维度为 978 维和 12328 维，根据实验类型选择相应的数据维度。详细的实验数值参考表 3.3，对于每一种维度数据，都进行三项比较：在 untreated 实验组中 MCF7 细胞系和 PC3 细胞系、MCF7 细胞系和 A375 细胞系，以及 MCF7 细胞系在 untreated 实验组和 vehicle 实验组的对比。数据的训练比例为 0.5。

表 3.3 实验数据比对说明

维度		978	
实验类型	untreated	untreated	Untreated/Vector
细胞系	MCF7**&**PC3	MCF7**&**A375	MCF7
维度		12328	
实验类型	untreated	untreated	Untreated/Vehicle
细胞系	MCF7**&**PC3	MCF7**&**A375	MCF7

实验评价标准：由于度量学习获得的是一个距离矩阵，因此不能直接比较距距离矩阵的大小。故而，第一个衡量标准是基于距离的分类准确率。先利用各种度量算法获得所需的距离矩阵，接着运行 KNN 算法衡量度量效果，度量效果依据分类的准确度高低来判定，在不考虑分类算法本身影响的情况下，准确度越高就表明度量效果越好。另外一个衡量标准则是 ROC 曲线，通过曲线的具体表现来反馈模型的性能表现。

3.4.2 实验一：度量算法和 GSEA 比对

本实验通过选择几种常规的度量模型 LMNN，LFDA，LSML，SDML 和 GSEA 算法做比对，验证常规度量模型和 GSEA 做比对的度量效果。本实验使用 MCF7 细胞系和 A375 细胞系基因数据。

通过图 3.1 可以看出，常规的度量学习方法在度量性能上面和 GSEA 还是有些差距，但是 NCA 算法在分类准确率上比 GSEA 算法高 4 个百分点左右，NCA 度量学习算法比较适合 LINCS 基因表达谱数据。时间消耗如表 3.4 所示。可以看出，耗时上面 GSEA 是其他算法的数倍，LFDA 在时间消耗上面最少，其次是 NCA 算法，度量学习算法有较大改进空间。

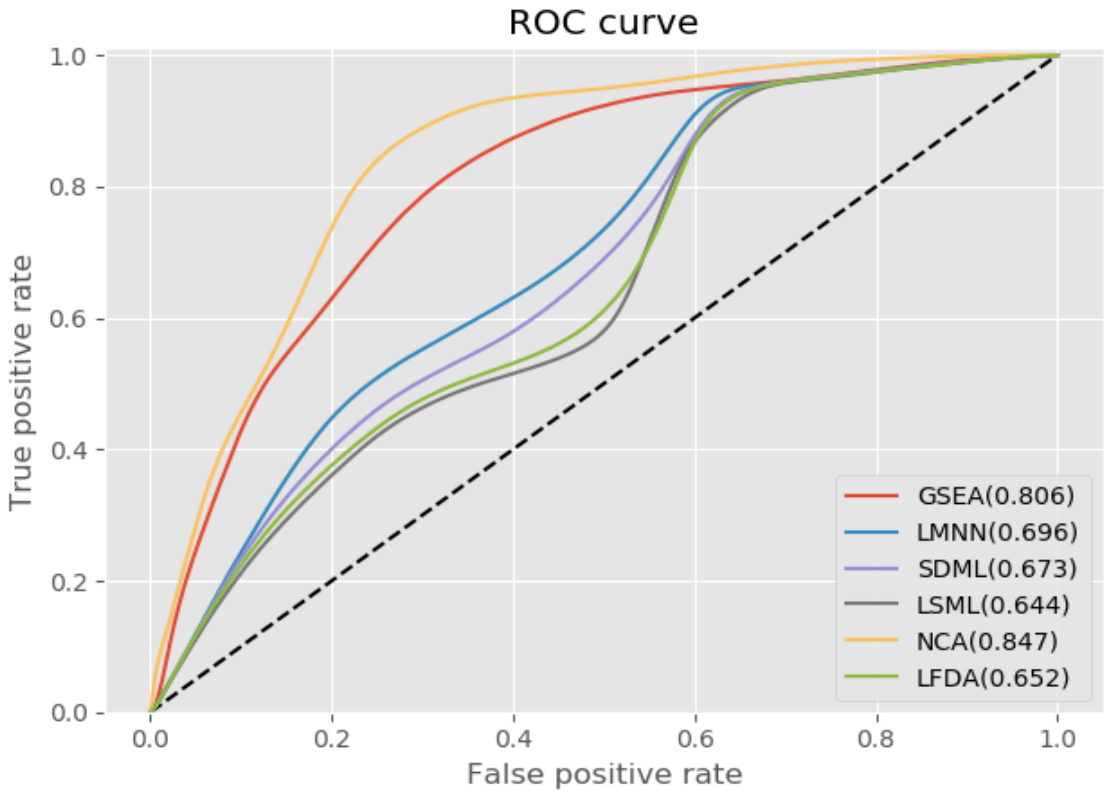

图 3.1 常用度量学习模型和 GSEA 算法分类准确率

表 3.4 各种算法运行耗时（s）

算法	GSEA	LMNN	SDML	LSML	NCA	LFDA
耗时	533.6028	48.3662	21.6228	43.6748	21.2348	7.525

3.4.3 实验二：距离度量函数性能比对

本实验通过选取不同的相似度函数在 MCF7 细胞系基因表达谱下，直接度量矩阵来得出距离矩阵，通过 KNN 分类准确度来衡量不同的相似度函数对于基因表达谱数据的相似度。距离度量函数在基因表达谱数据上直接使用的效果如图 3.2 所示。

从图中实验不难看出，Cosine 相似度相比较其他的度量距离（包括欧式距离、第二范式距离、曼哈顿距离、切比雪夫距离、汉明距离、杰卡德距离、闵可夫斯基距离），在基因表达谱度量上具有较大优势，分类准确率达到 77.2%,性能超过排名第二的第二范式距离 8.6%。性能最低的是汉明距离，只有 53.1%。可以看出，余弦距离比较适用于基因表达谱的相似度计算。

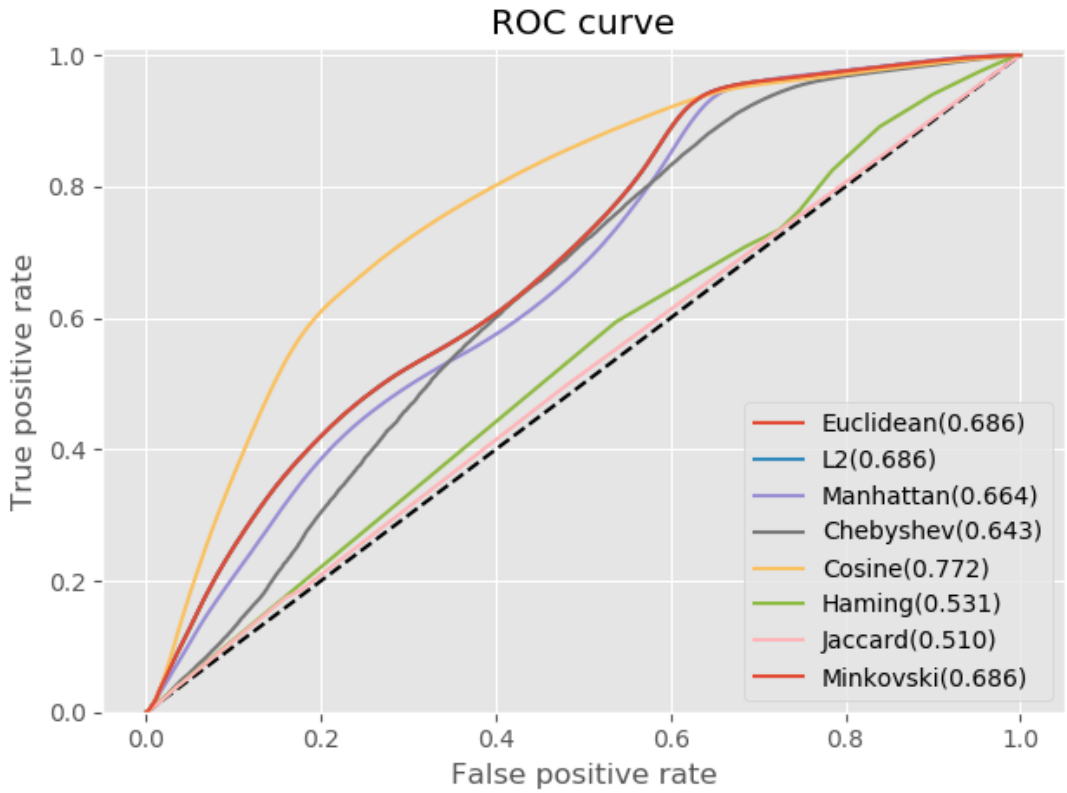

图 3.2 距离度量函数的性能比较

3.4.4 实验三：PC-NCA 算法的性能评估

本实验四种算法做对比：GSEA，NCA，NCA+cosine，PC-NCA，分别在 978 维度、12328 维度数据下做了两组实验对比，每组实验测 MCF7 细胞系分别和 PC3 细胞系和 A375 细胞系做对比、MCF7 在两种不同的扰动实验下做对比，具体实验 ROC 曲线和 KNN 分类准确率如图 3.3 所示。

从图中不难看出，在 978 维的实验组中，MCF7 细胞系和 PC3 细胞系的分类效果没有 MCF7 和 A375 效果好。而对于同一组细胞系在不同的实验组中，度量分类还是比较困难的。而算法 PC_NCA 在实验中，度量分类效果明显好于 GSEA。GSEA 算法和 PC_NCA 算法时间对比如表 3.5 所示，可以发现同一组实验中，PC_NCA 算法所用时间不到 GSEA 的 1/20。

分析实验结果表，PC_NCA 算法明显优于 GSEA，和未优化的 NCA+cosine 算法对比，在区分难度大的数据实验中，往往优化的余弦距离近邻成分分析算法效果更加显著。所有实验组中，PC_NCA 度量性能几乎都是最优的，在不是最优的那一组实验中，和最优结果也是非常接近的，算法性能还是得到了有效验证。

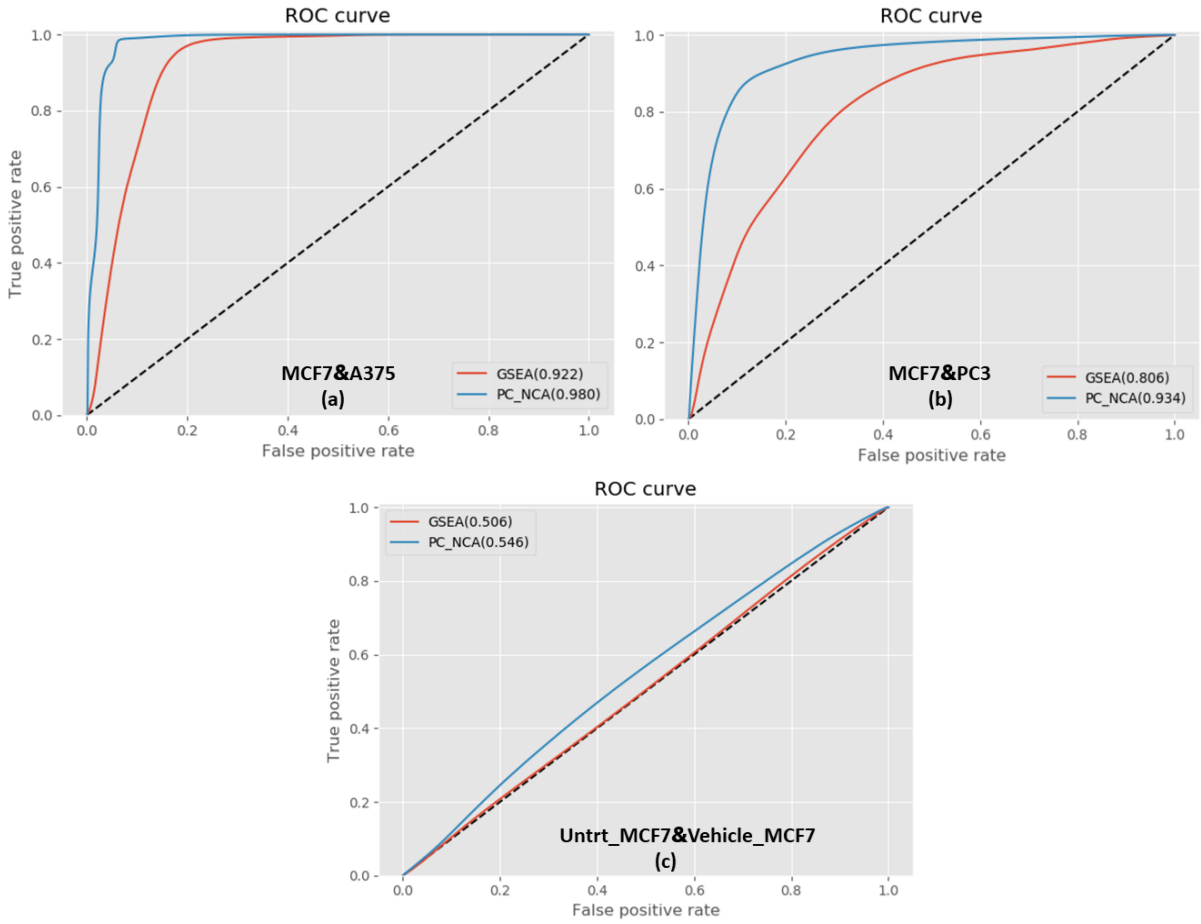

图 3.3 不同实验组 GSEA 和 PC_NCA 性能比较

表 3.5 GSEA 和 PC_NCA 算法时间对比（s）

细胞系/算法	GSEA	PC_NCA
MCF7&PC3	575.4144	22.3967
MCF7&A375	587.4102	10.8390
MCF7&MCF7	612.7979	29.6620

各部分实验总的汇总如表 3.6:

表 3.6 各组实验的分类准确率（%）

细胞系/维度	978			
	NCA	NCA+cosine	GSEA	PC-NCA
MCF7&PC3	77.5	89.6	80.6	93.4
MCF7&A375	97.7	98.1	92.2	98.0
MCF7&MCF7	52.6	53.2	50.6	54.6
细胞系/维度	12328			
	NCA	NCA+cosine	GSEA	PC_NCA
MCF7&PC3	62.3	68.6	68.7	70.2

MCF7&A375	79.7	87.3	88.1	91
MCF7&MCF7	67.2	72.2	70.2	74.6

3.5 本章小结

本章内容是对结合改进余弦距离的近邻成分分析度量学习算法（PC-NCA）的介绍。首先，简介了度量学习的基本原理，然后对于最新的度量学习的研究现状进行了分类别描述。接着，针对典型的度量学习模型如 LMNN,LFDA,ITML 等进行了简单介绍。然后引入 NCA 方法，以及改进的余弦距离的介绍，详细描述了对余弦距离进行改进的必要性和需求。最后，文章对常用的度量学习模型和 GSEA 算法在 LINCS 基因表达谱数据上进行了比对，并对 PC-NCA 算法和常规 NCA、NCA+COSINE、GSEA 算法在度量性能和运行耗时上进行了详细比对，实验证明，提出的 PC-NCA 和 GSEA 算法对比，无论在度量性能上还是运算耗时上都有了较大提升，是一种较为适合基因表达谱相似性计算的度量方法。

第四章 基于深度学习的表达谱度量学习算法

4.1 引言

深度学习能够从样本中自动获取数据特征，因此本章采用深度卷积神经网络自动提取特征展开对 LINCS 基因表达谱数据的研究。深度学习经过搭建多个具备数据处理能力的的计算模型来实现对特征的学习，进而实现对样本的抽象表示。深度学习能够实现高层次、抽象表达来发掘数据复杂结构当中的隐藏信息和关系，实现的方法就是组合大量而简单的非线性模型用于特征的获取和学习[73]。深度学习的关键点在于，学习到的特征是通过一个通用的学习过程从数据中学到的，而不是利用人工工程来设计的。

近些年以来，深度学习结合其他领域的研究获得了较大突破，然而，结合生物学信息学的应用起步却比较缓慢，直到 2015 年后才被应用于 DNA 测序分析。但是伴随多种生物大数据数据集的公开（比如 NIH LINCS 计划）以及多类开源工具的广泛应用（例如 TensorFlow，Pytorch，Kares 等等），生物信息学上的深度学习应用前景非常广阔。深度学习的学习结构通常来说是具有层次性的，它还具有模拟人大脑皮层的较为完整的理论基础，由于大脑皮层在某些部分比如视觉方面同样也是分层次的，故而潜在的视觉皮层对那些潜在的特征会比一般的特征要敏锐一些。综上所述，深度学习在人工智能领域中必将发挥一定的作用，因为有比较多的应用需求和生物神经的理论支撑。本文针对 LINCS 数据的维度高非线性的特点，设计了一种以 DenseNet[75]为特征提取器，通过 Center Loss[76]和 CrossEntropy Loss 进行优化，然后结合 Siamese 度量架构和余弦函数的深度孪生网络（DeepCDNet,Deep Cosine Distance Dense-Siamese Network）。实验证明，相比较传统的基因表达谱相似度计算的富集方法 GSEA，这种方法显著提高了数据之间的距离度量效果，并且可以有效的减少基因表达谱分析时候的人工干预，避免深度网络易发生的过拟合现象，该算法模型有较强的可迁移性和泛化能力。

4.2 卷积神经网络

卷积神经网络 CNN 属于一类比较典型的而且使用范围较广的深度学习方法，Hubel 和 Wiesel[77]在试验动物视觉皮层的时候首次提出来。权值数量的减少以及模型网络的复杂程度的降低都是通过权值的共享来实现的，基于此，网路模型和生物的神经网络变得更加相似，CNN 已经在图像的识别和语音分析上运用比较成功，下采样和权重共享以及局部感受野是 CNN 的三个核心方法，通过这些方法

能够实现位移和放缩等功能[73]。多层 CNN 一般的构成包括卷积层、池化层、全连接层和分类层。局部感受野意味着每层神经元仅连接到上层的小邻域中的神经元。 有了感受野的存在，单个神经元就能获得低层次数据特征。权重共享让卷积神经网络有着较少的参数和较少的训练数据。下采样降低了特征的分辨率，并使得位移、缩放和其他形式的失真不变[74]。

4.2.1 卷积神经网络的结构和作用

CNN 的核心组成部分包括数据输入层、卷积计算层、激励层、池化层和完全连接层[78]。

数据输入层：需要进行的操作一般是对原始图像的预处理，具体是平均化、白化以及归一化。 作为 CNN 名称的由来以及核心结构，卷积计算层包含两个核心处理：一是窗口滑动（receptive field），二是本地关联[79]。

池化层处于多层卷积层之间，依据参数以及压缩数据量，来缩小过度拟合的程度。完全连接层，作用是用权重把层级之间的神经元连接起来，一般位于卷积神经模型的末端。传统的卷积神经网络参考图 4.1[79]：

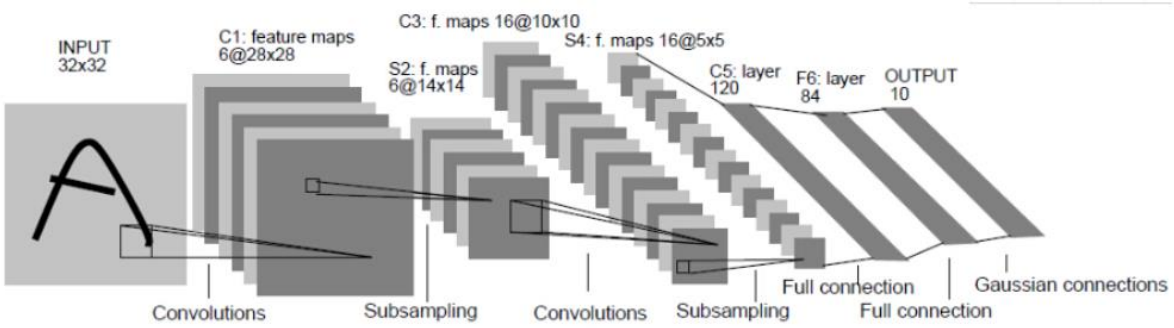

图 4.1 卷积神经网络构成图

从图 4.1 中可以看出，输入样本首先与一组可训练的卷积核进行卷积，以获得一组特征映射 C1 层。通过将前一层的特征图的局部区域与卷积核卷积来获得特征图的 C1 层中的每个神经元。故而，特征图中的每个神经元仅与前一层特征图的局部区域有关系，而且是该区域的特征的线性组合。如果特征的线性组合只能在神经网络中完成，则多级网络结构将退化为主要网络结构。因此，在特征图上的卷积运算之后，将激活函数层添加到所获得的特征映射 C1，即，对特征进行非线性变换以增强网络的特征提取能力。在对原始输入样本进行卷积后，获得相应的特征映射，并对每个特征映射开展子采样，即获得图中 Xn 的像素区域的最大值（或平均值），其值由价值取代，获得子采样层 S1 层。在二次取样操作之后，要素图的数量不会改变，但要素图的大小将减小到原始的 1/N。正是由于网络中的卷积运算和子采样运算这两种结构，CNN 对输入图像具有很高的抗失真能力。然后，对子采样的 S1 层进行卷积以获得特征映射 C2 层，并且再次对 C2 层进行

子采样以获得子样本层 S2 层。最后，将这些二维特征图连接成一维向量，并输入 Softmax 分类器以获得最后的分类结果。

4.2.2 卷积神经网络的训练过程

CNN 的主要结构是卷积层和子采样层。这种构成能够提取高级判别特征并降低计算复杂度，同时确保输入图像结构的不变性。

（1）卷积层的计算

卷积层的计算过程是将前一层和卷积核的特征映射卷积得到相应的线性特征，然后通过激活函数，即该层的特征映射获得非线性特征。这一层的特征图与前一层的特征图相关，并且经过前一层的所有特征图的非线性组合获得。卷积层的运算流程如公式 4.1[77]：

$$f\left(x_j^l\right)=f\left(\sum_{i\in M_j} f\left(x_i^{l-1}\right)*k_{ij}^l+b_j^l\right) \tag{4.1}$$

其中，l代表层数，k代表卷积核，M_j定义为传输特征图的选项，$f\left(x_i^{l-1}\right)$代表前一层的特征图，偏置b可以根据最后输出的对应特征图而得出。

（2）子采样层的计算

子采样表示要输入要素图的子采样操作。对要素图进行子采样后，数字不会更改，并且大小会更改。子采样操作的公式是 4.2[77]所示：

$$f\left(x_j^l\right)=\mathrm{down}\left(f\left(x_j^{l-1}\right)\right) \tag{4.2}$$

其中，l表示层数，$f\left(x_i^{l-1}\right)$定义为上一层特征图，$down(.)$代表子采样的函数。子采样操作是指在输入图像中找到 n×n 区域的最大值或平均值，并且替换原始 n×n 区域。因此，原始要素图的大小是输入要素图的 1/n。

（3）交叉熵损失函数的运算

在卷积神经网络中，存在两种常见的损失函数，简单的均方误差损失函数和交叉熵函数。如果均方差损失函数和 Sigmoid 激活函数都出现在网络中，则网络中的误差在执行反向传播时容易出现梯度消失。交叉熵损失函数可以在一定程度上避免均方误差损失函数的这一缺点。故而，它通常用作网络中最后一层的误差计算功能。交叉熵损失函数[77]的表示如下：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\log\frac{e^{w_t^T f(x_i)}}{\sum_{l=1}^{T}e^{w_l^T f(x_l)}} \tag{4.3}$$

其中，传输的第i个数据样本代表为$(f(x_i),y_i),f(x_i)$，指的是模型末端的层得到的特征，表示实际标签的是y_i，$f(x_i)$和y_i是相互对应的状态，w代表网络中末端层（全连接层）的参数，$1\{.\}$代表示性函数，假设第i个数据运算得出的类型与已有的真实标签t一样，那么示性函数$1\{.\}$=1，否则$1\{.\}$=0，T表示类，最后的样本数可以

用 N 来表示。

前向传播和反向传播属于 CNN 训练进程的两个主要部分。 前向传播是通过各种网络层计算输入样本，最终输出层给出网络获得的分类结果。将样本的实际标签与网络计算结果之间的差异用作反向传播中使用的误差，而且依据反向传播更新结构里的参数，直到达到最大迭代次数，并停止更新网络参数。

4.3 Siamese CNN 网络

Siamese 网络是由 Yann Lecun 于 2005 年所提出。两张图片取代传统的一张图片作为输入是其算法思维的核心点[80]。Siamese 网络是衡量相似性的方法，如果有许多类别时，它就能用于分类或者类别识别等，但每个类别的样本数量很少。它的提出原因就在于每个类别的样本太少，从而导致无法获取想要的结果，这个时候训练数据的新的方法需要被提出来，从而产生了 Siamese 网络。它的核心思想是：

定义的函数能取得目标空间中的输入数据映射，并且利用简单距离（欧几里德距离等）在目标空间中比较他们的相似性。在训练阶段，来自同一类样本对的损失函数值被最小化，并且来自不同类样本对的损失函数值被最大化。

$G_W(X)$代表给出的映射函数，那么 W 表示参数，这样做的目的是想要得到参数 W，能够在X_1和X_2是同一类型的条件下，相似性度量的结果尽量小；如果X_1和X_2是不同类型的时候，相似性度量的结果$E_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\|$可能会较大。训练过程的输入数据或者样本都是成对的，假设X_1和X_2是同一类数据，就要使得损失函数最小化；而假设X_1和X_2是不同类别的时候，则需要损失函数非常大。$G_W(X)$除去可微存在的需要之外，不需要其他前提，因为对于成对的样本输入，这里的两个相同的函数 G 具有相同的参数 W，这种结构是两边对称的，它被称为暹罗网络或者孪生网络[81]。

具体的网络结构：

Siamese 网络的主要结构如图 4.2[81]所示：左侧和右侧的两个网络是相同的网络结构，权重 W 被左右所共享，（X1，X2，Y）代表传入的数据对，而 Y = 0 的时候，X1 和 X2 属于同一个类型，Y = 1 代表它们是相反的类型。 也就是说，对于两个不同的输入，X1 和 X2 同类对是（X1，X2，0），异类对则是（X1，X2'，1），最后得到$G_W(X_1)$和$G_W(X_2)$，它们属于低维空间，是经过 X1 和 X2 映射而获得。最后将他们和最后得到的能量函数$E_W(X_1, X_2)$展开比对[82]。

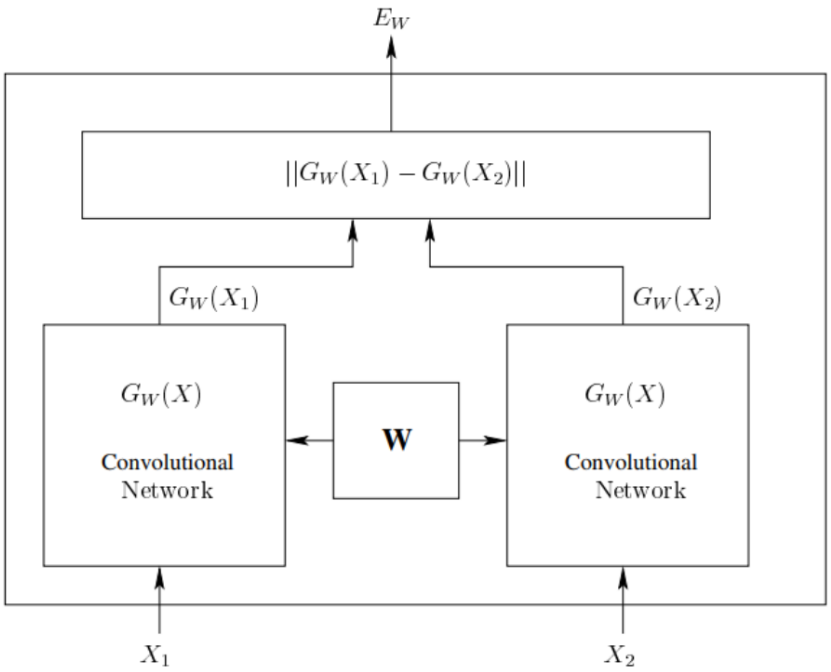

图 4.2 Siamese Network 的主要结构

4.4 DenseNet 网络

在深度学习网络中，梯度消失的问题会根据模型结构的深入而变得严重。近些年，大量研究都针对这个问题提出了自己的解决方案，其中比较认可的比如 ResNet [84]，虽然这部分算法的网络结构不同，但核心两者都是为了使连接层之间的距离更近。DenseNet 能够延续这个想法，就是保证各层次网络相互间能传输最大的信息量的前提下直接连接所有层。DenseNet 模块的结构展示图参考图 4.3[84]。在一般的卷积神经网络中，有一个 L 层，会产生 L 连接，但 densenet 结构有 L(L+1)/2 连接。通俗来说，前面层的所有结果都会传入下一层的网络中。如下所示：输入x_0，H1 输入为x_0（输入），H2 输入为x_0和x_1（x_1为 H1 的输出）。

参数少而且网络窄是 DenseNet 的显著特征。很大一部分原因是由于密集块的设计。密集块中每个卷积层的输出特征映射的数量非常小（小于 100），不会像某些网络那样有成千上万个宽度。另外，这种连接能够让梯度和特征的传输更加有效，而且网络训练比较简单。如前所述，当网络深度更深时，更可能发生梯度消失问题。究其缘由是梯度信息和输入信息在许多层之间传输。现在，这种密集连接相当于直接连接输入和每一层。故而梯度消失能够减轻，网络的层数不再是个限制条件。同时，密集连接所产生的正则化效应对过拟合能起到抑制作用。

Huang 等[83]提出的方法中具有两个公式，被用于解释 DenseNet 和 ResNet 之间的关联。原则上理解这两个网络有着巨大的意义。

前面一个是 ResNet，参考公式（4.4）[83]。其中，l表示层，x_0代表层l的输出，

H1 指的是非线性的变换。

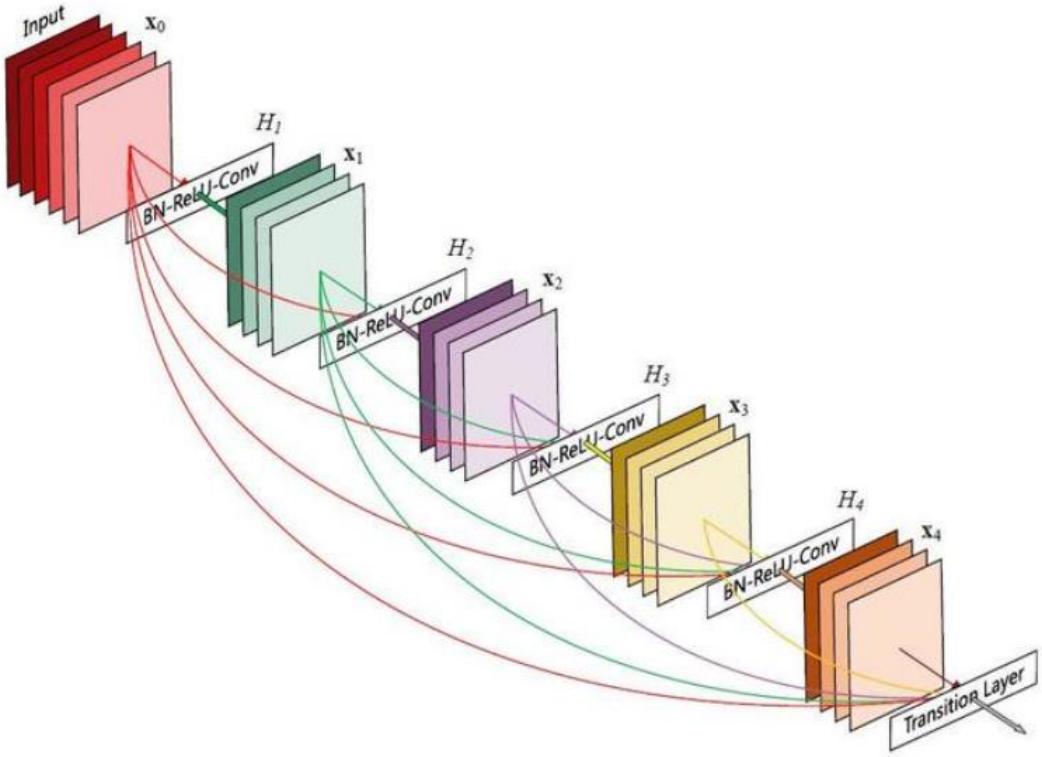

图 4.3 DenseNet-block 结构图

因而在 ResNet 中，层$l-1$的输出和$l-1$的输出非线性转换会成为层l的输出。

$$x_\ell = H_\ell(x_{\ell-1}) + x_{\ell-1} \tag{4.4}$$

公式（4.5）[83]表示的是 DenseNet 将 0 到$l-1$层的输出 feature map 做 concatenation，而前面 resnet 运行的是值的相加，通道数并不发生变化。Hl包括 BN 卷积、RELU 卷积和 3*3 的卷积。

$$x\ell = H\ell([x_0, x_1, \ldots, x_{\ell-1}]) \tag{4.5}$$

4.5 DeepCDNet 距离度量算法

由于 Siamese 网络结构是一种度量学习框架，其适应性很强，故而已经成为图像识别以及语音分析方面的热点问题。DenseNet 网络的优势在于参数更少而且网络更窄。另外，梯度和特征传输的高效性由此种连接来保证，网络训练的简单性亦由其确保。本文针对 LINCS 数据的维度高非线性的特点，设计了一种以 DenseNet[75]为基本特征提取器，通过 Center loss[76]和 Crossentropy loss 进行优化，然后结合 Siamese 度量架构和余弦函数的深度暹罗网络（DeepCDNet,Deep Cosine Distance Dense-Siamese Network）。

4.5.1 网络结构

DeepCDNet 的主要构架如图 4.4 所示：

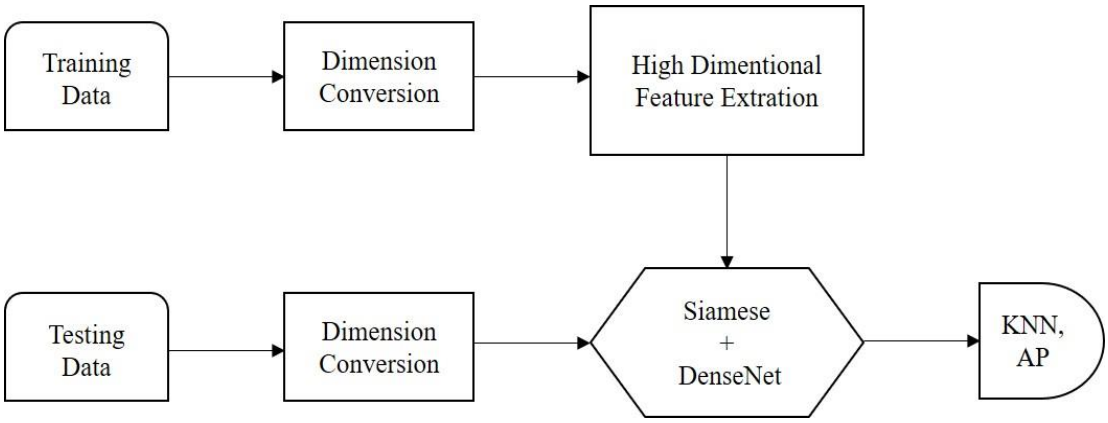

图 4.4 DeepCDNet 的总体框架

DeepCDNet 网络主要包括数据处理模块、高层次特征提取模块，Siamese 距离度量模块和度量验证模块。

在数据处理模块中，我们将样本对转成图片形式的基因矩阵，基因矩阵长度根据表达谱数据的维度决定，对于维度为 N 的样本，转换为 x*x 的矩阵，其中 x 通过公式$x = \sqrt{N}$获得。若 x*x>N，则对多出来的像素位置做补 0 处理，接着进行归一化减均值等数据预处理操作。将不同类别的基因矩阵分别赋予不同的类别标签，并划分训练和测试样本集。

高层次特征提取模块如图 4.5 所示：

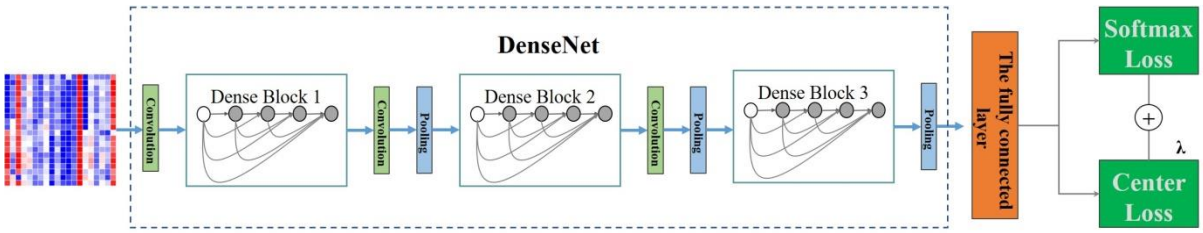

图 4.5 特征提取模块结构图

文章的结构采用三个 dense block 结合，每个 block 里面设计了 16 层卷积，压缩系数设为 0.5，模型增长率设为 12。为了缩小类内间隔以及扩大特征的类间间隔，我们拓展了隐式的度量学习，采用 Center loss 和 crossentropy loss 结合计算损失。Center loss 是 Yandong Wen[85]等人在 ECCV2016 中发表的用于深度人脸识别的损失函数，能确保类内特征间隔的最小化。通过 Center loss 与 Crossentropy loss 的结合，我们提高了模型学习到的高层次特征表达的判别性。我们的最终的损失函数$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_C$[85],具体内容如公式 4.6 所示：

$$\mathcal{L} = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2 \tag{4.6}$$

其中 C_{yi} 是每个类别的聚类中心，λ是训练的时候学习出来的参数。

特征提取采用单向训练模式，训练出来的参数作为 Siamese 距离度量模块的共享权值，距离度量模块的构成参考图 4.6：

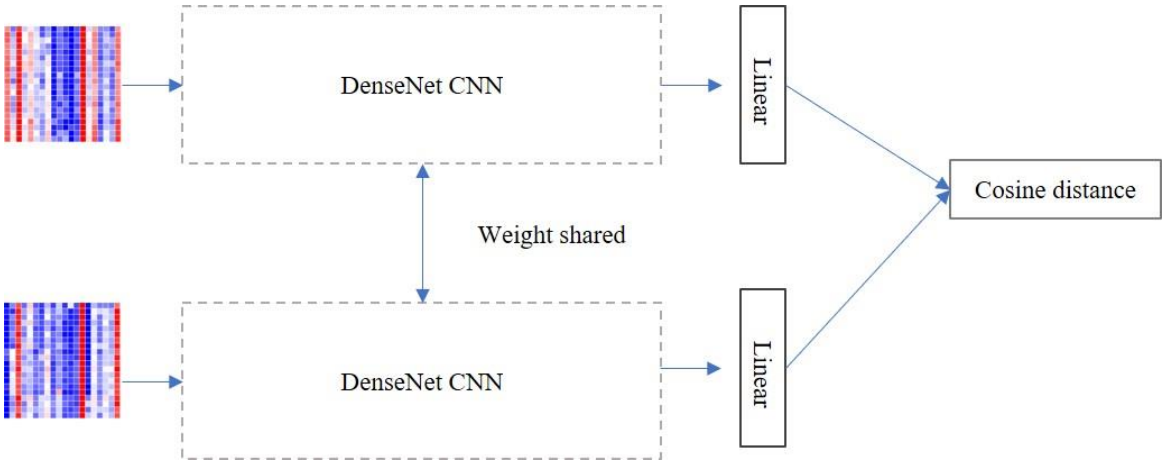

图 4.6 距离度量模型结构

利用两个相同权值的已经训练好的卷积神经网络构建 Siamese 结构的度量计算框架。每次输入两个基因矩阵得到他们 Center loss 中已经定义好长度的的高级特征表达，并求得其特征余弦距离,其距离计算如公式 4.7。

$$\cos(\theta) = \frac{\sum_{i=1}^{M} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}} \tag{4.7}$$

运算获取基因表达谱之间的距离矩阵，通过 KNN 或者 AP 聚类算法，查看分类和聚类性能，与传统的度量学习方法和 GSEA 做比对，得出 DeepCDNet 的度量优越性。

4.5.2 训练过程与收敛

参数训练过程如公式（4.8）所示：

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{4.8}$$

其中，J是损失，θ是网络参数，η是学习率，本文采取隐式度量学习的思路对其进行训练。

在训练过程中，我们分别传入了 978 维度的 LINCS 标量基因和 12328 维度的基因表达谱数据进行测试，收敛效果如图 4.7 所示。

图 4.7 训练了 7 组不同的网络过程，都迅速达到了收敛状态，从途中可以看

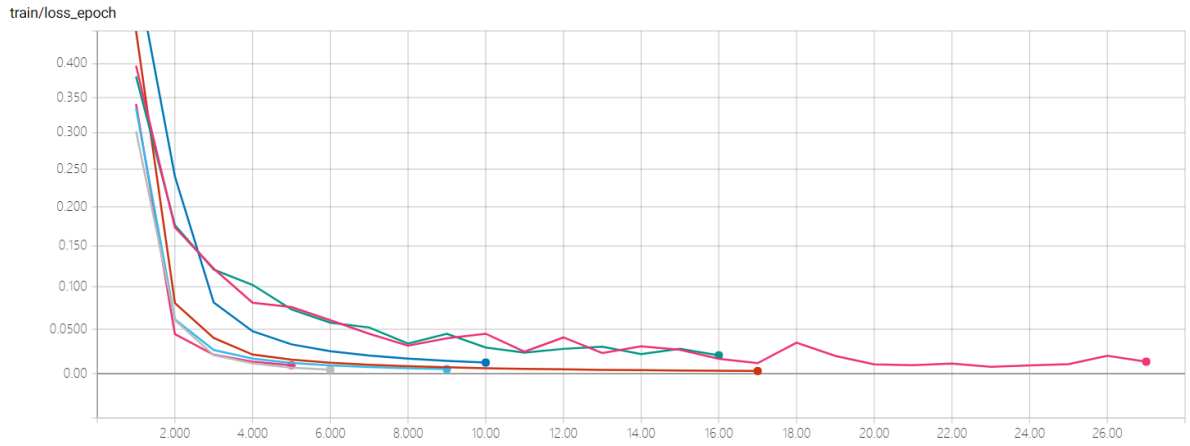

图 4.7 训练收敛过程图

出，训练次数最多的一组 28 次迭代网络收敛，训练次数最少的一次是 5 次达到收敛。可以看出损失函数收敛迅速，网络效果较好。迭代次数最多的洋红色曲线代表的是 MCF7 样本在不同的实验刺激下的收敛情况，蓝色的 16 次迭代曲线和红色的 17 次迭代曲线，表示的是 MCF7 和 PC3，MCF7 和 A375 细胞系两两之间的相似度比对。而后面的四条曲线表示的是样本数量减少为 1000 时候的收敛次数。

网络训练环境配置为：ubuntu 16.04 系统，Pytorch3.5(GPU)+Tensorboard，主板的型号是英特尔 i7-8700K,显卡的型号是英伟达 GTX 1080Ti，内存为 64GB,固态硬盘 256GB。训练的所需具体耗时和设置如图 4.8 所示。在所展示的几组训练过程中，训练耗时最多为 10m29s，在本章实验所有训练网络组中，耗时最多的一组时间为 47 分钟。

Name	Smoothed	Value	Step	Time	Relative
Mar01_16-07-08_zhanRG	0.01650	0.01650	16.00	Fri Mar 1, 16:17:14	9m 22s
Mar01_20-37-50_zhanRG	0.02034	0.02034	16.00	Fri Mar 1, 20:49:29	10m 59s
Mar05_14-16-00_zhanRG	4.3624e-3	4.3624e-3	6.000	Tue Mar 5, 14:20:31	3m 45s
Mar05_14-53-40_zhanRG	0.01223	0.01223	10.00	Tue Mar 5, 14:54:46	56s
Mar05_14-55-10_zhanRG	3.1215e-3	3.1215e-3	16.00	Tue Mar 5, 14:56:46	1m 26s
Mar05_14-59-18_zhanRG	4.9271e-3	4.9271e-3	9.000	Tue Mar 5, 15:01:17	1m 43s
Mar05_15-02-16_zhanRG	9.2953e-3	9.2953e-3	5.000	Tue Mar 5, 15:03:41	1m 5s

图 4.8 训练值和耗时

4.6 实验评估

为了保证实验的准确性，实验在 LINCS 数据集上提取了三种典型肿瘤细胞系：A375（皮肤癌细胞）、MCF7（乳腺癌细胞）、PC3（前列腺癌细胞）在不同的实验条件、不同刺激、不同的数据维度和样本下进行了的多组实验，实验得出 DeepCDNet 度量效果远远好于 GSEA 算法和常规的度量学习算法。

4.6.1 数据集

本文在 LINCS 数据中提取了两组实验数据：

1）基因表达谱都是基础的 978 维度的数据，提取了 A375、PC3、MCF7 在实验类型 untrt、control vector 和 vehicle control 下的 2000 个样本，在基因沉默（untrt）下进行了 A375 和 MCF7 的度量实验，MCF7 和 PC3 度量实验。另外还比较了 MCF7 在不同的对照组实验中的度量实验，测试和验证比例为 1：1。

2）基因表达谱都是基础的 12328 维度的数据，提取了 A375、PC3、MCF7 在实验类型 untrt、control vector 和 vehicle control 下的 2000 个样本，在基因沉默（untrt）下进行了 A375 和 MCF7 的度量实验，MCF7 和 PC3 的度量实验。另外还比较了 MCF7 在不同的对照组实验中的度量实验，测试和验证比例为 1：1。

所有的实验数据都采取随机抽取，并且经过 5 倍交叉验证。

4.6.2 实验结果和分析

实验的验证算法是 GSEA 富集分析和 LMNN 等几种典型的度量学习算法，比对各种算法获得距离矩阵后输入 KNN 的分类效果，查看分类准确率。第一组实验 ROC 曲线和 KNN 分类准确率如图 4.9 所示。

通过第一组实验结果，可以看出本在不同细胞系度量对比计算中，本文算法具有非常良好的度量性能，分类准确率维持在 97%以上。在同一种细胞系不同实验组的基因表达谱对比实验上，其他算法的度量性能普遍下降只有百分之六十多的情况下，DeepCDNet 分类准确率依然维持在 98.2%。三个实验中，本文算法较 GSEA 相比，分类性能比 GSEA 提升率分别为 8.24%，21.72%，42.11%。分析其中原因，不难发现在 MCF7 和 A375 对比实验中，由于基因表达谱差距较大，所有算法普遍性能提升了，而本文算法准确率可以达到 99.8%，已经接近百分之百的准确率。而在基因表达谱数据差距不大，常规度量方法性能普遍下降的情况下，DeepCDNet 性能依旧高达 98.2%,说明本文方法不依赖数据的度量容易程度，而是从数据本身的维度上挖掘基因之间的内在关联和提取可靠的特征来支持良好的度量性能。具体的准确率和时间消耗如表 4.1 所示。

在时间开销上，可以发现，GSEA 富集方法时间花费普遍在 600 秒左右，而常规的度量学习算法时间开销上往往小的多，尤其是 LFDA 算法，在 k=2，dim=50 的参数设置下，训练和运行时间不到 4 秒，而文本算法由于需要训练网络，因此在训练网络过程中消耗了一些时间，基本一次训练网络需要半个小时，而数据验证耗费的时间较少。

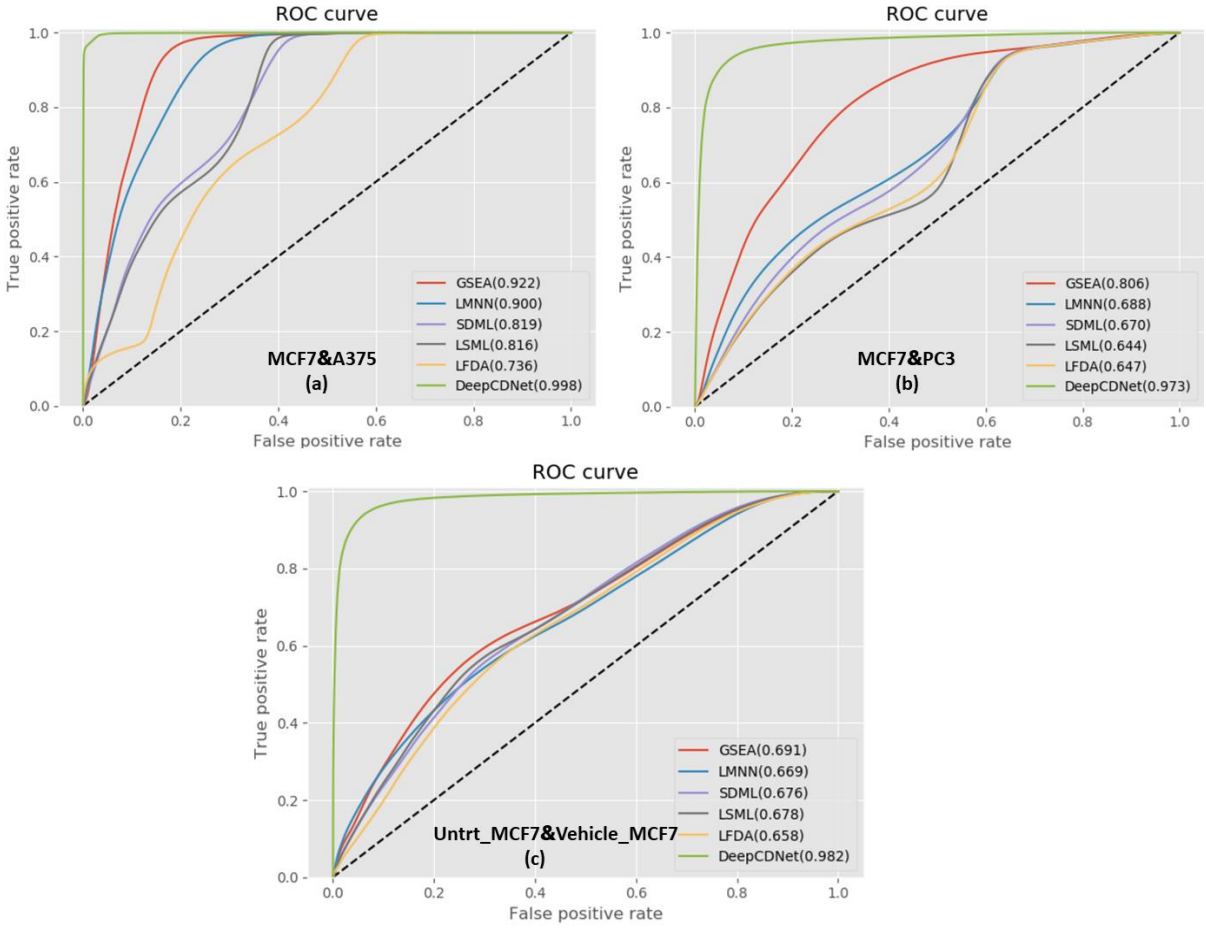

图 4.9 978 数据维度下各类算法分类准确率

表 4.1 978 数据维度下算法准确率和时间消耗

算法	LMNN	SDML	LSML	LFDA	GSEA	DeepCDNet
实验组			MCF7&A375			
准确度（%）	90	81.9	81.6	73.6	92.2	99.8
时间（s）	49.49	3.38	66.62	3.80	613.81	1833.32
实验组			MCF7&PC3			
准确度（%）	68.8	67.0	64.4	64.7	80.6	97.3
时间（s）	56.16	3.78	16.50	3.85	589.24	1890.40
实验组			Untrt_MCF7&Vehicle_MCF7			
准确度（%）	66.9	67.7	67.8	65.8	69.1	98.2
时间（s）	51.74	3.44	8.24	3.81	576.76	1756.66

为了展示本文算法的性能，本章在第二组实验数据上做了进一步实验，第二组 ROC 曲线和 KNN 分类准确率如图 4.10 所示，具体的时间消耗准确率和如表 4.2 所示。

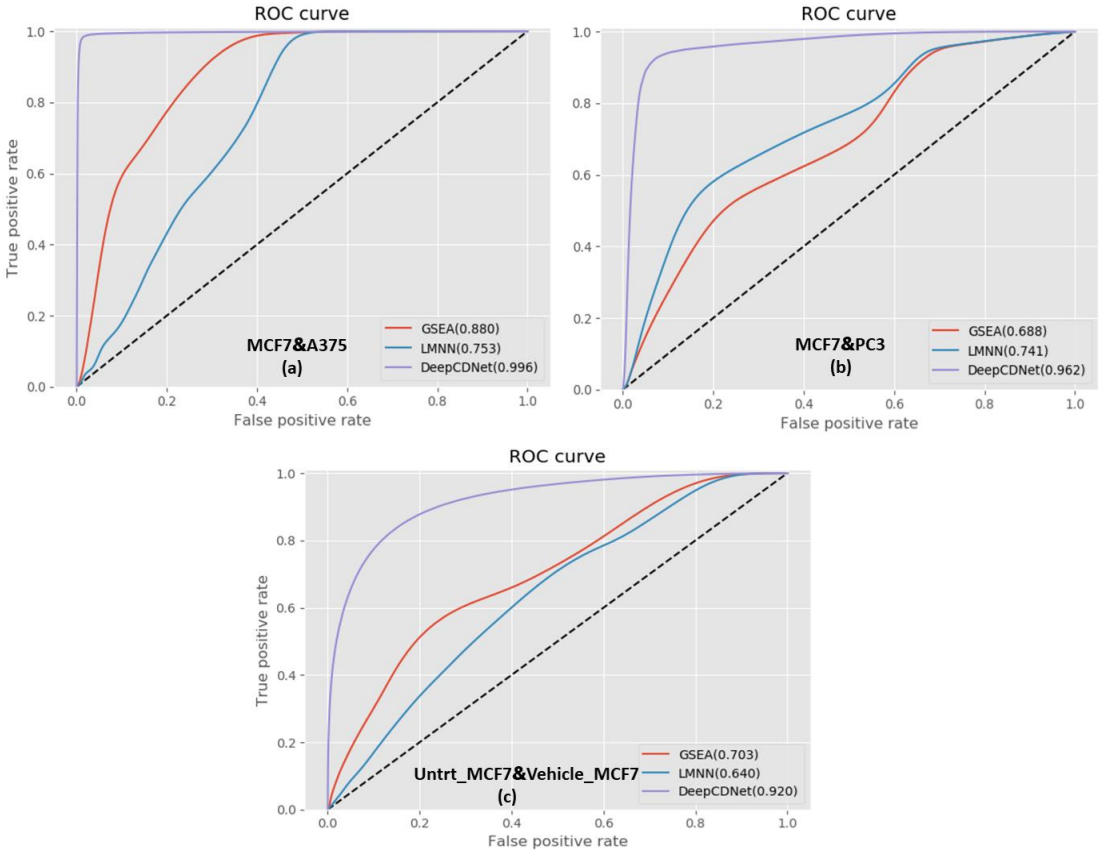

图 4.10 12328 数据维度下各类算法分类准确率

表 4.2 各组实验准确率和时间消耗（%，s）

实验组	MCF7&A375		MCF7&PC3		MCF7&MCF7	
	准确率	耗时	准确率	耗时	准确率	耗时
LMNN	75.3	5832.5	74.1	5770.1	64	6003.5
GSEA	88.0	2426.1	68.8	2447.7	70.3	2515.6
DeepCDNet	99.6	2120.3	96.2	2108.9	92.0	2180.7

通过第二组实验，不难得出以下几点：一是随着数据维度由 978 变成 12328 维度，DeepCDNet 算法依旧保持的较高的度量分类正确率。在三个实验中，比 GSEA 算法分类性能的提升率依次为 13.2%，39.8%，31.4%。同第一组实验类似，当数据之间的相似度较低，容易度量距离的时候，本文算法依旧在分类准确率中都是最高的，实验获得 99.6%的准确率。

二是在数据之间相似度很高，很难取得好的度量效果的时候如 untrt 实验组中 MCF7 细胞系和 vehicle 实验组中 MCF7 细胞系做度量学习，常规的度量算法和 GSEA 的度量效果都有所下降，分类准确率都不高的情况下，分类准确率还是高达 92%。

三是随着数据维度的增加，常规的度量学习算法如 LMNN 时间消耗成倍数增加，所消耗的时间是同等实验中 GSEA 的两倍以上，基本上需要 5700 秒以上。而本文提出的算法虽然需要训练网络模型，但是相比较 978 维度的数据，训练 12328 维度的数据所需要的时间并没有增加多少，基本维持在 2100 秒左右，也比常规的 GSEA 算法的时间要少一些。故而，不难得出，数据维度的增长，DeepCDNet 算法在维持较好的 ROC 曲线以及较高的分类准确率的同时，训练所需要的时间并没有大幅增加，算法的优势随着数据维度的增长而逐渐凸显，实现了在时间上和分类准确率上的优势体现。

另外，虽然算法在训练阶段耗费一定时间，但是在如今硬件配置价格普遍下降的情况下，这种算法还是有良好的拓展性和参考性。

4.7 本章小结

本章首先针对目前深度学习在 AI 和各领域的火热应用和生物学上的应用做了介绍。接着分析卷积神经网络的基本构架，剖析了卷积神经网络的结构和各种网络层的效果。然后，对 Siamese 度量学习构架进行了介绍和分析，简述了 DenseNet 网络模型的基本情况。在引入二者分析的情况下，提出二者结合的原因和构想，同时，优化了损失函数，为了扩大特征的类间间隔以及缩小类内间隔，模型拓展了隐式的度量学习，采用 Center loss 和 Crossentropy loss 结合计算损失。使收敛更加迅速。在距离函数的选择上，采用比较适用于表达谱数据的余弦函数。最后提取 LINCS 数据在 MCF7、A375、PC3 细胞系上 978 维度和 12328 维度进行了实验验证，证明的该算法的在度量学习上的有效性和优越性，而且随着数据维度的增加，算法在保证度量性能的同时在时间消耗上的优势逐渐凸显。

第五章 基于字典学习的表达谱分类算法

5.1 引言

LINCS生物大数据一个重要的应用就是针对基因表达谱开展的分类和聚类研究。由于 LINCS 数据大部分都是各种人体肿瘤细胞系基因表达谱，因此研究其分类性具有十分重大的意义。经过 LINCS 表达谱的分类探究，我们可以从分子层面认识肿瘤的致病机理，从本质上认识肿瘤，并为彻底治疗肿瘤提供基因层次的解决方案。同时，借助基因表达谱数据利用计算机技术对肿瘤进行鉴定和分类还有许多常规医学分类方法不可比拟的优点。首先，利用机器学习等算法借助基因表达谱进行肿瘤分类可以避免医学工作者的主观臆断，实现分类全自动化，基本的目标是提高准确率，同时需要在实验消耗和时间开销上有所减少。其次，利用基因表达谱数据对肿瘤进行分类可以同时对多个病人的表达谱进行分类,速度更快，效率更高。最后，对肿瘤基因表达谱进行分类还能对那些尚未发生明显病变的组织做出及早预测，使病人尽早治疗，提高病人生存率。深度学习的快速发展提供了更好的平台给稀疏字典学习。学习到一个完备字典是字典学习的首要任务，从字典中挑选出少量的字典原子，给定的原始信号是通过对所挑选的字典原子的线性组合来近似体现的。

本章立足于 LINCS 数据，着重分析了稀疏字典学习中的有监督学习算法，探究了其根本原理和特点，并将之与典型的字典学习方法作对比，从而引出基于共享字典学习的 LINCS 数据分类算法，为 LINCS 基因表达谱相似性研究的拓展应用开辟了新的方向。

5.2 稀疏表示和字典学习

5.2.1 稀疏表达模型优化

稀疏表示因为具有良好的特征提取和表征能力,慢慢进入相关研究者的视野，在部分场景使用取得了良好成绩,也引发出来越来越多的与之相关的探究和运用。稀疏字典作为稀疏表示的要点之一，它的提取工作是稀疏编过程中的重要步骤和内容，而字典学习是稀疏理论的构成核心之一[86]。

稀疏表示的根本思想是，假定信号可以用先期定义的某些原子线性组合来体现。假定初始的信号$x \in R^N$可以用一个数据矩阵$D = [d_1, d_2, ..., d_L] \in R^{N \times L}(N < L)$来表达[86]，可以得到：

$$x = D\alpha \tag{5.1}$$

其中$\alpha=[\alpha_1,\alpha_2,\ldots,\alpha_L]^{\mathrm{T}}\in \mathrm{R}^L$代表稀疏信号，即：仅包括$K$个非零的值，其余的都表示成零，K 的稀疏性能体现在 K 的非 0 个数上。详细表述如图 5.1 所示，图中α中的深色是非零构成，其余的都表示成零。其他的通过D获取的信号都具有稀疏性。

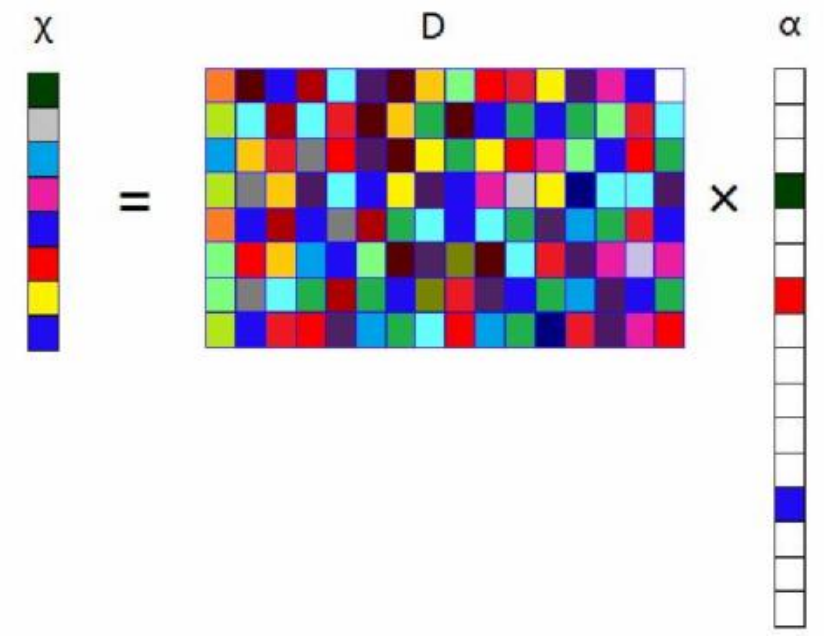

图 5.1 稀疏示意图

稀疏表示具体表示可以看成：

$$\min \|\alpha\|_0, \quad \text{s.t.} \quad x=D\alpha \tag{5.2}$$

式（5.2）中α作为系数，所代表的是x，D就是通常应用的稀疏变化矩阵，就是我们通常所说的字典。d_i是字典里面的原子表示。$\|\cdot\|_0$定义为l_0范数，其具体表示的非 0 的元素量。故而不难得到，从信号所在的角度出发来看稀疏表示的特点具体包括 2 个，即稀疏性和过完备性。具体的完备性表现在存在的矩阵里面的原子数量要比信号所具有的维度数高，正交基字典在稀疏压缩进程中能表现平稳的特性。在理想情况下，公式（5.2）里提到的稀疏编码并不难得到，然而这个目标的本质可归纳为 NP-Hard。知名学者 Donohol[87]和 Taol[88]已经证明，范数l_0的凸优化能够由其所在的局部问题转化而来[86]。

$$\min \|\alpha\|_1, \quad \text{s.t.} \quad x=D\alpha \tag{5.3}$$

为了表述方便，上述两个公式被表示为：

$$\min \psi(\alpha), \quad \text{s.t.} \quad x=D\alpha \tag{5.4}$$

上面提到的内容是比较理想的状态，即约束的条件是一样的，但在现实的使用场景中，噪声会起到非常大的干扰作用，故而可以将$x=D\alpha$转换为:

$$x=D\alpha+z \tag{5.5}$$

式子中的z可以看成是现实情况下的白噪声。故而（5.5）可以转变成公式（5.6），

参考如下：

$$\min \psi(\alpha), \quad \text{s.t.} \quad \|x - D\alpha\|_2^2 \leq \varepsilon \tag{5.6}$$

其中 ε体现噪声的强度亦或稀疏偏差。通常$\psi(\alpha) = \|\alpha\|_p, p \in \{0,1\}$，上述模型具备多种表现形式，像稀疏约束的表现形式:

$$\min \|x - D\alpha\|_2^2, \quad \text{s.t.} \quad \psi(\alpha) \leq s \tag{5.7}$$

其中 s 体现出稀疏性目标，上述模型也可能过渡成为正则化表现模式：

$$\min \|x - D\alpha\|_2^2 + \lambda\psi(\alpha) \tag{5.8}$$

式（5.8）是正则化的参数表现，能够让矩阵的稀疏更加均衡体现，同时具体地展现这个过程中的误差存在。为了尽可能减小重构误差和提高数据分类准确性，除了在稀疏编码搜索算法上进行优化之外，对字典的学习也有重大的作用，好的字典结构可以使稀疏编码以更简洁的方式重构原始信号，并且使得重构误差更小[93]，能对原始信号的实质特色进行本质的描绘[89-91]。

字典学习基本能够归纳为两类，第一类是以重构原始信号为目标的字典学习[92]，该类学习算法不考虑分类信息在字典学习中的作用，称为非监督字典学习方法[93]；第二类是在重构原始信号的过程中，加入训练数据的分类信息，形成以原始信号重构为目标，以分类为导向的监督的字典学习方法[94]。在基于 LINCS 数据相似性拓展应用上，用到的是有监督字典学习算法，故其是本章要介绍的目标。

5.2.2 字典学习分类算法

通过学习数据集中最本质的特征，可以训练出一组具有很强线性表示能力的字典原子，这组字典原子理论上也可以对任何测试样本进行还原度比较高的稀疏线性表示。

训练这组字典原子的过程就被称为字典学习。类比其他相似的算法，稀疏算法是直接利用原始训练数据集，使之作为字典原子表达出测试数据的稀疏线性，而字典学习算法还多出一个训练字典原子的过程，训练好的字典不仅线性表示能力更强，而且占用空间更少。相关字典学习流程如图 5.2 所示。

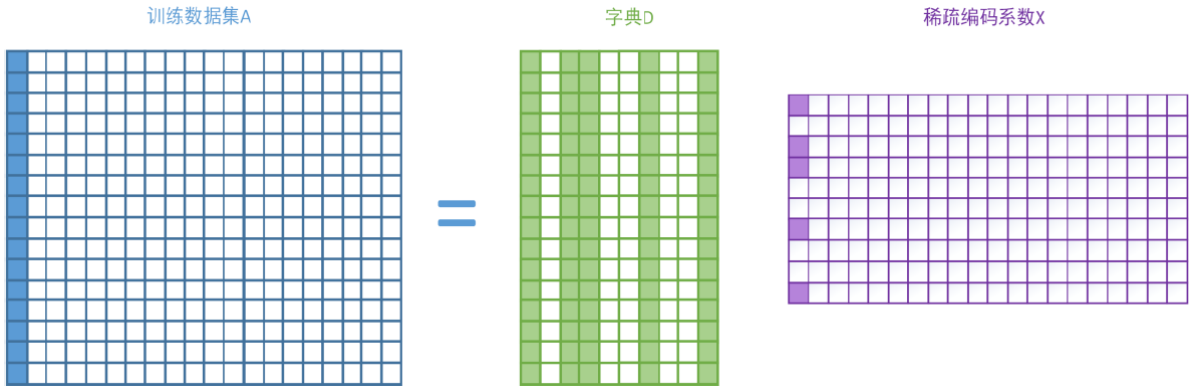

图 5.2 字典学习范例展示

训练集是图中蓝色方块表示，样本由方块中的列表示。绿色的代表字典，列表示的是字典原子。紫色的代表稀疏编码系数，一类对应训练数据中的一列。更具体的讲，字典的学习过程可用式（5.9）[86]来表示：

$$J_{(D,X)} = \underset{(D,X)}{\operatorname{argmin}} \left\{ \|A - DX\|_2^2 + \lambda \|X\|_1 \right\} \tag{5.9}$$

其中，最小化第一项$\|A - DX\|_2^2$代表字典D和系数表达稀疏X的乘积要尽可能的拟合训练数据A，也就是字典D要对训练集A中每一个样本都能误差很小的线 性表示出来。第二项，代表稀疏编码相关系数X要维持均衡的稀疏性，系数λ控制系稀疏程度。常用于解决上述优化函数的方法有 MOD（Method of Optimized Directions）算法，ILS-DLA（iterative least squares dictionary learning algorithms）算法以及 K-SVD 算法[95]等。无论用哪种优化方法，其优化思路都可以用图 5.3 来表示。

如图所示，在初始化字典时可以用随机数列或者是原始数据的相应转换来获得一个原始字典，然后固定住字典 D，即把 D 看成一个已知矩阵，然后计算稀疏编码矩阵 X，即求解一个 LASSO 问题。X 计算完毕后，就可以把 X 固定住，即把 X 看成一个已知矩阵，更新字典 D。在字典更新的过程中，通常方法是一个原子一个原子逐列进行更新，这相当于解一个凸 QP 问题。字典 D 更新完毕后，就可以判断字典 D 和系数 X 是否已能很好的还原训练集 A 或是否已达到最大迭代次数。循环会持续进行在没有达到目标情况下，经过多轮 X 和 D 迭代更新退出循环后，字典矩阵 D 就会被获得，字典训练完成。

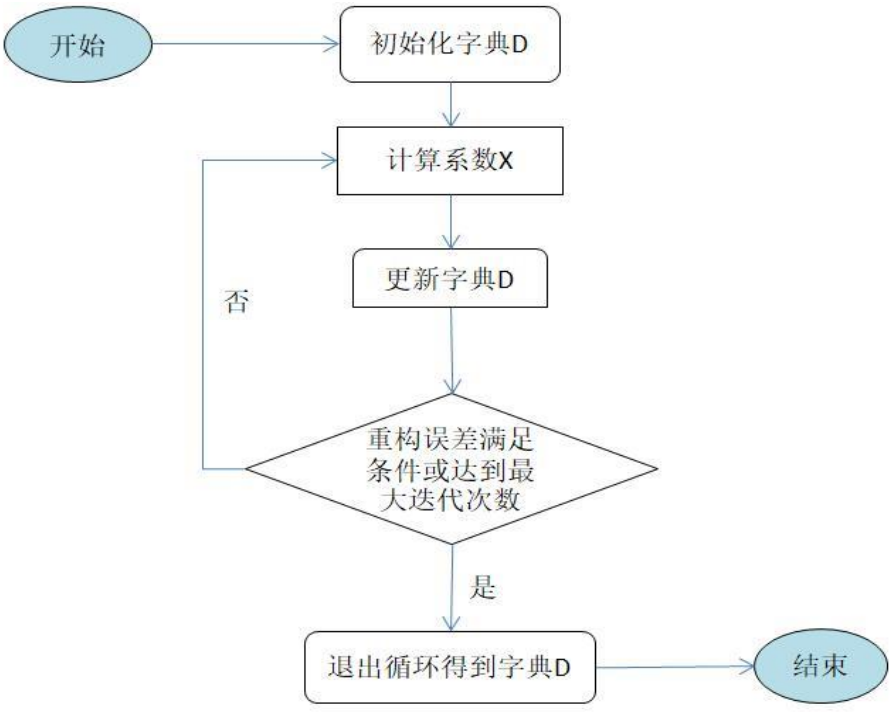

图 5.3 字典学习处理流程

在字典训练完成后，接下来面临的就是分类的问题。利用学习好的字典进行分类有两种常用的方法。其一是为所有的训练集训练出一个公共的字典 D 以及对用这些训练集的稀疏编码矩阵 X，接着使用字典 D 对一个测试样本 y 开展稀疏编码，得到编码系数x，接着用矩阵X和x向量当做特征传入 SVM 等分类器进行分类[96]。但这种措施不运用字典的分类性能，分类效果也较差。还有一种方法是为每一类训练样本都训练出一类子字典，然后合并多类子字典获得所需要字典，测试样本稀疏编码是使用组合的字典对得出的，最后得到不同类别字典对测试样例的构造偏差，以此来判定测试样本的所属种类[97]。这种方式很好地利用了已知标签信息，分类效果也较好，已经成为了目前最通用的字典学习分类方法。

5.3 DPSDL 算法

LINCS 生物数据的基因表现差异性比较大，而且数据维度比较高，部分细胞系的数据量目前还是比较小，和图像识别在模式上还存在较大差距，可能已经存在的图像领域的方法对于 LINCS 数据来说有点不适用。面对这种状况，本论文针对 LINCS 数据所具有的特点，提出共享的字典分类学习方法。该算法训练出一个带判别特征的共享字典，该字典能够表征数据所有类别的样本。在训练子字典的过程中，该方法注重加强字典对同类别样本的线性表示能力，削弱其对非同类样本的线性表示能力。在训练字典的同时，该算法能学习出特定的投影矩阵 P。用 P 投影测试样本可以拉大不同类别的测试样本之间的距离。本文提出的方法被命名为基于判别投影的共享字典学习算法（DPSDL,Discriminant projection shared dictionary learning）。

5.3.1 DPSDL 模型

DPSDL 算法的全称为基于判别投影的共享字典学习算法(Discriminant projection shared dictionary learning)。之所以学习一个共享字典，其原因在于共享字典可以表征每个类别的样本，也就是说，期望Y_c能够经过特定的字典D_i与共享字典D_0的联合来良好地展示。在分类步骤中，将测试样本 y 分解为两部分：由共享字典D_0X_0表示的部分和由剩余字典D_x表示的部分。由于预计D_0X_0不包含特定类的特征，因此可以在进行分类之前将其排除。假设有基于表达谱数据集$Y=[Y_1,Y_2,...,Y_c]$ $\mathrm{Y}\in\mathrm{R}^{\mathrm{d}\times\mathrm{N}}$ ，其中 c 代表总的类别数，$\mathrm{Y_c}\in\mathbb{R}^{d\times nc}$代表类别为 c 的训练集数据的子集，d 代表样本的基因数目，$\mathrm{N}=\sum_{c=1}^{C}\mathrm{n}_c$代表总的样本个数。投影矩阵$\mathrm{Y_c}\in\mathrm{R}^{d\times nc}$，用于将训练集样本以及测试集样本投影到一个更容易分类的 p 维空间。字典$\mathrm{D}=[\mathrm{D}_1,\mathrm{D}_2,\ldots\ldots,\mathrm{D}_\mathrm{C}]$ $\mathbf{D}\in\mathbb{R}^{p\times m}\mathbb{R}$,其中$D_c$为对应训练样本$Y_c$的子字典。

稀疏编码的系数表示为$\mathrm{X}=[\mathrm{X}_1,\mathrm{X}_2,\ldots.,\mathrm{X}_\mathrm{C}],\mathrm{X}\in\mathbb{R}^{m\times N}$，$\overline{D}=[\mathrm{D},\mathrm{D}_0]$表征整个字

典，X^i是Y在D_i上的稀疏系数，$\mathrm{X_c} \in \mathbb{R}^{m \times x_{nc}}$是$Y_c$在$D$上的稀疏系数，$X_c^i$是$Y_c$在$D_i$上的稀疏系数，$\overline{X} = \left[X^T, (X^0)^T\right]^T$，$\overline{X}_\mathrm{c} = \left[(X_c)^T, (X_c^0)^T\right]^T$。具体地，基于最小化以下成本函数来学习判别字典$\overline{D}$和稀疏系数矩阵$\overline{X}$：

$$\overline{J}(\overline{D}, \overline{X}, P) = \frac{1}{2}\sum_{c=1}^{c} \overline{r}\left(Y_c, \overline{D}, \overline{X}_c, P\right) + \lambda\|\overline{X}\|_1 + f(X) \tag{5.10}$$

其中$\overline{\mathrm{r}}\left(Y_c, \overline{D}, \overline{X}_c, P\right)$是判别保真项，它的作用是尽可能地保证$DX$能够最大程度地还原$PY$，如图 5.4 所示，在训练字典的时候，该算法能学习出特定的投影矩阵 P，用 P 投影测试样本可以拉大不同类别的测试样本之间的距离。判别保真项定义为：

$$\left\|\mathrm{PY_c} - \overline{\mathrm{DX}}_c\right\|_F^2 + \|\mathrm{PY_c} - \mathrm{D_c X_c^c} - \mathrm{D_0 X_c^0}\|_F^2 + \sum_{i=1, i\neq c}^{C} \|\mathrm{D_i X_c^i}\|_F^2 \tag{5.11}$$

在式子中，$\overline{\mathrm{r}}\left(Y_c, \overline{D}, \overline{X}_c, P\right)$也可以记作$\mathrm{r}\left(\overline{\mathrm{Y}}_c, \mathrm{D}, \mathrm{X}_c, \mathrm{P}\right)$，其中$\overline{Y}_\mathrm{c} = Y_c - D_0 X_c^0$。通过最小化第一项$\left\|PY_\mathrm{c} - \overline{DX}_c\right\|_F^2$可以使得整个字典$\overline{D}$对都能误差很小地还原出来投影后的每一类训练样本。最小化第二项$\|PY_\mathrm{c} - D_c X_c^c - D_0 X_c^0\|_F^2$可以使得每一类的子字典都能对投影后的同一类训练样本体现出很强的表示能力。因为对训练样本Y_i进行稀疏线表示时，除了D_i以外的子字典的重构贡献要最小化，所以通过最小化最后一项$\sum_{i=1,j\neq c}^{C} \|D_i X_c^i\|_F^2$，便削弱了每一类子字典对非同类训练样本的表现力度，详细的模型流程可参考图 5.4。

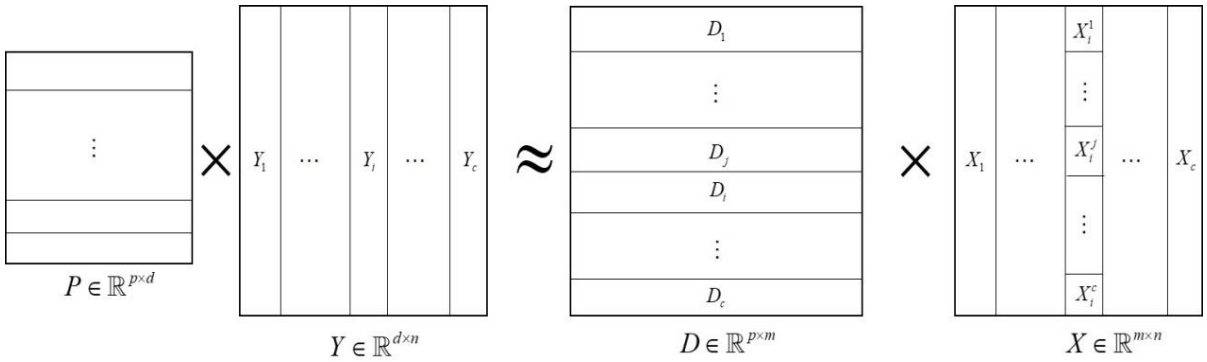

图 5.4 DPSDL 模型

在公式（5.10）中，$\lambda\|\overline{X}\|_1$为稀疏项，它的作用是让稀疏系数矩阵$\overline{X}$保持一定的稀疏性，通过参数$\lambda$来调整矩阵$\overline{X}$的稀疏程度。

为了使字典$\overline{D}$对训练样本 Y 具备相关的辨别能力，我们通过让训练样本 Y 对字典$\overline{D}$的编码系数$\overline{X}$具备判别能力，这能够经过最大化疏编码系数$\overline{X}$的类间分布和最小化稀疏编码系数$\overline{X}$的类内分布来完成。所以在公式（5.10）中，我们加入了判别系数项$f(X)$，$f(X)$定义为：

$$f(\mathrm{X}) = \mu\sum_{i=1}^{C}\left(\|\mathrm{X}_i - \mathrm{M}_i\|_F^2 - \|\mathrm{M}_i - \mathrm{M}\|_F^2\right) - \alpha_1\|\mathrm{PY}_t\|_F^2 - \alpha_2\|\mathrm{PY}_b\|_F^2 + \beta\|\mathrm{X}\|_F^2 \tag{5.12}$$

其中$Y_t = Y - M$，$Y_b = [M_1 - M, \cdots, M_i - M, \cdots, M_c - M]$。M 是训练集 Y 的向量均值表示，$M_i$为第 i 类训练样本$Y_i$的向量均值表示。通过最小化 $f(X)$中的第一项，也就是最小化$\|X_i - M_i\|_F^2$，最大化$\|M_i - M\|_F^2$，使得稀疏编码系数 X 的类内的分布尽可能的小，类间的分布尽可能的大。

对于 $f(X)$中的第二项，我们通过最小化 $-\alpha_1\|PY_t\|_F^2$，也就是最大化$\|PY_t\|_F^2$，会使得所有的训练集样本在经过投影后都尽可能地拉大与整个样本的均值向量的距离，使得样本之间更为发散。通过最小化$-\alpha_2\|PY_b\|_F^2$，即最大化$\alpha_2\|PY_b\|_F^2$，也就使得每一类训练样本在经过 P 投影后，其均值向量也会变得更加发散，类间的间距变大。$-\alpha_1\|PY_t\|_F^2 - \alpha_2\|PY_b\|_F^2$共同作用训练出一个能拉大样本类间距离的投影矩阵 P。由于此时 $f(X)$是非凸的并且是不稳定的，所以我们在 $f(X)$中加入弹性项 $\beta\|X\|_F^2$来解决这个问题。

5.3.2 DPSDL 模型优化

目标函数 J 中包含 3 个变量：X，D，P要想同时得到这三个变量的最优解是很难的，因为这是一个非凸优化问题。幸运的是，固定三个变量中的两个，然后求得另外一个变量的最优解，这个效果转换为了一个临时凸优化问题。因此，本文的 DPSDL 模型可以通过分别固定目标函数的两个变量求第三个变量的方式来迭代求解。

首先用随机数序列使字典初始化D，用训练集Y的 PCA 转换矩阵初始化投影矩阵P。此时字典D和投影矩阵P可被暂时看做已知矩阵，待求的变量只有稀疏编码系数矩阵$\overline{X}$，此时的问题就变成了一个 SRC 中常见的稀疏编码问题。在本文中，我们采用一类一类计算$\overline{X}$的策略，当计算$\overline{X_i}$，即计算相对应训练数据子集Y_i的稀疏的编码向量的时候，$\overline{X}$中其他的编码系数$\overline{X_j}$，$j \neq i$都被固定住。此时目标函数$J(\overline{D}, \overline{X}, P)$变成了下面的模式：

$$\overline{J}\left(\overline{X_i}\right) = \frac{1}{2}\sum_{i=1}^{C}\left\{\left\|PY_i - \overline{DX}_i\right\|_F^2 + \left\|PY_i - D_iX_i^i - D_0X_i^0\right\|_F^2 + \sum_{j=1,i\neq j}^{C}\left\|D_jX_i^j\right\|_F^2\right\} + \lambda\left\|\overline{X}_i\right\|_{\mathrm{h}} + \mu\sum_{i=1}^{C}\left(\|X_i - M_i\|_F^2\right) + \beta\|X_i\|_F^2 \tag{5.13}$$

可以看出式（5.13）中只有一个变量$\overline{X_i}$，其余的P，Y等都可以看做常数，对于这个稀疏编码问题，本文采用了投影迭代法来计算每一类样本的编码系数$\overline{X_i}$，最后把每一类样本的$\overline{X_i}$组合成稀疏编码矩阵$\overline{X}$。

在计算完编码系数矩阵$\overline{X}$后，我们把投影矩阵P和$\overline{X}$看成已知常量，然后更

新字典$\overline{D}$。同计算系数$\overline{X}$一样，我们更新字典$\overline{D}$也是一类一类轮流更新的，此时目标函数$J(\overline{D},\ \overline{X},\ P)$变成了式（5.14）的形式：

$$\overline{J}\left(\overline{D}_i\right)=\frac{1}{2}\sum_{i=1}^{c}\left\{\left\|PY_i-\overline{D}_i\overline{X}_i\right\|_F^2+\left\|PY_i-D_iX_i^i-D_0X_i^0\right\|_F^2+\sum_{j=1,j\neq j}^{C}\left\|D_jX_i^j\right\|_F^2\right\} \tag{5.14}$$

在字典$\overline{D}$和系数$\overline{X}$完成更新以后，还需要计算和更新投影矩阵P，把字典$\overline{D}$和系数$\overline{X}$当成已知的常数，此时目标函数$J(\overline{D},\ \overline{X},\ P)$变成了式(5.15)的形式：

$$\begin{aligned}\overline{J}(P)=\frac{1}{2}\sum_{c=1}^{C}\left\{\left\|PY_{\mathrm{c}}-\overline{DX}_c\right\|_F^2+\left\|PY_{\mathrm{c}}-D_cX_c^c-D_0X_c^0\right\|_F^2+\sum_{i=1,i\neq c}^{C}\left\|D_iX_c^i\right\|_F^2\right\}\\ -\alpha_1\left\|PY_t\right\|_F^2+\alpha_2\left\|PY_b\right\|_F^2\ \text{s.t.}\ \mathrm{PP^T}=\mathrm{I}\end{aligned} \tag{5.15}$$

本文所概括的模型 DPSDL 优化过程如表 5.3 所示：

表 5.1 DPSDL 模型优化过程

DPSDL优化过程：
输入：训练样本集$\mathbf{Y}=\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_C$。
输出：稀疏编码系数矩阵$\overline{X}$，字典D和投影矩阵P。
1. 采用随机的方式初始化字典$\overline{D}$，并通过训练集Y的 PCA 转换矩阵初始化投影矩阵P。
2.重复迭代：
3.通过使字典$\overline{D}$以及投影矩阵P保持固定，计算出每一类样本的子系数矩阵$\overline{X}_i$。
4.组合每类$\overline{X}_i$成为$\overline{X}$。
5. 通过使系数矩阵$\overline{X}$以及投影矩阵P保持固定，计算出每一类样本的子字典$\overline{D}_i$。
6.组合每类$\overline{D}_i$成为$\overline{D}$。
7.使$\overline{X}$系数以及$\overline{D}$字典保持不变，经过多次小步的更新操作让P逐渐接近最理想值。
8.迭代停止（当重构误差平稳或者迭代次数最大的时候）
9.输出（$\overline{X}$，D，P）

5.3.3 分类判定标准

经过以上两小节内容展示的 DPSDL 方法的训练和优化过程，输出的是训练好的字典矩阵D、投影矩阵P以及稀疏的系数矩阵$\overline{X}$。对于输入的数据验证集 y，在分类过程中，先依据P让 y 投影到一个新的空间维度，这个维度中同类之间的距离更近，异类之间的距离更远，这样更容易进行分类操作。在投影空间中的样本$\hat{y}$经过D的线性稀疏表示后，会获得对应的稀疏向量$\overline{x}$。$\hat{y}$分类的标准是用$\overline{x}$和重构误差的距离来判定，具体情况如下：

$$\arg\min_i \left\{ \|\hat{y} - D_i\overline{x}\|_2^2 + \omega \|\overline{x} - M_i\|_2^2 \right\} \tag{5.16}$$

上述公式中，M_i代表第i类训练样本A_i和系数矩阵$\overline{X}$对应的系数的均值向量，ω表示权值参数，本章内容设置为 0.5。

5.3.4 模型的收敛

本节实验验证了 DPSDL 模型和常规字典学习方法在收敛性上的比较，具体的收敛过程如图 5.5 所示，不难发现 DPSDL 在收敛的迭代次数以及损失函数的值上都有比较大的优势。

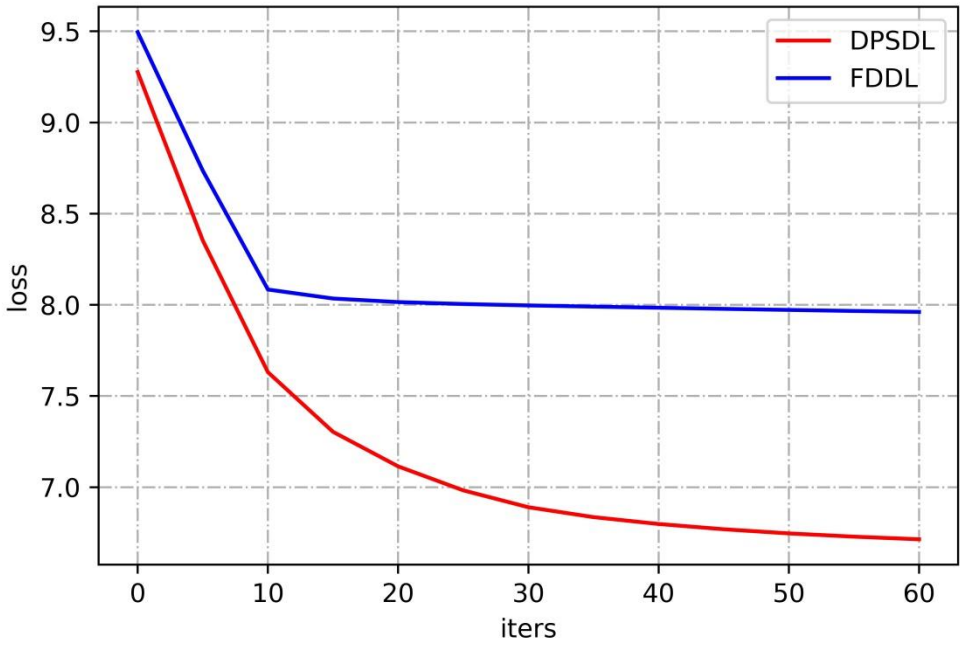

图 5.5 损失函数收敛过程

5.4 实验评估

5.4.1 实验平台和数据集

本章实验在基于 win10 专业版系统的的电脑进行，硬件配置如下：CPU 为英特尔 i7 的处理器型号 6700HQ，,GPU 英伟达型号 GTX970，电脑内存 32GB，固态硬盘 256GB。软件配置为 Pycharm+tensorflow（GPU）版本。

本章实验的数据集包含了 LINCS Level 3 和 LINCS Level 4（GSE92742）两个阶段的数据，第三阶段和第四阶段的区别如图 5.6 所示，如图所示，左边代表第三阶段数据，右边表示的是第四阶段数据。三列代表一个细胞系，三行代表一种小分子化合物刺激的分布情况，如图所示第一幅图代表两种细胞系在两种刺激下的表达谱分布情况，每个细胞系选取 100 个样本做了散点分布图，具体的对比左右对比发现，第三阶段数据还是有区分度的，第四阶段数据很多重合，区分度

低，分类困难，另外，由于前面提到过的 LINCS 数据相比较 GEO 和 NCI 数据在

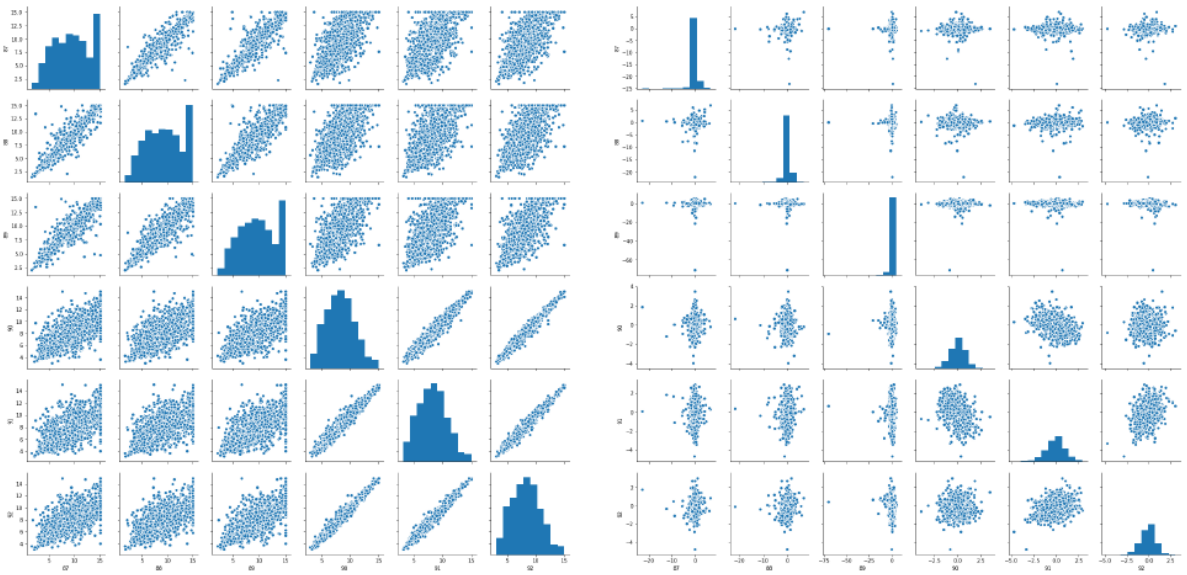

图 5.6 Level3 和 Level4 数据分布对比

来源、标准、处理上的严格控制和统一，用 LINCS 数据第四阶段数据进行分类也相比较更加可靠，验证结果也更加具有说服力。

实验数据分为两组，第一组数据实验数据集包含 Level3 和 Level4 阶段数据，分别提取了五种肿瘤细胞系，包含乳腺、皮肤、大肠、肺、卵巢，对应的类别分别是 6，4，13，11，6，其中乳腺皮肤和肺还包含了一种正常细胞系（非肿瘤细胞）；第一组具体数据情况可以参考表 5.2，T 代表 Tumor（肿瘤细胞系），N 就是 Normal（正常细胞系）。

表 5.2 第一组分类数据集具体描述

细胞类型	基因数	样本数	类别	具体型号
乳腺	978	300	6	5T1N
皮肤	978	200	4	3T1N
大肠	978	650	13	13T
肺	978	550	11	10T1N
卵巢	978	300	6	6T

第二组提取自 Level3 阶段肺细胞系，包括 10 种肺肿瘤细胞系：A549、CORL23、DV90、HCC15、NCIH1694、NCIH1836、NCIH2073、NCIH596、SKLU1、T3M10 和 1 种正常肺细胞系：HCC515。这 11 个细胞系每种提取 100 个在化合物小分子刺激的实验条件（trt_cp）的实验样本作为实验的数据集。根据 12328 维度的数据分别对实验数据进行 2 分类、6 分类和 11 分类的分类实验。数据集的分类可以参

考表 5.3，其中类别一栏 T 代表 Tumor（肿瘤细胞系），N 就是 Normal（正常细胞系）。在本实验中肺部正常细胞系为 HCC515 细胞系。提取实验细胞系的 100 个样本作为数据，训练和测试比例为 1:1。

表 5.3 第二组分类数据集具体描述

维度		12328	
类别数	2	6	11
类型	2T	5T1N	10T1N

5.4.2 实验结果及分析

本小节通过实验来验证判别投影的共享字典学习算法 DPSDL 的效果，对比算法采用了多种算法，除了 SRC 算法[98]之外，还有 COPAR[99]算法、DLSI[100]算法、LRSDL[101]算法以及常规的 KNN 算法。算法的性能由分类精度决定。第一组实验数据的分类结果如表 5.4 所示。

表 5.4 第一组数据分类准确度（%）

	KNN	SRC	CPOAR	DLSI	LRSDL	FDDL	DPSDL
Lv3_乳腺肿瘤	98.55	100	100	99.33	100	100	100
Lv3_皮肤肿瘤	97.5	99.5	99.5	97	99.5	99.5	99.5
Lv3_大肠肿瘤	98.31	98.69	98.0	98	98.0	98.31	98.77
Lv3_肺肿瘤	99.09	99.09	98.73	98.55	98.55	98.73	99.09
Lv3_卵巢肿瘤	96.67	99.33	87.67	99.33	99.0	99.33	99.0
Lv4_乳腺肿瘤	31	83.67	82.67	83.67	83.0	80.67	89.67
Lv4_皮肤肿瘤	36	98.5	98.5	98.0	97.5	96.0	98.0
Lv4_大肠肿瘤	19.08	80.46	77.08	75.23	77.23	69.08	89.07
Lv4_肺肿瘤	27.64	88.18	84.73	85.82	82.91	88.45	89.27
Lv4_卵巢肿瘤	30.33	93.33	94.0	94.0	94.0	90.33	95.66

从第三阶段数据结果可以看出，虽然我们的算法性能中的来说有优势，但是优势并不大。在第四阶段数据结果中我们能发现两点，首先是和常规 KNN 分类算法做对比，发现字典学习对于难以区分分类的情况下尤其表达谱数据具有巨大优势，另外我们的算法普遍在分类性能上具有优势。第二组拓展实验的分类准确度结果如表 5.5 所示。从实验结果来看，在数据维度为 12328 的时候，虽然在二分类中准确率比 LRSDL 算法高，但是结果还是比较接近的，而多分类的时候效果普遍好于常规的方法。

表 5.5 第二组数据分类准确度（%）

维度	12328		
类别	2	6	11
KNN	73	38.67	19.64
COPAR	92	82.33	72.55
DLSI	91	82.4	73.2
LRSDL	92	81.33	70
SRC	89	84	75.27
DPSDL	92	85.24	78.32

综上所属，对于 LINCS 生物信息数据的分类问题，本文提出的 DPSDL 算法具备优良的分类能力。

5.5 本章小结

在本章中，首先对把字典学习算法引入 LINCS 数据分类研究的必要性和可行性进行了分析，然后对系数表达模型进行了介绍和模型的优化。从而引入字典学习，对字典学习的流程和运算推导进程进行了详细剖析。接着针对基因表达谱数据遇到种类多但是样本很少的情况，提出一种新的方法，即判别投影的共享字典学习算法 DPSDL，该算法的本质创新点在于共享机制，所创立的字典可以用来表征所有的类别样本，具有良好的适应性，能大幅提高识别率。文章从运算推理上对 DPSDL 算法进行了验证，然后针对 LINCS 第四阶段的数据特性，做了多组实验验证，通过对肺肿瘤细胞数据的多维度、多分类的情况进行验证，对比传统的 SRC 算法、COPRA 算法、DLSI 算法、LRSDL 算法发现，本文提出的 DPSDL 算法在分类性能具有较大优势，用于 LINCS 数据时获得了良好的实验结果。

结 论

本章是论文的总结，首先回顾论文的主要内容，针对 LINCS 生物大数据中基因表达谱相似度研究的问题，总结了基于改进余弦距离的近邻成分分析度量算法的研究，分析了基于深度学习的度量学习算法。另外，针对肿瘤细胞表达谱的分类算法，创新地应用共享字典学习进行研究，最后针对本文工作中存在的不足和未来 LINCS 数据相关研究的发展趋势，对接下来的工作进行了展望。

工作总结

随着生物技术的迅速发展，多个生物大数据项目（如 NIC LINCS）应运而生，这些资源逐渐变成了生物信息领域研究重要数据资源，但是针对这些数据资源的研究方法和研究技术手段却是相当缺乏的。以此为契机，本课题以 LINCS 数据基因表达谱数据分析为出发点，对基因表达谱数据之间的相似度比对算法进行了深入探究。由于这项工作往往容易被忽视但却是其他研究的基础，因此以目前研究 LINCS 数据相似度的常用算法 GSEA 为衡量标准。GSEA（Gene Set Enrichment Analysis，基因探针富集分析）是普遍使用的表达谱相似度算法。但受限于其本身复杂的计算过程，目前已有的实现工具都难以达到需求的计算相似度精度和速度。本文研究了改进余弦距离和近邻成分分析算法相结合的度量学习算法，另外，随着深度学习在生物学领域的广泛应用，本文还研发了基于深度学习的度量学习算法应用于 LINCS 数据之间相似性研究。而且，由于 LINCS 数据大部分是人体常见的癌细胞系数据，而基于基因表达谱数据的研究为肿瘤分类提供了新的思路，本文还提出了基于共享字典学习的基因表达谱分类算法。本文主要工作包括如下三个方面：

1.提出了结合改进余弦距离的近邻成分分析度量学习算法（PC-NCA, Promotional Cosine Distance Neighbourhood Components Analysis）。本文分析了不同的相似度系数和方法，对基因表达数据相似度计算较为合适的余弦距离进行了改进，除了着力于向量夹角，还通过中心化和归一化使得算法对于在各维度上的值也更加敏感，这样设计符合数据分析的真实情况。对常用的度量学习模型和 GSEA 的算法在 LINCS 基因表达谱数据上进行了比对，并对 PC-NCA 算法和常规 NCA、NCA+COSINE、GSEA 算法在度量性能和运行耗时上进行了详细比对，实验证明，提出的 PC-NCA 和 GSEA 算法对比，无论在度量性能上还是运算耗时上都有了较大提升。

2.提出了基于深度学习的表达谱度量算法。针对目前深度学习在 AI 和各领域尤其是生物学上的广泛应用，本文希望能通过深度学习的算法来学习表达谱数据的特征。因此提出了将 Siamese 度量学习构架和 DenseNet 网络模型相结合的设想，同时，优化了损失函数，使收敛更加迅速。算法中的一个基础但是关键的步骤是数据的处理，通过把基因表达谱数据的维度转换为方块矩阵并且设定为 3*3 的卷积核使得在训练过程中迅速收敛的目的。最后提取不同细胞系的 LINCS 基因表达谱数据开展了多组实验测试，证明该算法的距离度量性能在分类效果精度上高达 99.8%，和 GSEA 相比有了较大的提升。

3.提出一种基于字典学习的 LINCS 数据分类算法。首先对把字典学习算法引入 LINCS 数据分类研究的必要性和可行性进行了分析，从而引入字典学习，针对基因表达谱数据遇到种类多但是样本很少的情况，提出共享字典学习这样一种方法，运用能读取所有数据的样本这样一种思想，提高了识别率。通过对肺肿瘤细胞数据的多维度、多分类的情况进行验证，对比传统的 SRC 算法、COPRA 算法、DLSI 算法、LRSDL 算法发现，改进的算法在分类性能上得到了一定的提高。

本文提出的三种方法均属于 LINCS 数据相似性研究及其拓展应用。其中，基于优化余弦距离的近邻成分分析算法（PC-NCA）和基于深度学习的表达谱度量学习算法（DeepCDNet）用于度量 LINCS 基因表达谱数据之间的距离，二者具有互补性。PC-NCA 运算速度快，比较适用于低维度的基因表达谱数据，虽然度量性能没有 DeepCDNet 好，但是和 GSEA 比，在度量性能和时间开销上有较大进步。而且在许多使用场景中，度量性能要求并没有那么严格，需要迅速的分析出基因表达谱之间的相似度从而应用于差异性表达，PC-NCA 适用于小样本低维度数据和时间开销比较少的场景中。而 DeepCDNet 算法在相似性度量性能上和 GSEA 相比较有绝对优势，可以针对目前 LINCS 数据在第三阶段由基础 978 个基因推导出 20000 多个人类细胞系基因过程中存在的不足进行修正，从系统上改进 LINCS 数据的外推准确度，而且该方法随着基因表达谱维度的提升，时间消耗并没有显著性增长，比较适用于高维度和高度量性能要求的数据分析。进一步的，对于基因推断表达、药物重定位、基因数据特征提取等应用而言，DeepCDNet 都是一个非常好的方法，极具参考意义。基于共享字典的基因表达谱分类算法（DPSDL）属于 LINCS 基因表达谱相似度比对基础上的一个拓展应用，相似性度量研究是基因表达谱数据的一项基础研究，基于这个工作上的拓展应用主要有分类和聚类，这两种拓展都需要用到基因表达谱数据之间的距离，而且所得距离的优劣直接关系到聚类或者分类的效果，所以探究相似性度量在各种分类和聚类方法中的适用性是拓展 LINCS 数据相似性应用的一个重要方向。

综上所述，本文在 LINCS 基因表达谱数据的相似性度量研究和分类上取得了阶段性的进展，但后续仍有很多地方需要改进。

未来展望

针对本课题的不足之处，未来可以从如下几个方面展开深入研究工作：

对于基因表达谱相似性研究的度量距离设定。文章虽然对余弦距离进行了优化并且结合近邻成分分析在 LINCS 数据上取得了良好的度量效果，但是在算法的迁移性上的相关验证还需进一步加强。同时，针对逐渐增加的基因表达谱数据，开发出自适应的度量学习算法应对逐步增长的数据是很有必要的。

虽然本文在第四章提出的基于深度学习的度量学习算法在数据集上取得了良好的度量效果，并且超过 GSEA 一大截。但是，经过数据训练和特征提取后，得出的特征是 50 个提取特征，这些特征来自那些基因及其所占的比例我们是未知的。换言之，特征提取目前只是单向的，并不能反过来推测出哪些基因是我们所需要的对于基因表达起关键作用的基因。这是比较遗憾的一点，如果能够确定这些基因，那么对于我们后续药物定位、关键因子发现等的研究将会是一个质的提升。同时，深度学习因为需要去训练网络，故而如果想要提升效率需要一定的硬件支持，和 GSEA 比对虽然时间有所缩减。虽然随着样本数量的增加，训练网络所花费的时间随着维度的增加并不明显，但是还是需要后期优化网络，尽量减少训练时间。

针对 LINCS 数据的应用，分类只是其中的一小部分，第五章只是针对共享学习这样一种思维做了优化，如何把特征提取与稀疏表示分类相结合也是值得深入探究的。LINCS 数据的应用不仅仅局限于分类，还有降维、聚类分析、序列比对、基因网络的构建等都是值得研究的方向。

参考文献

[1] Brum A M, van de Peppel J, Nguyen L, et al. Using the connectivity map to discover compounds influencing human osteoblast differentiation. Journal of cellular physiology, 2018, 233(6): 4895-4906.

[2] Xiao S, Zhu X, Deng H, et al. Gene expression profiling coupled with connectivity map database mining reveals potential therapeutic drugs for Hirschsprung disease. Journal of pediatric surgery, 2018, 53(9): 1716-1721.

[3] Mathieu S, Manneville J B. Intracellular mechanics: connecting rheology and mechanotransduction. Current opinion in cell biology, 2019, 56: 34-44.

[4] 黄昕, 何松, 刘阳, 等. LINCS——面向转化医学的细胞反应大数据计划. 生物化学与生物物理进展, 2017, 11:91-95.

[5] Afgan E, Lonie A, Taylor J, et al. CloudLaunch: Discover and deploy cloud applications. Future Generation Computer Systems, 2019, 94: 802-810.

[6] 尹晓尧.LINCS 生物大数据解读与分析: [国防科技大学硕士论文]. 长沙: 国防科技大学生物系生物医学研究所. 2015. 12-16.

[7] Kelm J M, Lal-Nag M, Sittampalam G S, et al. Translational in vitro research: Integrating 3D drug discovery and development processes into the drug development pipeline. Drug discovery today, 2019, 24(1): 26-30.

[8] 向丽娟, 汪圣毅, 包楚阳, 等. 胃癌脂代谢通路基因表达的转录组学高通量分析. 安徽医科大学学报, 2019 (1): 2.

[9] Capuzzi S J, Thornton T E, Liu K, et al. Chemotext: a publicly available web server for mining drug–target–disease relationships in PubMed. Journal of chemical information and modeling, 2018, 58(2): 212-218.

[10] 宋欣雨, 文昱琦, 刘祯, 等. 基于 LINCS 转录组大数据的药物诱导基因共表达网络构建. 军事医学, 2018 (6): 8.

[11] Chen W, Zhou X. Drug Signature Detection Based on L1000 Genomic and Proteomic Big Data. Bioinformatics and Drug Discovery. Humana Press, New York, NY, 2019: 273-286.

[12] Otvos R A, Still K B M, Somsen G W, et al. Drug Discovery on Natural Products: From Ion Channels to nAChRs, from Nature to Libraries, from Analytics to Assays. SLAS DISCOVERY: Advancing Life Sciences R&D, 2019, 24(3): 362-385.

[13] 熊志勇, 吴珏堃, 梁豪. 基于癌症基因信息数据库分析载脂蛋白 F 在肿瘤中生物

学意义. 中华实验外科杂志, 2018, 35(3):570.

[14] 朱明敏, 刘三阳. 一种基于改进 Bhattacharyya 距离的高斯网络协方差矩阵灵敏度分析方法. 浙江大学学报 (理学版), 2019, 46(1): 9-14.

[15] Goudarzi Z, Adibi P, Grigat R R, et al. Making metric learning algorithms invariant to transformations using a projection metric on Grassmann manifolds. International Journal of Machine Learning and Cybernetics, 2019: 1-10.

[16] 詹增荣, 程丹. 基于 LDA 与距离度量学习的文本分类研究. 湖南师范大学自然科学学报, 2016, 39(5):70-76.

[17] Oh Song H, Xiang Y, Jegelka S, et al. Deep metric learning via lifted structured feature embedding.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 4004-4012.

[18] Nguyen B, Morell C, De Baets B. Supervised distance metric learning through maximization of the Jeffrey divergence. Pattern Recognition, 2017, 64: 215-225.

[19] Lei L, Fithian W. AdaPT: an interactive procedure for multiple testing with side information. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2018, 80(4): 649-679.

[20] http://support.lincscloud.org/hc/en-us/articles/202062163-L1000-Code-via-GitHub-

[21] https://amp.pharm.mssm.edu/L1000CDS2/#/index

[22] https://en.wikipedia.org/wiki/GitHub

[23] http://api.lincscloud.org/

[24] Duan Q, Reid S P, Clark N R, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. Npj Systems Biology & Applications, 2016, 2(1):16015.

[25] Wang Z, Clark N R, Ma'Ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics, 2016, 32(15):2338.

[26] https://pypi.org/project/h5py/

[27] Kuntz E M, Baquero P, Michie A M, et al. Targeting mitochondrial oxidative phosphorylation eradicates therapy-resistant chronic myeloid leukemia stem cells. Nature medicine, 2017, 23(10): 1234.

[28] Nakahara S, Medland S, Turner J A, et al. Polygenic risk score, genome-wide association, and gene set analyses of cognitive domain deficits in schizophrenia. Schizophrenia research, 2018, 201: 393-399.

[29] Zhang W, Bojorquez-Gomez A, Velez D O, et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nature genetics, 2018, 50(4): 613.

[30] Reddy N H, Kumar E R, Reddy M V, et al. Bioinformatics and Image Processing—Detection of Plant Diseases.First International Conference on Artificial Intelligence and Cognitive Computing. Springer, Singapore, 2019: 149-154.

[31] 曾玮, 刘孟刚, 刘宏鸣,等. 基因表达谱结合共表达网络分析发现胰腺导管腺癌预后分子机制. 肿瘤研究与临床, 2014, 26(9):583-586.

[32] Wang X Y, Hua G, Han T X. Discriminative tracking by metric learning. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer, 2010, 200-214.

[33] Chen J H, Zhao Z, Ye J P, Liu H. Nonlinear adaptive distance metric learning for clustering. In: Proceedings of the 2007 International Conference on Knowledge Discovery and Data Mining. California, USA: ACM, 2007, 123-132.

[34] 仰迪, 白延琴, 李倩. 半监督距离度量学习的内蕴加速投影梯度算法. 运筹学学报, 2018, 22(2):66-78.

[35] 郑宝芬, 苏宏业, 罗林. 无监督特征选择在时间序列数据挖掘中的应用. 仪器仪表学报, 2014, 35(4):834-840.

[36] Wang B, Jiang J Y, Wang W, Zhou Z H, Tu Z W. Unsupervised metric fusion by cross diusion. In: Proceedings of the 2012 Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012, 2997-3004.

[37] Mignon A, Jurie F. CMML: a new metric learning approach for cross modal matching. In: Proceedings of the 11th Asian Conference on Computer Vision. Daejeon, Korea: Springer, 2012, 14-27.

[38] Cao B, Ni X C, Sun J T, Wang G, Yang Q. Distance metric learning under covariate shift. In: Proceedings of the 22nd International Joint Conference on Articial Intelligence. Barcelona, Spain: AAAI, 2011, 1204-1210.

[39] Guillaumin G, Verbeek J, Schmid C. Multiple instance metric learning from automatically labeled bags of faces. In:Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer, 2010, 634-647.

[40] 酆勇, 熊庆宇, 石为人,等. 深度非线性度量学习在说话人确认中的应用. 声学学报, 2018(1):112-120.

[41] Yang L, Jin R, Sukthankar R. Bayesian active distance metric learning. In: Proceedings of the 23th Conference on Uncertainty in Articial Intelligence. Vancouver, Canada:AUAI Press, 2007, 442-449.

[42] Yang L, Jin R. Distance metric learning: A comprehensive survey. Michigan State Universiy, 2006, 2(2): 4.

[43] 仰迪, 白延琴, 李倩. 半监督距离度量学习的内蕴加速投影梯度算法. 运筹学学

报, 2018, 22(2):66-78.

[44] 徐晓祥, 李凡长, 张莉,等. 范畴表示机器学习算法. 计算机研究与发展, 2017, 54(11):2567-2575.

[45] Xing E P, Jordan M I, Russell S J, et al. Distance metric learning with application to clustering with side-information.Advances in neural information processing systems. 2003, 521-528.

[46] Shental N, Hertz T, Weinshall D, et al. Adjustment learning and relevant component analysis.European conference on computer vision. Springer, Berlin, Heidelberg, 2002, 776-790.

[47] Schultz M, Joachims T. Learning a distance metric from relative comparisons. Advances in neural information processing systems. 2004, 41-48.

[48] Goldberger J, Hinton G E, Roweis S T, et al. Neighbourhood components analysis. Advances in neural information processing systems. 2005: 513-520.

[49] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 2009, 10(Feb): 207-244.

[50] Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning. Proceedings of the 24th international conference on Machine learning. ACM, 2007: 209-216.

[51] Jin R, Wang S, Zhou Y. Regularized distance metric learning: Theory and algorithm. Advances in neural information processing systems. 2009: 862-870.

[52] Yang P P, Huang K Z, Liu C L. A multi-task framework for metric learning with common subspace. Neural Computing and Applications, 2013, 22(7-8): 1337-1347

[53] Yang P P, Huang K Z, Liu C. Geometry preserving multi-task metric learning. Machine Learning, 2013, 92(1):133-175.

[54] Jin R,Wang S J, Zhou Y. Regularized distance metric learning: theory and algorithm. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems.Vancouver, Canada: MIT Press, 2009, 862-870

[55] Hoi S C H, Liu W, Chang S F. Semi-supervised distance metric learning for collaborative image retrieval. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Alaska, USA: IEEE, 2008, 1-7.

[56] Shen C H, Kim J, Wang L. Scalable large-margin Mahalanobis distance metric learning. IEEE Transactions on Neural Networks, 2010, 21(9): 1524-1530

[57] Shen C H, Kim J, Wang L. A scalable dual approach to Semide nite metric learning. In: Proceedings of the 24th Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2011, 2601-2608.

[58] Huang K Z, Ying Y M, Campbell C. GSML: a unied frame-work for sparse metric learning. In: Proceedings of the 9th International Conference on Data Mining. Florida, USA:IEEE, 2009, 189-198

[59] Huang K Z, Ying Y M, Campbell C. Generalized sparse metric learning with relative comparisons. Knowledge and Information Systems, 2011, 28(1): 25-45

[60] Liu W, Hoi S C H, Liu J Z. Output regularized metric learning with side information. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France:Springer, 2008, 358-371.

[61] Weinberger K Q, Blitzer J, Saul L K. Distance metric learning for large margin nearest neighbor classification.Advances in neural information processing systems, 2005, 1473-1480.

[62] Cheng G, Han J, Zhou P, et al. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Transactions on Image Processing, 2019, 28(1): 265-278.

[63] Sugiyama MLocal Fisher discriminant analysis for supervised dimensionality reduction.Proceedings of the 23rd Intenation-al Conference on Machine Leaming, Pittsburgh, Pennsylvania, 2006, 905-912.

[64] 孙渡, 李永强, 吴珍珍,等. 基于多回波及Fisher判别的陡坡点云滤波研究. 地理与地理信息科学, 2018, 34(2).

[65] 杜树新, 吴铁军. 模式识别中的支持向量机方法. 浙江大学学报(工学版), 2003, 37(5):521-527.

[66] Davis JV, Kulis B, Jain P, Sra S, Dhillon I S. Information-theoretic metric learning. In:Proceedings of the 24th Inter-national Conference.Oregon, USA: ACM, 2007, 209-216.

[67] Kostinger M, Hirzer M, Wohlhart P, Roth P M, Bischof H. Large scale metric learning from equivalence constraints.In: Proceedings of the 2012 Computer Vision and Pattern Recognition. Providence. RI: IEEE.2012,2288-2295.

[68] Yang Z, Hu X, Dai F, et al. Person re-identification by discriminant analytical least squares metric learning. Machine Vision and Applications, 2018, 1-13.

[69] Y 丁勇, 叶大炜, 袁方,等. 根本原因分析法(RCA)在医疗不良事件分析中的应用. 中国医院, 2015(5):41-43.

[70] Qi G , Tang J, Zha Z J, et al. An efficient sparse metric learning in high-dimensional space via l 1-penalized log-determinant regularization.Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, 841-848.

[71] Alvarez G, Li S. Cryptanalyzing a nonlinear chaotic algorithm (NCA) for image

encryption. Communications in Nonlinear Science & Numerical Simulation, 2009, 14(11):3743-3749.

[72] Deming T J. Polypeptide and Polypeptide Hybrid Copolymer Synthesis via NCA Polymerization. Cheminform, 2010, 38(5).

[73] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.

[74] 赵志宏，杨绍普，马增强. 基于卷积神经网络 LeNet-5 的车牌字符识别研究. 系统仿真学报，2010, 03：638-641.

[75] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 4700-4708.

[76] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking. European conference on computer vision. Springer, Cham, 2016, 850-865.

[77] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349(6245): 255-260.

[78] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展. 软件学报, 2006, 17(9):1848-1859.

[79] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述. 计算机应用, 2016, 36(9):2508-2515.

[80] Lecun Y, Muller U, Ben J, et al. Off-road obstacle avoidance through end-to-end learning. International Conference on Neural Information Processing Systems. 2005.

[81] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking. 2016, 850-865.

[82] Rama Varior R, Haloi M, Wang G. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. Computer Vision – ECCV 2016. Springer International Publishing, 2016, 791-808.

[83] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 4700-4708.

[84] Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition, 2019, 90: 119-133.

[85] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition. European conference on computer vision. Springer, Cham, 2016, 499-515.

[86] PATEL V M, CHELLAPPA R. Sparse Representations, Compres-sive Sensing and

dictionaries for pattern recognition.Asian Conference on Pattern Recognition. Beijing: IEEE,2011,325-329.

[87] Donoho D L. For most large underdetermined systems of linear equations the minimal 11-norm solution is also the sparsest solution . Communications on pure and applied mathematics,2006,59(6):797-829.

[88] Candes E J, Tao T. Near-optimal signal recovery from random projections: Universal encoding strategies. IEEE Transactions on Information Theory,2006,52(12): 5406 -5425.

[89] QIU Q, JIANG Z, CHELLAPPA R. Sparse dictionary-based rep-resentation and recognition of action attributes.IEEE Inter-national Conference on Computer Vision. Barcelona, Spain:IEEE, 2011, 707-714.

[90] 张新鹏, 王朔中. 基于稀疏表示的密写编码. 电子学报, 2007, 35(10):1892-1896.

[91] MAIRALJ, BACH F, PONCE J. Task-driven dictionary leaming . IEEE Transactions on Pattem Analysis & Machine Intelli-gence,2012,34(4):791-804.

[92] ENGANK, AASES O, HUSOY J H. Method of optimal directions for frame design. IEEE International Conference on Acous-tics, Speech, and Signal Procesing, vol.5. Barcelona, Spain:IEEE, 1999, 2443-2446.

[93] 孙利雷, 秦进. 稀疏字典学习方法综述. 贵州大学学报(自然科学版), 2018, 35(05):87-92.

[94] 练秋生, 石保顺, 陈书贞. 字典学习模型、算法及其应用研究进展. 自动化学报, 2015, 41(2):240-260.

[95] 朱杰, 杨万扣, 唐振民. 基于字典学习的核稀疏表示人脸识别方法. 模式识别与人工智能, 2012, 25(5):859-864.

[96] Scholkopf B, Platt J, Hofmann T. Sparse representation for signal classification.In: Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December. 2006, 609-616.

[97] 罗涛, 冯玉田, 唐子成,等. 基于加权稀疏表示分类的车辆识别. 电子测量技术, 2018(6).Processing.2010,p.1601-1604.

[98] 那天, 宋晓宁, 於东军. 基于主元分析和线性判别分析降维的稀疏表示分类. 南京理工大学学报, 2018, 42(03):32-37.

[99] Kong S, Wang D. A dictionary learning approach for classification: separating the particularity and the commonality. European conference on computer vision. Springer, Berlin, Heidelberg, 2012, 186-199.

[100] Ramirez I, Sprechmann P, Sapiro G. Classification and clustering via dictionary

learning with structured incoherence and shared features. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, 3501-3508.

[101] 蔡泽民, 赖剑煌. 一种基于超完备字典学习的图像去噪方法. 电子学报, 2009, 37(2):347-350.

[102] Liang S, Wang Y, Liu Y. Face recognition algorithm based on compressive sensing and SRC. 2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control. IEEE, 2012, 1460-1463.

[103] 练秋生, 王小娜, 石保顺,等. 基于多重解析字典学习和观测矩阵优化的压缩感知. 计算机学报, 2015, 38(6):1162-1171.

[104] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(2): 210-227.

[105] 余发军, 周凤星, 严保康. 基于字典学习的轴承早期故障稀疏特征提取. 振动与冲击, 2016, 35(6):181-186.

[106] Feng Z, Yang M, Zhang I et al. Joint discriminative dimensionality reduction and dictionary learning for face recognition, Pattern Recognition, 2013, 46:2134-2143.

致 谢

不经意之间，为期两年的研究生生涯即将接近尾声，对于即将重新踏入社会面对各种可能和未知的机遇与挑战的我来说，真的是感慨万千，有期待，有失落，有迷茫，有感动，然而更多的是存在于心头的那抹温暖和感恩之情。回头望去，已经发现在不知不觉中收获颇多，于学业，于技术，于思维，于成长。在此，我要感谢所有指导和帮助过我的人。

由衷感谢我的导师彭绍亮教授。彭老师博学而睿智，待人亲和而富有感染力，他超脱的气质和精神一直在影响和促进周围的人尤其是我。在学术上，他耐心教导，启发思考，提供了许许多多的开阔眼界增长见识紧扣学术前沿的平台和机会，真切而又无私地带领着我们前行。在生活上，更是像一个家人的角色在关注学生的生活问题，急学生之所需，任何困难都能不辞辛劳地帮助学生解决。我很庆幸能够遇到彭老师这样的人生好导师，在给了我关怀和温暖的同时时刻促进我的成长，在此我要真挚的感谢彭老师！

衷心地感谢军事医学研究院的李非老师，李非老师温文儒雅，淡泊而善师，从大到小，由浅入深地指导我课题的研究。当我有不成熟想法而不知道是否可行的时候，李非老师不会立马否定来打击我的自信心，而是会从我的想法出发，引导我去思考其可行性。很多时候，李非老师晚上十一、二点下班之后还会不知疲倦的尽量先指导我们的课题研究，这让我非常感动，谢谢李非老师，您辛苦了！

感谢那些被我不断骚扰的师兄们，包括国防科大的崔英博师兄、张志强师兄，北京林大的洪浩师兄、实验室的杨亚宁师兄、还有学院的刘清星师兄、徐永宝师兄，师兄们谦逊而刻苦，理论知识和代码能力异常扎实，而当我遇到解决不了技术问题，总是去骚扰他们，他们很多时候也放下手中的事情来支援我的研究。

感谢实验室的同门孙哲、程敏霞、胡星和实验室的各位学弟学妹们，很多时候大家一起分工完成工作，一起讨论问题、汇报工作、一起打球和运动。

感谢研究生小分队的小伙伴们施程、李娜、王江干、杨韵霞、李海舟，大家一起去爬山、吃饭、看烟花（虽然我没去过），总是能帮我走出低落的情绪。正如娜姐说的那样，没有一顿美食解决不了的问题，如果有，那就两顿。

感谢三一圆梦班的小飞侠，认识 10 年来，仍然保持着这段来之不易的友情，作为好哥们，很多时候不需要多言就能帮忙解决我的困境。

感谢我的父母、我的姐姐，正是因为你们在背后的支撑，才能让我没有后顾之忧，能够任何时候分担我的痛苦悲伤，能让我重新燃起奋斗的信念。

最后特别感谢各位教授和专家百忙之中抽空评阅我的论文，参与答辩，谢谢！

刘伟　2019 年 4 月于超算中心 0 号楼

附录 A 攻读学位期间所发表的学术论文

学术论文：

[1] Shaoliang Peng,**Wei Liu**, Yaning Yang, Fei Li,Hao Hong，Kenli Li, Shulin Wang. A Deep Metric Learning Algorithm for Similarity Measure of Gene Expression Profile. Nucleic Acids Research (NAR) （SCI 一区，TOP 期刊，IF: 11.561,已投稿）.

[2] Shaoliang Peng,**Wei Liu**, Yaning Yang, Fei Li, Kenli Li, Xiangke Liao. Discriminant Projection Shared Dictionary Learning for Classification of Tumors Using Gene Expression Data. IEEE-ACM Transactions on Computational Biology and Bioinformatics(TCBB)（CCF-B 类 SCI 期刊，IF:2.428 ,已投稿）.

发明专利：

[3] 一种基于深度学习的基因表达谱距离度量方法.(申请号：201910296276.1)

[4] 一种基于共享字典学习的基因表达谱分类方法.(申请号：201910296287.X)

软件著作权：

[5] LINCS 生物大数据相似性度量分类系统.(登记号：2019SR0159294)

附录 B 攻读学位期间参与的研究项目

[1] 国家自然科学基金项目：《面向大规模异构体系结构的生物医药大数据并行算法及优化关键技术研究》，编号：61772543。

[2] 国家重点研发计划项目：《精准医学大数据的有效挖掘与关键信息技术研发》，编号：2018YFC0910405。

[3] 深圳市科技计划项目：《面向生物大数据药物重定位的小样本机器学习方法》，编号：JCYJ20170818110101726。