硕士学位论文

基于特征提取和特征选择的级联深度学习模型研究

STUDIES ON CASCADED DEEP LEARNING MODEL BASED ON FEATURE EXTRACTION AND FEATURE SELECTION

王维智

哈尔滨工业大学

2015 年 6 月

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：612.3　　密级：公开

工学硕士学位论文

基于特征提取和特征选择的级联深度学习模型研究

硕士研究生：王维智
导　　师：张大鹏教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2015 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.4
U.D.C: 612.3

Dissertation for the Master Degree in Engineering

STUDIES ON CASCADED DEEP LEARNING MODEL BASED ON FEATURE EXTRACTION AND FEATURE SELECTION

Candidate: Weizhi Wang
Supervisor: Prof. David Zhang
Academic Degree Applied for: Master of Engineering
Speciality: Computer Science and Technology
Affiliation: School of Computer Science and Technology
Date of Defence: June, 2015
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

特征学习是机器学习中一个重要研究方向，好的特征可以提供数据的语义和结构信息，使简单的任务模型也能取得良好的学习效果。区别于使用浅层的观测特征和人工设计特征的方法，深度学习是一种可以从数据中自动学习特征的方法，其所学习到的特征往往具有较高的抽象性和语义性。通过逐层变换，深度模型可以在不同层抽象出数据的不同表示，从而很好的适应于机器学习的各种任务中。现有的深度学习方法大多采用特征变换、非线性操作和特征选择（约简）的多层迭代模型，所解决的问题也多集中在具有二维空间结构意义的领域，而对于一般的向量特征却尚无有效的应对方法。

为解决普通向量特征的学习问题，本文构建起一个通用有效的级联深度模型。为此本文首先对模型中用到的特征选择方法进行了研究， 提出了基于 $L_{2,p}$ 范数约束的非凸正则化特征选择模型，给出了非凸问题的求解算法。由于发掘并利用了样本特征空间内相互表示的特性，该方法可以有效的进行特征选择的任务。

其次，通过组合一种通用的特征变换方法和所提出的特征选择方法，本文实现了一个多层的级联深度特征学习模型，并通过分类结果考察其学习性能。为利用模型各层学习得到的特征，本文提出了合理有效的特征组合方式，充分地提取了不同层特征间所包含的互补信息，从而显著提升了模型的分类性能。

为探究有效训练数据数量的增加是否可以提升深度模型学习性能的问题，本文继而提出了一种基于手写体数字图像的数据扩展方式，并研究分析了扩展参数对模型性能的影响。实验结果表明数据增强可以显著提高模型分类结果。

关键词：特征提取；特征变换；深度学习；级联模型

Abstract

Feature learning is an important branch in machine learning, via which ‘good’ features can be extracted to provide semantic and structural information of data, leading to better results even for simple classification models. Instead of extracting features using raw observations or hand-craft methods, deep learning aims to extract features automatically, which can commonly provide high-level semantic and structural information. Moreover, with the help of multi-layer structure, deep learning models can abstract the data from different scales, meeting the needs for kinds of machine learning tasks. Currently, all the deep learning methods grow up based on the multi-layer scheme, which are roughly composed of three steps, i.e., feature transformation, non-linear operation and feature selection. However, most of the existing deep learning methods only concentrate on the two dimensional data with spatial structural information, leaving the general vector-form features unresolved.

In this work, we propose a generalized cascade deep learning model to solve the feature learning problem with the general vector-form. To achieve this, firstly we propose a general unsupervised feature selection model, where $L_{2,p}$ norm is introduced to investigate the self-representation property of feature space more deeply together with its non-convex solutions, thus leading to an effective feature selection model.

By combining generalized feature transformation methods with the proposed feature selection method, a cascade deep learning model for feature learning is finally obtained in our work. To fully utilize the features learned from each layer of the deep learning model, we propose a reasonable feature combination strategy, which adequately exploits the complementary information from different layers, thus improving the classification performance significantly.

As to the training of the proposed cascade deep model, we propose a data augmentation strategy on the handwritten digitals dataset for better performance. During training the deep model, we study the performance of the proposed model with different data augmentation parameters. Experimental results have shown that the proposed data augmentation strategy boosts the performance notably.

Keywords: feature extraction, feature transformation, deep learning, cascade model

目 录

第1章 绪 论

1.1 课题背景及研究的目的和意义

人工智能（Artificial Intelligence，AI）是计算机领域的一个重要研究方向，其愿景就是更好地理解我们身边的世界[1]。基于统计的机器学习是人工智能的重要组成部分。在机器学习中，面对不同的学习任务，包括数据挖掘、计算机视觉、文本分类、基因工程等，一个重要的难点就是如何提取和选择合适的特征来处理这些问题[2]。 例如：在基于上下文语境的两类问题的监督学习中，经常需要学习一个可以区分两类问题的超平面。原始的特征并不总能满足这样的要求，事实上大部分情况都不能直接从原始的特征空间学习到这样的超平面，因此，寻找原始特征的映射空间，使得数据在新的特征空间内线性可分的任务就变的十分重要。

映射空间的求解一般有几种方法：（1）人工设置特征生成方法；（2）设置合适的核函数，使用隐式的内积空间表示特征，这也被简单的称作从数据中自动学习。这两种方法中，第一种方法需要研究者根据数据或问题领域自己设计特征，也已经有一些针对特定领域的优秀特征设计方法，如图像领域里的 Textons、SIFT、HOG 特征等。这些优秀的人工设计特征具有很好的特性，如对尺度、大小和旋转等变换的不变性适应，同时对不同类别的数据又有良好的区分性。但是这种特征设计方法有着一个很大的缺点：设计过程中需要专业的先验知识，整个过程需要耗费巨大的人力和物力，尤其当原始数据是分布在高维空间的时候（如图像数据、基因数据），这一任务更显得难以完成。与此同时，人工设计的特征也充满着更多的随机性，大量的人工设计特征的有效性难以得到保障。由于可以减少对人力的依赖，并且能较好地提取出相关的特征，因此自动学习特征的算法使得人工智能的愿景更令人期待。而核函数特征空间映射在解决非线性问题中也确实取得了一定的效果。

普通的核函数可以看作是简单的浅层特征映射方式，随着最近研究的进展，第三种方法——一种从深度结构中自动学习特征的方法，被认为是能够学习到数据更加抽象的本质特征。受人类视觉系统的信息处理机制[3]的启发，Hinton 等人于 1986 年提出了基于反向传播（Back-Propagation，BP）算法的人工神经网络[4]，其主要思想是使用了特征的分层抽象思想和感知器模型。深度学习模型作为一种组合了非线性处理的多层结构，对于一些高层次的非线性信息，可以比浅层的结构（如 SVM）有着更接近真实的表示，与此同时，深度结构还可能有着更简洁的参数形式。例如对含有 n 个比特位输入的奇偶校验函数，使用前馈神经网络进行编码时，网络的代价是 $O(\log n)$个隐含层和 $O(n)$个神经元，然而如果前馈神经网络的隐含层只有一层时，就需要 n 的指数次方个神经元来达成相同的任务[5]。此外，在

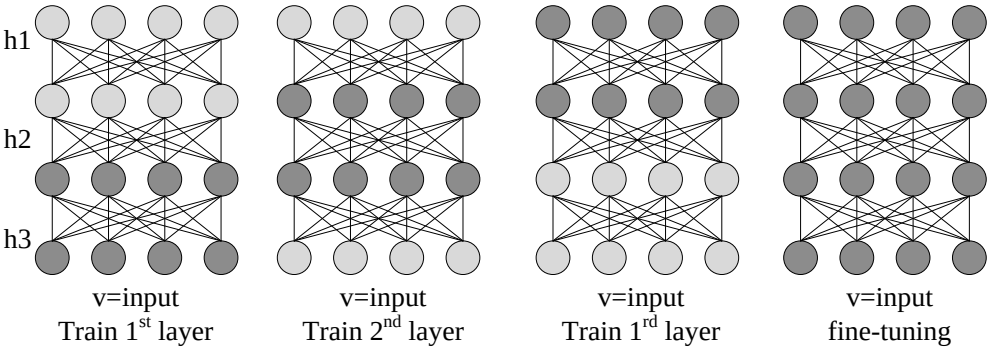


图 1-1 基于预训练的非监督分层优化深度结构训练模型[1]

不同的函数下，完全基于局部泛化的算法会不可避免的造成维数灾难[6]。对于这一难题，深度结构也可以通过使用分布式表示和构建更合适的特征选择等方法进行解决。

深度学习的能力与优点让人憧憬，然而，深度框架的训练是一项艰巨的任务，而传统的、已经被证明有效的浅层训练方法在移植到深度学习中又不能确保其有效性（事实是基本都无效）。另一个现实的问题是增加层结构与获取更好的特征表示也不存在必然的联系。例如，在神经网络中，隐含层越多，反向传播算法中第一层所受的影响就越小，使用梯度下降算法时也会陷入局部最优而失去继续传递的效果[7]，这也是为何研究者们之前总是只设计一层或两层神经网络的原因。

深度结构的训练难题在 2006 年取得突破性进展，多伦多大学的 Hinton 教授介绍了一种基于预训练的非监督的分层优化深度结构的训练方法[8]。在该方法中，深度结构的每一层都通过贪心算法成功的进行训练，前一层训练所得到的输出结果作为下一层的训练的输入数据，这样一直逐层完成特征的训练，在最后一层通过使用有监督的策略对整个网络进行反向微调，这一过程简单描述为图 1-1。接着，卷积神经网络（Convolutional Neural Networks, CNN）[9]与自动编码模型[10]作为另外两种深度学习模型的成功实现也引起广泛的关注，并一起带动了深度研究的高潮。与此同时，随着研究的深入与对于深度理论认识的加深，研究者们也开始关注更简单的深度学习模型。DeepBoosting 通过级联 Gabor 滤波和 Boosting 算法构建了图像分类的深度模型[11]，DeepFisher 通过级联 Fisher 特征构建了图像的分类模型[12]。

典型的深度学习模型都是基于“特征变换——非线性操作——特征选择（约简）”的多层迭代模型，特征变换通过设计滤波器或是其他特征提取方法，成功的提取当前阶段的特征信息，同时也升高了数据的维数；非线性操作模拟了人类神经元具有激活与抑制两个状态，将变换后的特征二值化或使用逻辑回归函数处理；特征选择将对分类或其他机器学习任务起作用的特征进行挑选，同时达到降维的

作用，使得深度网络模型的规模维持在一定范围内。模型中，特征变换和特征选择的合理设计对于深度模型的学习能力至关重要。以 CNN 为例，CNN 是一个包括了卷积滤波层、非线性变换、池化的过程的迭代网络模型。其中卷积滤波层使用线性滤波器对图像进行卷积操作，经过滤波操作后，可有效提取出图像数据中的边缘等信息，有着良好的特征变换作用；非线性变换一般使用 sigmoid 函数，对变换后的特征进行增强和区分；池化操作通常采用最大池化或平均池化，不论最大池化或是平均池化，都针对了图像数据局部区域信息的不变性，这是由于图像是二维空间的数据，同类数据间存在平移旋转等不变性，而池化操作的设计正是基于此特点，提取图像局部区域内的不变性特征，同时也对特征数进行了降维。通过迭代的进行这些操作，图像数据的特征被逐层提取出来，从像素到边缘，从边缘到局部结构再到整体结构，随着层数的增高，特征的语义信息越来越抽象，越来越接近认知领域。

CNN 的成功，离不开滤波的思想和池化的作用，但是其只能针对在二维空间内含有语义信息的任务进行建模，对于没有二维空间语义信息的数据集合，卷积神经网络就无能为力。那么是否像 CNN 可以有效提取图像信息一样，深度模型也可以对一般的向量数据有较强的特征学习能力？是否可以构造一种通用的深度学习模型对普通向量数据进行特征学习？基于此出发点，我们构建了一个简单的泛化能力更强的深度特征学习框架，达到对普通向量数据进行深度特征学习的任务，从而获取数据中的更有效的信息。我们首先对框架中所使用的特征变换和特征选择相关算法进行具体的研究和分析，提出一种有效的特征选择算法，最终实现一个有效的级联深度学习模型，并通过分类准确率来衡量模型的学习性能。由于我们提出的模型不使用到数据的任何先验信息，因此，对于具体数据分类任务，我们可以根据先验知识对训练数据进行数据增强，基于此出发点，我们也提出一种有效的手写体数字图像数据的增强方式。由于本文所实现的不仅是一个深度模型，而且提供了一个深度学习的框架，因此研究者们也可以通过针对不同领域的问题对本文所提出的模型进行改进，从而扩展其应用范围。

1.2 国内外研究现状

近年来，国际和国内的一些优秀学者们投身到深度学习中来，并推动着这一领域的发展。常见的深度学习模型训练方式有两种：有反馈调节的深度学习模型和无反馈的级联深度模型。包括反馈调节的深度模型如卷积神经网络、深度信念网络和自动编码表示等，无反馈的深度模型有 DeepBoosting[11]、DeepFisher[12]和 DeepPCA[13]等。这两种深度学习模型都是特征提取和特征选择的有规律组合，因此，如何设置特征变换和特征选择方法及其组合方式是在前人基础上进行研究的重要环节。

1.2.1 深度学习模型

作为有反馈且是最成功的深度模型之一，卷积神经网络是一个主要用来进行 2 维信号（如图像，语音信号等）特征学习和分类的多层神经网络模型。具体的卷积神经网络包括两个主要部分：多个用于特征学习的卷积层和一个用于分类的全连接层。卷积层包括三个步骤，卷积变换，非线性变换和池化操作。其中卷积变换是一个线性变换，定义一个固定大小的模板，模板值随机初始化，通过对 2 维信号进行卷积操作，完成信号的线性变换，这是特征的一个映射，定义多个模板，分别对 2 维信号进行卷积，可以得到多个不同的特征变换。对变换后的特征，使用 Sigmoid 函数进行非线性变换，抽取信号的非线性信息。接着再对信号进行池化，常用的池化方法包括 Max-pooling 或是 Mean-pooling，其中 Max-pooling 可以看作是一个特征选择的过程，Mean-pooling 是一种维数约简方式，模型中的这两种方法都可以有效的针对 2 维变换信号中的局部不变性。多个卷积层中，每层的输出是下一层的输入。最后一个卷积层生成的特征连接到全连接层，此处全连接层同普通神经网络的全连接层相同。在训练中，分为前向训练和后向反馈。前向训练中，卷积模板和全连接层权重随机初始化，然后从低层到高层，每层中使用当前的卷积模板和 Sigmoid 函数及下采样规则进行特征的抽象，多层卷积结束后将最后一层所得的抽象特征输入到全连接层并按连接权重进行操作。后向反馈中，从高层到低层，通过求解最小二乘误差，使用梯度下降法逐层更新全连接层权重和卷积层的卷积模板值。如此重复直至收敛，从而完成模型的训练。与普通神经网络相比，卷积神经网络在卷积层中使用了局部连接和权值共享的机制代替普通神经网络中全链接操作，同时，非线性变换后还加入了下采样操作。这样做的好处是不仅可以有效的提取样本变换时一些抽象特征的不变性，同时也大大减少了训练模型中的参数。卷积神经网络在图像处理和语音方面都取得了骄人的成果。

与卷积神经网络略有不同，另一种深度模型没有后向反馈的步骤。林倞等人提出了一种基于 Boosting 的逐层特征挖掘的深度模型来对一般图像进行分类[11]。该模型是一个没有反馈调节的多层学习模型。通过逐层的特征学习，直接将最后一层特征输入到分类器进行分类。其中每层特征学习模型中都包含了以下几个步骤：Gabor 滤波，Boosting 特征选择，近邻特征组合。其中 Boosting 特征选择的同时对特征进行了 Sigmoid 非线性操作。在这三步中，Gabor 滤波对图像的特征提取有着较好的作用；而通过使用基于桩函数的 Boosting 算法，可以选择对于分类有效的特征，同时可给出特征在分类过程中的置信度；近邻特征组合，通过规定近邻区域，使用 Boosting 挑选出的特征进行加权线性组合，权值为区域内特征的置信度。通过这种方法，逐层进行特征的变换和选择，最后将变换后的特征使用 Boosting 分类器进行分类。模型最后在真实数据库上取得了良好的性能。

上述两种模型的成功固然与深度结构的使用有很大关系，然而，是否随便搭建一个多层的结构就可以学习到数据的抽象表示，并能很好的完成相应的机器学习任务？我们是否应该考虑结构中所使用到的特征学习方法？基于这种考虑，需要对特征提取和特征选择方法进行了研究与分析。

1.2.2 特征提取

特征提取是从数据特征中获取信息的过程。通过对特征进行变换获取数据的信息。最简单特征变换方式有 PCA 变换、线性判别分析[14]等，PCA 变换是寻找数据的主轴方向，使数据投影后用尽可能少的特征表示尽可能多的信息；线性判别分析是通过线性投影，使得新空间下的数据具有更大的类间距离和更小的类内距离；Gabor 滤波[15]变换是图像领域中一种重要的变换方式，通过设置滤波器中高斯核函数的不同的参数，可以有效的提取图像中的边缘信息；通过在多尺度空间采用 DOG 算子检测关键点，SIFT 特征提取图像的局部信息[16]，使其对旋转、尺度缩放、亮度变换保持着很好的不变性；通过将整幅图像分成若干个连接区域，并统计每个区域的直方图，HOG 特征[17]可以有效的应对图像的旋转和尺度不变性。

1.2.3 特征选择

特征选择是一种从数据特征中挑选出对任务有帮助的特征子集的方法。常见的特征选择方法可以被分为 3 类：Filter[18]、Wrapper[19]和 Embedded 方法[20]。Filer 方法是根据每个特征所对对应的特征估值指标大小逐一选择特征，这些指标常常是数据的一些统计特性。Wrapper 方法是根据选择出的特征子空间进行分类时的分类准确率来进行特征选择，由于特征子空间不确定，因此 Wrapper 方法需要进行多次训练才能给出选择的特征空间。Embedded 方法首先确定了使用特征的模型，然后从特征空间中搜索可以提高模型性能的特征子空间。从样本是否含有类别信息，特征选择又可以被分为监督学习和非监督学习。早期的非监督学习首先定义一些度量特性，然后逐一计算特征的度量特性，并按顺序选取，这些度量特性可能是聚类效果、信息冗余度等，其中一些代表性的度量包括 Laplacian Score、Trace ratio 等。然而这种依靠搜索的特征提取方法需要巨大的计算量，因此研究人员开始考虑不再需要搜索空间的聚类算法，基于特征的相似特性，一系列谱聚类算法被提出。基于稀疏表示，L_1 正则化和 $L_{2,1}$ 组稀疏的方法被广泛应用，L_1-SVM[21]和 $L_{2,1}$ 正则化组稀疏[22]都被用来进行特征选择。

通过合理的组合了合适的特征提取和特征选择方法，深度学习已经在各领域取得了一系列的成果。卷积神经网络的成功使用使得语音识别领域取得突破性进展[23-28]，同样卷积神经网络在目标识别领域也达到了当前最好的效果[29]。自然语言处理领域，深度卷积网络也被用来训练成语言模型[30, 31]，在工业界，深度学习

也显示了其优良的性能。2012 年，刊登在《纽约时报》上的谷歌大脑项目构建的深度神经网络从海量的 Youtube 数据中自动识别出了猫脸；2012 年，微软在天津的一次发布会上演示了全自动的同声传译系统，包括英文语音的识别，英中翻译和中文合成，效果惊人，该系统的核心算法就是卷积神经网络；2013 年 1 月，百度成立了深度学习研究院，致力于研究深度学习，并将其运用到图片、语音、自然语言的搜索中。

1.3 本文的主要研究内容

本课题研究内容主要来源于国家自然科学基金研究项目“舌脉合参中信号和特征的约简与协同分析方法研究”（批准号码：61271093）资助。本文将构建一种简单的可泛化的级联深度学习框架。通过对框架中使用到的特征提取和特征选择方法进行研究和分析，实现了一种有效的级联深度特征学习模型。为了更好的使用模型学习到的特征，提出了一种有效的组合各层特征的方法。在实现深度学习模型前，我们首先对模型中所使用到的特征变换方法和分类器方法进行了介绍，并对模型中使用到的基础算法——特征选择算法进行了深入研究。然后通过组合特征变换和特征选择算法，成功实现了一个用于分类的级联深度学习模型。由于深度学习的效果依赖于有效训练数据量的大小，而通过数据的先验知识我们可以对数据进行有效的增强与扩充，因此实现深度学习模型后，本文还提出了一种基于手写体数字图像的数据扩展方式，并通过实验分析了扩展参数对模型性能的影响，给出了一种有效的扩展方式和对应的参数选择。论文的主要框架如图 1-2 所示。

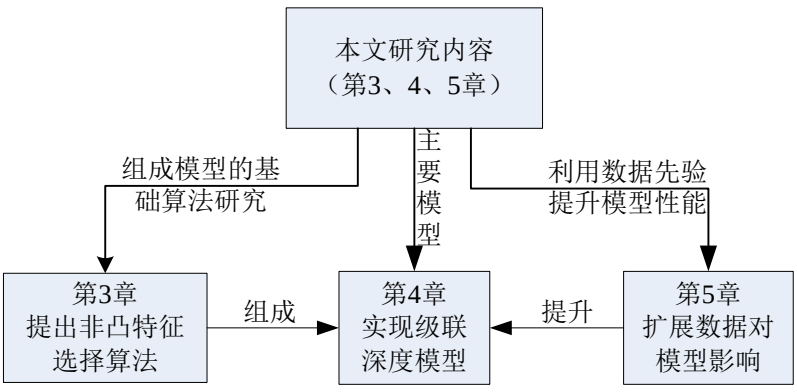


图 1-2 本文主要研究内容框架

基于以上研究内容，本文的具体章节安排如下：

第 2 章介绍了模型中会使用到的基础算法和前人的研究工作，包括主成分分析降维，基于消失成分分析的特征变换算法以及多分类器融合的 Boosting 算法。

第 3 章研究深度学习中的一个基础工作——特征选择。提出了一种基于特征

自表示的非凸非监督特征选择模型，给出了求解该非凸模型的迭代算法，并严格证明了算法在解空间的收敛性。最后通过详细的实验对比验证该特征选择模型的有效性。

第 4 章构建了级联深度学习框架。通过对框架中用到的特征变换、特征选择和分类方法进行详细的分析，成功实现了一个基于特征提取和特征选择的有效的级联深度模型。为充分利用模型学习到的各层特征，本文还提出了一种合理有效的特征组合方式，并最终通过分类实验验证模型在学习特征领域的有效性。

第 5 章研究了扩展样本对模型特征学习性能的影响。首先提出一种手写体数字图像数据的扩展方法，通过设置不同的扩展函数尺度及采样方法，着重分析了扩展函数的参数对模型性能的影响，并通过合适的参数对数据增强，成功提升了模型的性能。

第2章 基础算法介绍

2.1 引言

本文主要构建了一种泛化能力较强的级联深度学习模型。模型中我们不仅使用了自己研究的方法，也使用其他一些具有优秀性能的算法，这里统一对他们进行介绍，包括各阶段所使用的主成分分析[32]降维、基于消失成分分析[33]的特征变换以及 Boosting 分类器算法[34]。

2.2 主成分分析

主成分分析（Principal component analysis，PCA）是一种线性空间变换，目的是通过空间变换，使用较少特征却可尽量多保留数据信息的条件下表示原样本集。对于样本数据，PCA 的主要思想是寻找一个新的数据表示空间，使得该表示空间的坐标轴方向能刻画原始数据的主要方向，然后将原数据投影到新的坐标空间下。问题的关键是求解新空间的基，根据内积与投影的关系，就可以直接通过内积运算得到投影后的新样本值。为了达到最优的降维效果，就希望能在保留足够多信息的条件下将维数降到最低。为此，需要需要设置两个约束：（1）作为基的变换向量应该正交；（2）每次变换完的数据样本在投影后的方差应尽可能的大。因此，假设原始样本向量为 **x**，变换向量为 **w**，则主成分分析就是解决一个如公式（2-1）的准则函数的问题：

$$\mathbf{w}=\arg\max_{\mathbf{w}} G(\mathbf{w})=E\left(\left\|\mathbf{w}^T\left(\mathbf{x}-E(\mathbf{x})\right)\right\|^2\right) \tag{2-1}$$

假设训练样本为$\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$，G(**w**)可以写为：

$$G(\mathbf{w})=\mathbf{w}^T\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^T\right)\mathbf{w}=\mathbf{w}^T\mathbf{S}\mathbf{w} \tag{2-2}$$

其中，$\mathbf{S}=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^T$ 是$\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$的散度矩阵。假设 $\mathbf{W}_i$ 是 **W** 的列向量，则 PCA 的优化问题转化为：

$$\mathbf{w}=\arg\max_{\mathbf{w}}\left\{G(\mathbf{w})=\mathbf{w}^T\mathbf{S}\mathbf{w}\right\},\ \text{s.t.}\quad \mathbf{w}_i^T\mathbf{w}_j=0,\quad \forall i\neq j \tag{2-3}$$

对 **S** 进行特征值分解就能得到公式（2-3）的最优解，最大的 m 个特征值对应的特征向量就是变换空间的基向量，即 PCA 的投影矩阵$\mathbf{L}=\left(\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_m\right)^T$,投影后的新数据为$\mathbf{X}'=\mathbf{X}*\mathbf{L}$。至此，PCA 降维的工作完成，由于只是将数据用新的坐标空间进行表示，PCA 损失信息少而且一定程度上可以获取特征的主要方向信息。

2.3 消失成分分析

表示数据的方法有多种，最简单的方法是直接用观测值来表示，也可以提取数据的一些相关的统计特征，如均值、直方图信息等一阶特性或方差、散度矩阵等二阶特性，这些信息都对数据的表示和任务的完成有着一定的正面作用。除了这些简单的统计特性，还可以通过建模获取数据中的一些其他语义信息。本节将介绍并分析一种可获取数据的零空间多项式语义信息的特征提取方法——消失成分分析（Vanishing Component Analysis，VCA）[33]。

2.3.1 模型描述

假设数据样本为 x，可以构造一个作用在数据特征上的多项式函数 f，使得 $f(x)=0$。通常情况下，一个数据样本可以用无数个多项式来描述，这些多项式的集合称为多项式集，当约束部分条件时，这些多项式的数目也被限定。这里进行如下约束：（1）满足相同分布的同类数据都可以用相同的多项式来描述；（2）多项式集合中的多项式元素需要相互正交，没有冗余信息。条件（1）的限定描绘了该类数的语义信息，例如，假设圆周上的点集为一类数据点，那么使用多项式 $f(x) = (x-a)^2+(y-b)^2-r^2=0$ 就可以描述该类点。限定（2）直接约束了满足这样多项式的个数，例如对于圆周上的点，$x*f(x)=0$ 同样成立，但是却与 $f(x)$冗余。利用多项式的这一特征，可以进行特征提取，具体描述为：对于任意样本点 x 属于点集 S，寻找一组多项式集 $F=\{f_1(x),\ldots f_k(x)\}$使得对于所有的$i \in \{1,2,...k\}$ 和任意的 $x \in S$，都有 $f_i(x)\approx 0$（给出一定容忍度，增加多项式的鲁棒性）。而对不属于此类的数据，不能满足样本在多项式集 F 上的所有多项式上的投影都为 0。因此可以使用一个多项式集合来特定的表示一类数据。由于数据满足在所寻找的多项式集上的投影为 0，同时考虑到求解过程中与数据零空间的一致性，该方法被称之为消失成分分析，方法的目标是求解数据的零空间多项式。

2.3.2 算法求解

对数据样本进行消失成分分析时，每类数据都需要求解一个对应的多项式，为方便表示，这里的数据集特指一类数据，其他的类别使用同样方法求解即可。假设数据集 $\mathbf{X} \in \mathbb{R}^{m \times n}$，$m$ 为数据的样本数，n 为特征数，$x \in \mathbb{R}^n$ 是 $\mathbf{X}$ 中的一个数据样本。则定义在 x 上的单项式 $f_{mo}:\mathbb{R}^n \to \mathbb{R}$ 为：

$$f_{mo}(x) = \prod_{i=1}^{n} x_i^{\alpha_i}, \alpha_i \geq 0 \tag{2-4}$$

设 $x^{\alpha} = \prod_{i=1}^{n} x_i^{\alpha_i}$，则单项式次数为 $\|\alpha\|_1 = \sum_{i=1}^{n} \alpha_i$。多项式由单项式线性组合而成：

$$f_{poly}(x)=\sum_{j}\beta_j x^{\alpha(j)} \tag{2-5}$$

其中，β_j 是第 j 个单项式的系数。多项式次数指的是最大的 $\alpha(j)$ 值。为找到投影为零的数据集的所有多项式，算法将按次数由低到高有规律查找所有这些多项式。这里给出一种巧妙的多项式表示方法：k 次多项式是由 k 次基多项式（或 k 次基单项式）线性组合而成，其中 k 次基多项式的定义是：给定一组次数为 k 次的多项式，这些多项式相互正交，且所有的 k 次多项式都可以由这一组基多项式线性组合而成。由此，多项式满足如下 3 个特点：

（1）非 0 常数是 0 次多项式的基；

（2）将$\{x_1，x_2，\ldots x_n\}$看做基单项式，则一次多项式就是这些基单项式的线性组合，实际使用中，将 1 次和 0 次进行合并；

（3）k+1 次多项式基由所有的 k 次多项式基与所有的一次多项式基的笛卡儿积构成。

因此求解零空间的问题就可以转换为构造基多项式和求解表示系数的问题。同时注意到求解多项式的第 2 个限定条件，即：如果 $f(x)$=0，则 $x*f(x)$ 类型的多项式就没有再次计算的意义，因此每次求解 k 次基多项式时，需要将其划分为两个部分：已经是零空间的多项式集合，和非零空间多项式集合，在求解 k+1 次的多项式时，只用到 k 次和 1 次的非零空间多项式。定义候选基集合 C，零空间多项式集 V，非零空间多项式集合 F。形式化的给出 V 的求解方法：

（1）基多项式的构造

当单项式次数是 1 次时，候选基集合 C 即是所有一次项和非零常数。当单项式次数为 $k(k>1)$时，如果已知 k-1 次的非零空间 F_{k-1}，那么 $C_k=F_{k-1}\times F_1$ 其中，$\times$ 表示笛卡尔积。

（2）表示系数计算

给定候选单项式基集合 C 时，假设其大小为 N，则每个样本都可以将其原本的特征投影到该多项式上。对于样本 $X\in\mathbb{R}^{m\times n}$ ，则 m 个样本全部投影到该多项式集上可以形成一个 $\mathbb{R}^{m\times N}$ 大小的 N 列的列向量组，使这 N 列向量线性组合为 0 的系数即是多项式的表示系数。这一问题可以用 SVD 分解求解，只需要找出对应 0 特征值的特征向量即可。在具体操作中，为增加算法的鲁棒性，将特征值为 0 这一约束条件进行放缩，设置一个较小的容忍值 ε 即可。具体求解算法见 VCA 算法。

假设 $p_i^l(x)$ 为数据集 $\mathbf{X}$ 的次数为 l 的第 i 个零空间多项式，则次数为 l 的零空间多项式集为 $\left|p_{n_l}^l(x)\right|=\{p_1^l(x)\ldots p_{n_l}^l(x)\}$，对应于数据 x，特征变换后的结果为：

$$x\rightarrow\left(\left|p_{n_1}^1(x)\right|\ldots\left|p_{n_k}^k(x)\right|\right) \tag{2-6}$$

VCA 算法

输入： $X \in \mathbb{R}^{m\times n}, Y \in \mathbb{R}^{m}$

输出： F, V

1. $F = \left\{ f(\bullet) = 1/\sqrt{m} \right\}, V = \phi$
2. $C_1 = \{ f_1, \ldots f_n \}$ *where* $f_i(x) = x_i$
3. $(F_1, V_1) = \text{FindRangeNull}(F, C_1, S_m, \varepsilon)$
4. $F = F \bigcup F_1, V = V \bigcup V_1$
5. *for* $t = 2, 3 \ldots$
6. $\quad C_t = \{ gh : g \in F_{t-1}, h \in F_1 \}$
7. $\quad$ *if* $C_t = \phi$
8. $\quad\quad$ *break*
9. $\quad (F_t, V_t) = \text{FindRangeNull}(F, C_t, S_m, \varepsilon)$
10. $\quad F = F \bigcup F_t, V = V \bigcup V_t$
11. *end*

FindRangeNull 算法

输入： F, C, S_m, ε

输出： F, V

1. denote $k = |C|$ *and* $C = \{ f_1, \ldots, f_k \}$
2. *for* $t = 1, \ldots k$
3. Let $\tilde{f}_i = f_i - \Sigma_{g\in F} \langle fi(Sm), g(Sm) \rangle g$
4. Let $A = \left[\tilde{f}_1(S_m), \ldots \tilde{f}_k(S_m) \right] \in \mathbb{R}^{m,k}$
5. *decompose* $A = LDU^T$ *using SVD*
6. *for* $i = 1, \ldots k$
7. $\quad$ let $g_i = \Sigma_{j=1}^{k} U_{ji} \tilde{f}_j$
8. $F_1 = \left\{ g_i / \left\| g_i(S_m) \right\| : D_{i,j} > \varepsilon \right\}$
9. $V_1 = \left\{ g_i : D_{i,j} < \varepsilon \right\}$

原始数据集 **X** 由 $\mathbb{R}^{m\times n}$ 变为 $\mathbb{R}^{m\times \sum_i n_i}$，特征变换完成。对于不同类的数据，分别学习各自类别的零空间多项式，然后将全部类别数据在全部类别的零空间多项式上进行投影。

2.4 Boosting 分类算法

Boosting[34, 35]是当前最重要的分类算法之一。通过顺序的使用一系列的简单分

离散 AdaBoost 算法[34]

输入：训练样本 X，标签 Y

输出：分类器模型

初始化权重矩阵： $w_i=1/N, i=1,\ldots,N.$

1. 初始化权重矩阵： $w_i=1/N, i=1,\ldots,N.$
2. 开始循环： $m = 1,2,\ldots,M$:
3. 使用训练样本权重 w_i 值学习弱分类器 $f_m(x) \in \{-1,1\}$
4. 计算 $err_m = E_w\left[\mathbf{1}\left(y \neq f_m(x)\right)\right], c_m = \log\left((1-err_m)/err_m\right)$
5. 更新：$w_i = \exp\left[c_m \mathbf{1}\left(y_i \neq f_m(x_i)\right)\right]$ $i=1,2,\ldots N$, 重新标准化 w：$\sum_i w_i$
6. 循环结束

类器对加权的训练样本进行分类，然后对学习到的系列分类器进行加权投票，从而得到最终的分类结果[36]。对很多分类器算法，这种简单的策略就会使分类结果得到明显的提高。直观上理解，这是一种多轮迭代算法，每一轮中加入一个新的分类器，被前一个分类器错分的样本作为主要起作用的部分用来训练下一个分类器。每轮中的分类器可以是一个简单的分类器，且对于两类问题只要保证当前分类准确率大于 0.5 即可。在每一轮的训练中，AdaBoost 通过加权方式选取的训练样本代替随机选取的训练样本，从而将训练的的重点集中于较难区别的样本上[37]。训练过程中，每训练一个弱分类器，样本权重就需要更新一次，被准确分类的样本在下一个分类器的训练中被选中的概率降低，相反权重就会提高。从统计的观点来解释，Boosting 算法是加法模型和最大似然估计的组合。特别是对于两类问题，Boosting 算法可以看作是一种使用最大伯努利分布作为释然准则的逻辑尺度的加模型，算法具有严格的理论证明[36, 38]。

最常用的 Boosting 算法是离散 Adaboosting 算法[39]。对于一个两类数据的分类问题，假设训练样本为$(x_1,y_1),\ldots(x_N,y_N)$，其中 x_i 是第 i 个样本的特征值组成的向量，y_i 代表了样本的标签，其值为-1 或 1。定义函数：

$$F(x) = \sum_1^M c_m f_m(x) \tag{2-7}$$

其中每个 $f_m(x)$都是一个结果为-1 或是+1 的弱分类器，c_m 是常数，对应的预测结果是 $sign\left(F(x)\right)$，$F(x)$值的绝对值大小提供了分类的可信度。算法在加权的训练数据上训练得到弱分类器 $f_m(x)$，然后根据分类结果更新样本的权值，其中本次错分的样本权重变大。通过这种方式，可以得到一系列弱分类器，最后分类时，分类器函数是之前每一步得到的弱分类器的线性加权组合。详细算法描述见离散 AdaBoost 算法。

2.5 小结

本章分别介绍了基于 PCA 主成分分析的降维算法，通过设定保留维数，PCA 算法可以尽可能多的保留原始信息。VCA 算法是一种通用的特征提取方法，通过求解数据特征矩阵的多项式零空间，该方法可以在一定程度上获取数据的语义结构信息。Boosting 算法是一种有效的分类器算法，通过迭代的使用弱分类器，对各样本进行不同的加权采样，弱分类的线性组合起到良好的分类效果，并能有效的避免过拟合。

第3章 基于自表示的非监督特征选择

3.1 引言

本文旨在构建一个级联深度学习模型。一般的深度学习方法大多采用特征变换、非线性操作和特征选择（约简）的多层迭代框架，本文构建的模型也是基于这一框架。因此，我们首先对框架中的基础算法——特征选择算法进行了研究。随着电子传感器和社交媒体的广泛使用，出现了大量高维数据。更高的数据维数会导致时间和空间复杂度的指数量级的提升，而现有的机器算法基本都是只针对低维数据，不容易扩展到高维数据，这些问题被称之为维数灾难[6]。实际的高维数据中包含了大量的冗余信息，而针对不同的机器学习任务，又有大量的特征与任务不相关。因此，不论从空间和时间复杂性来讲，还是从任务相关角度来讲，都应该进行特征选择。作为机器学习任务的重要一步，特征选择已经被很多学者所研究，主要的方法有：Filter 方法[18]、Wrapper[19], 和 Embedded[20]方法。最近，稀疏表示在机器学习领域的巨大成功也激发了该方法在特征选择方向的应用。通过使用 L_1 范数来获得稀疏解，L_1-SVM [21] 模型被用来进行特征选择，根据样本间的相似性，$L_{2,1}$ 范数的组稀疏约束也被用来进行特征选择，并取得了不错的效果[22, 40]。不仅数据样本的相似性可以使用稀疏约束进行特征选择，同样，直接使用特征的相似性进行稀疏正则化从而选择特征或许会取得更好的结果，而 $L_{2,p}$ 有着较 $L_{2,1}$ 更好的稀疏约束特性，因此本章研究了使用非凸的 $L_{2,p}$ 正则化自表示约束来构建特征选择模型（Feature Selection by Regularized Self-representation Based on $L_{2,p}$ norm, $L_{2,p}$-RSR），直接求解对应的特征子集，消除所选特征子空间内的冗余性。

3.2 问题描述

令原始数据空间为 $\mathbf{X}\in\mathbb{R}^{m\times n}$,其中 m 是样本数，n 代表了样本的特征维数。将所有样本的第 i 个特征写为向量 $\mathbf{f}_i\in\mathbb{R}^m$ 的形式，则数据样本为：$\mathbf{X}=[\mathbf{f}_1,...,\mathbf{f}_i,...,\mathbf{f}_n]$。特征选择的目的就是从这 n 个特征中选出 k 个（假设目标是选取 k 个特征），以便进行后续的分类、聚类等机器学习任务。利用样本的相似性，可以构造一个对应的矩阵 $\mathbf{Y}$。特征选择的任务就是求解如式（3-1）的一个多输出的回归问题：

$$J_0\left(\mathbf{X}_k\right)=\min_{k\subset D,\mathbf{w}} l(\mathbf{Y}-\mathbf{X}_k\mathbf{W}) \tag{3-1}$$

其中 D={1, 2, …, n}代表了维数集，k 代表了被选中的 k 个特征子集，$\mathbf{X}_k$ 代表了 $\mathbf{X}$ 对应 k 列，$\mathbf{W}$ 是对应 k 个特征的表示系数矩阵（也叫权重矩阵），而 $l(\mathbf{Y}-\mathbf{X}_k\mathbf{W})$ 是作用在 $\mathbf{Y}-\mathbf{X}_k$ 上的损失函数。

对于上述离散的排列组合问题，如果使用遍历的方法求解，共需要进行

$C_n^k = n!/k!(n-k)!$ 次的特征子空间搜索，这是一个 NP 难的问题。通过观察发现，如果令 **W** 的大小为 $m*n$， 将特征选择问题转换为对 **W** 进行正则化约束，就可以等价的求解上述问题（3-1），因此，可以得到问题一般化的等价形式，如公式（3-2）：

$$\min_{\mathbf{w}} l(\mathbf{Y}-\mathbf{XW})+\lambda R(\mathbf{W}) \tag{3-2}$$

其中 $l(\mathbf{Y}-\mathbf{XW})$ 是损失项， $R(\mathbf{W})$ 是作用在 **W** 上的正则化项， λ 是正则化参数，用来平衡损失项和正则化项。

3.3 损失项和正则化项

参照一些同类型的成功的非监督特征选择方法[41]，本文使用原始数据矩阵 **X** 作为对应的构造数据 **Y**，也即 **Y=X**。在一定的误差允许下，数据的每个特征都可以由数据中所有特征中的部分线性表示得到（可能包括被表示的特征自己）。因此，数据矩阵 **X** 的特征 $\mathbf{f}_i$ 可以写成如下的线性表示形式：

$$\mathbf{f}_i = \sum_{j=1}^{n} \mathbf{f}_j w_{ji} + \mathbf{b}_i \tag{3-3}$$

其中， w_{ji} 是表示系数矩阵 **W** 中的第 i 行、第 j 列元素， $\mathbf{b}_i \in \mathbb{R}^m$ 是偏差向量，对于所有特征将其写成矩阵的表示形式：

$$\mathbf{X} = \mathbf{XW} + \mathbf{B} \tag{3-4}$$

其中： $\mathbf{X} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n] \in \mathbb{R}^{m \times n}$， $\mathbf{W} = \left[w_{ji} \right] \in \mathbb{R}^{n \times n}$ ， $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n] \in \mathbb{R}^{m \times n}$ 。

本模型希望学习到一个可以表示不同特征重要程度的表示矩阵 **W**。在线性表示中，如果某个特征被频繁的使用，并且具有较大的表示系数，那么可以直观的认为该特征的有较强的重要性，在特征选择的任务中应该作为重点被选择出来，相反，如果一个特征在表示其他特征时都可以被忽略，而它本身又可以被其他特征线性表示，那么就可以认为该特征的重要性小，甚至为 0。这一现象可以通过表示系数矩阵 **W** 来反应，即如果第 i 个特征如上陈述的一样重要，$\|\mathbf{w}_i\|_2$ 将会明显较大，相反，$\|\mathbf{w}_i\|_2$ 就会很小甚至为 0，因此特征选择在模型中对应的目标就是得到一个行稀疏的正则化表示矩阵 **W**。为达到这一目的，使用 $L_{2,p}$（0<p<1）范数正则化对包含 **W** 的正则化项进行稀疏约束，为了减少奇异值的影响，使用 $L_{2,1}$ 范数约束损失项。综上，可以得到模型的最终目标函数：

$$J(\mathbf{W}) = \arg\min_{\mathbf{w}} \|\mathbf{X}-\mathbf{XW}\|_{2,1} + \lambda \|\mathbf{W}\|_{2,p}^{p} \tag{3-5}$$

当 $0 < p \le 1$ 时， $L_{2,p}$ 范数的定义为:

$$\|\mathbf{W}\|_{2,p}^{p} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} w_{ij}^2 \right)^{p} = \sum_{j=1}^{n} \|\mathbf{w}_i\|^{p} \tag{3-6}$$

其中 $\mathbf{w}_i$ 是 **W** 的第 i 行， λ 是一个正平衡常数。

3.4 迭代再加权最小二乘算法

在模型（3-5）中，损失项是非光滑的，而正则化项甚至是非凸的，这就使目标函数成为一个非凸的最优化问题，为解决这一难题，本节基于迭代再加权最小二乘算法（Iterative Reweighted Least-Squares，IRLS）[41, 42],提出了一种改进算法，并证明该算法会收敛到一个固定点。

算法中，给定当前 $\mathbf{W}^t$ 值，可以构造对角矩阵 G_B^t 和 G_W^t，他们对角线元素构成如下：

$$g_{B,i}^t = \frac{1}{2\left\|\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t\right\|} \tag{3-7}$$

$$g_{\mathrm{w},j}^t = \frac{p}{2}\left\|\mathbf{w}_j^t\right\|_2^{p-2} \tag{3-8}$$

其中，$\mathbf{x}_i$ 和 $\mathbf{w}_j$ 是 $\mathbf{X}$ 和 $\mathbf{W}$ 第 i 和第 j 行。权重系数的迭代解 $\mathbf{W}^{t+1}$ 可以通过构造如下代理函数求解：

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} Q(\mathbf{W}\,|\,\mathbf{W}^t) = \arg\min_{\mathbf{W}} \begin{Bmatrix} tr((\mathbf{X}-\mathbf{X}\mathbf{W})^T\mathbf{G}_B^t(\mathbf{X}-\mathbf{X}\mathbf{W})) \\ +\lambda tr(\mathbf{W}^T\mathbf{G}_W^t\mathbf{W}) \end{Bmatrix} \tag{3-9}$$

这是一个凸优化问题，可以直接通过求偏导得到解析解。即令 $\frac{\partial}{\partial\mathbf{W}}Q(\mathbf{W}\,|\,\mathbf{W}^t)=0$，有：

$$(\mathbf{X}^T\mathbf{G}_B^t\mathbf{X}+\lambda\mathbf{G}_W^t)\mathbf{W}-\mathbf{X}^T\mathbf{G}_B^t\mathbf{X}=0 \tag{3-10}$$

进一步得到 $\mathbf{W}^{t+1}$ 的解析解：

$$\mathbf{W}^{\mathrm{t}+1} = ((\mathbf{G}_W^t)^{-1}\mathbf{X}^T\mathbf{G}_B^t\mathbf{X}+\lambda\mathbf{I})^{-1}(\mathbf{G}_W^t)^{-1}\mathbf{X}^T\mathbf{G}_B^t\mathbf{X} \tag{3-11}$$

为了增加算法的稳定性，引入一个极小值 ε，并对 $g_{B,i}^t$ 进行如下修正：

$$g_{B,i}^t = \frac{1}{\max\left(2\left\|\mathbf{x}_i-\mathbf{x}_i\mathbf{W}^t\right\|,\varepsilon\right)} \tag{3-12}$$

在得到最终权重矩阵 $\mathbf{W}$ 值后，最大的 K 行 $\left\|\mathbf{w}_i\right\|_2$ 值所对应特征就是被选择的特征子集。详细算法描述见 IRLS 算法。

每次的迭代中，第 3 步和第 4 步的时间复杂度分别是 $O(m^2n)$ 和 $O(m^2)$。第 5 步更新 $\mathbf{W}$ 时，时间复杂度是 $O(m^3 + m^2n)$。在上述表示中，m 和 n 分别代表样本数和特征数。假设我们共需要做 T 步迭代，总的算法时间复杂度是 $O(T(m^3 + m^2n))$。

对算法的收敛性进行证明：引入一个代理函数 $Q(\mathbf{W}\left|\mathbf{W}^t\right.)$（即公式（3-9），定义函数 $F(\mathbf{W}) = J\left(\mathbf{W}\right) - J\left(\mathbf{W}\left|\mathbf{W}^t\right.\right)$，则通过求解公式（3-9），可以得到一系列迭代

IRLS 求解算法

输入: 样本矩阵 $\mathbf{X}\in\mathbb{R}^{m\times n}$，和 λ>0

输出: 特征权重矩阵 $\mathbf{W}$

1 令 $t=1$，初始化 $\mathbf{W}^1$

2 循环：

3 使用等式（2-12）更新 $\mathbf{G}_B^t$

4 使用等式（2-8）更新 $\mathbf{G}_W^t$

5 更新：$\mathbf{W}^{t+1}=((\mathbf{G}_W^t)^{-1}\mathbf{X}^T\mathbf{G}_B^t\mathbf{X}+\lambda\mathbf{I})^{-1}(\mathbf{G}_W^t)^{-1}\mathbf{X}^T\mathbf{G}_B^t\mathbf{X}$

6 $t=t+1$

7 直到收敛

8 $\mathbf{W}=\mathbf{W}^t$

$\mathbf{W}$ 值 $\mathbf{W}_{optimal}=\left[\mathbf{W}^1,\mathbf{W}^2,\ \ldots,\mathbf{W}^T\right]$。首先证明 $F(\mathbf{W})$ 在 $\mathbf{W}_{optimal}$ 是一个递减函数，然后再证明对于任意的 $t=1,2,\ldots,T\text{-}1$, 都有 $J(\mathbf{W}^{t+1})\le J(\mathbf{W}^t)$。

引理 1：$Q(\mathbf{W}|\mathbf{W}^t)$ 是如式(3-9)的代理函数，定义 $F\left(\mathbf{W}\right)=J\left(\mathbf{W}\right)-Q\left(\mathbf{W}\middle|\mathbf{W}^t\right)$, 则当 $\mathbf{W}=\mathbf{W}^t$ 时，$F(\mathbf{W})$ 取得最大值。

证明：上述引理等价于证明对于任意 $\mathbf{W}$, 都有 $F\left(\mathbf{W}^t\right)-F\left(\mathbf{W}\right)\ge 0$。

$$F(\mathbf{W}^t)-F(\mathbf{W})=J(\mathbf{W}^t)-Q(\mathbf{W}^t\middle|\mathbf{W}^t)-J(\mathbf{W})+Q(\mathbf{W}\middle|\mathbf{W}^t) \tag{3-13}$$

将 $J(\mathbf{W})$ 和 $Q(\mathbf{W}\middle|\mathbf{W}^t)$ 带入公式 (3-13), 则（3-13）可写为：

$$\sum_i\frac{\left(\left\|\mathbf{X}_i-\mathbf{X}_i\mathbf{W}^t\right\|_2-\left\|\mathbf{X}_i-\mathbf{X}_i\mathbf{W}\right\|_2^2\right)^2}{2\left\|\mathbf{X}_i-\mathbf{X}_i\mathbf{W}^t\right\|_2}+\sum_j\left(\left(1\text{-}\frac{p}{2}\right)\left\|\mathbf{W}_j^t\right\|_2^p-\left\|\mathbf{W}_j\right\|_2^p+\frac{p\left\|\mathbf{W}_j\right\|_2^2}{2\left\|\mathbf{W}_j^t\right\|_2^{2-p}}\right) \tag{3-14}$$

由于 2 次函数 $\sum_i\frac{1}{2\left\|\mathbf{X}_i-\mathbf{X}_i\mathbf{W}^t\right\|_2}\left(\left\|\mathbf{X}_i-\mathbf{X}_i\mathbf{W}^t\right\|_2-\left\|\mathbf{X}_i-\mathbf{X}_i\mathbf{W}\right\|_2^2\right)^2\ge 0$ 恒成立，因此只需要证明：

$$\sum_j\left(\left(1\text{-}\frac{p}{2}\right)\left\|\mathbf{W}_j^t\right\|_2^p-\left\|\mathbf{W}_j\right\|_2^p+\frac{p\left\|\mathbf{W}_j\right\|_2^2}{2\left\|\mathbf{W}_j^t\right\|_2^{2-p}}\right)\ge 0 \tag{3-15}$$

式（3-15）是一个多项式的不等式问题，假设 $a_j=\left\|\mathbf{W}_j^t\right\|_2, y_j=\left\|\mathbf{W}_j\right\|_2$ ，则：

$h\left(y_j\right)=\left(1-\frac{p}{2}\right)\left\|\mathbf{W}_j^t\right\|_2^p-\left\|\mathbf{W}_j\right\|_2^p+\frac{p\left\|\mathbf{W}_j\right\|_2^2}{2\left\|\mathbf{W}_j^t\right\|_2^{2-p}}$ ，即：

$$h\left(y_j\right)=(1-\frac{p}{2})a_j^p-y_j^p+\frac{p}{2}y_j^2a_j^{p-2},\ \forall a_j>\ 0,0\ <p<\ 1 \tag{3-16}$$

不等式 (3-16) 是关于 y_j 的多项式。当 $y_j = 0$, $h(y_j) > 0$；当 $a_j > 0, y_j > 0$, 分别对 $h(y_j)$求解 y_j 的一阶和二阶导数：

$$h'\left(y_j\right)=\frac{\partial h}{\partial y_j}=py_j(a_j^{p-2}-y_j^{p-2}) \tag{3-17}$$

$$h''\left(y_j\right)=\frac{\partial^2 h}{\partial y_j^2}=pa_j^{p-2}-\left(p-1\right)y_j^{p-2} \tag{3-18}$$

由于 $a_j > 0, y_j > 0$ 和 $0 < p < 1$, 则 $h''\left(y_j\right)>0$，$h'\left(a_j\right)=0$ 并且 $h\left(a_j\right)=0$. 根据凸优化理论以多项式单调性相关知识，可以得到：

$$h(y_j)\geq h\left(a_j\right)=0 \tag{3-19}$$

因此$\sum_j h\left(y_j\right)\geq 0$，即不等式（3-15）恒成立，即 $F\left(\mathbf{W}^t\right)-F\left(\mathbf{W}\right)\geq 0$。显然当 p 值为 1 时，上述结论也成立。证毕。

引理 2：令 $\mathbf{W}^{t+1}=\arg\min_{\mathbf{W}}Q(\mathbf{W}\mid\mathbf{W}^t)$，则有 $J(\mathbf{W}^{t+1})\leq J(\mathbf{W}^t)$。

证明：

$$\begin{aligned}
J(\mathbf{W}^{t+1}) \quad &= J(\mathbf{W}^{t+1})-Q(\mathbf{W}^{t+1}\left|\mathbf{W}^t\right.)+Q(\mathbf{W}^{t+1}\left|\mathbf{W}^t\right.)\\
&= F(\mathbf{W}^{t+1})+Q(\mathbf{W}^{t+1}\left|\mathbf{W}^t\right.)\\
\overset{F(\mathbf{W}^{t+1})\leq F(\mathbf{W}^t)}{\Rightarrow} \quad &\leq F(\mathbf{W}^t)+Q(\mathbf{W}^{t+1}\left|\mathbf{W}^t\right.)\\
&= J(\mathbf{W}^t)-Q(\mathbf{W}^t\left|\mathbf{W}^t\right.)+Q(\mathbf{W}^{t+1}\left|\mathbf{W}^t\right.)\\
\overset{\mathbf{W}^{t+1}=\arg\min_{\mathbf{W}}Q(\mathbf{W}|\mathbf{W}^t)}{\Rightarrow} \quad &\leq J(\mathbf{W}^t)-Q(\mathbf{W}^t\left|\mathbf{W}^t\right.)+Q(\mathbf{W}^t\left|\mathbf{W}^t\right.)
\end{aligned}$$

即：

$$J(\mathbf{W}^{t+1})\ \leq J(\mathbf{W}^t) \tag{3-20}$$

证毕。

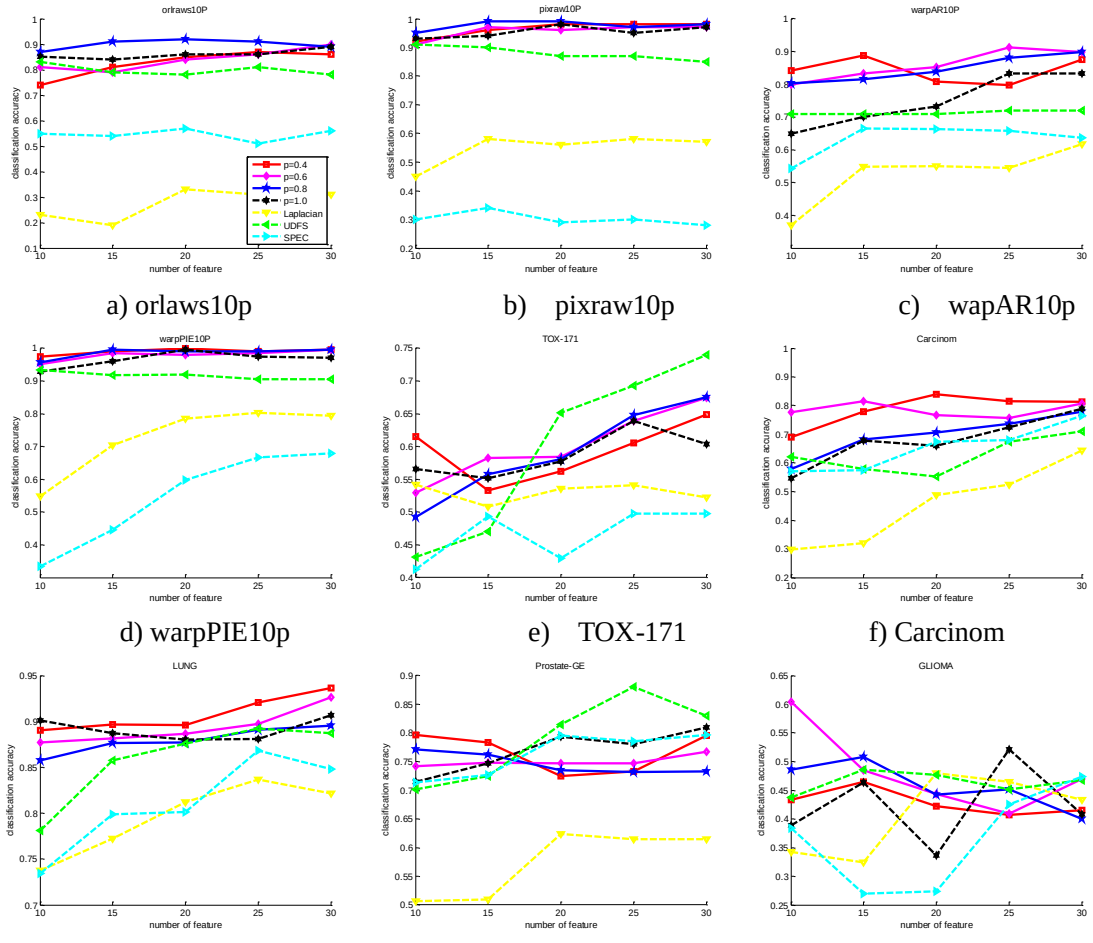
因此，通过改进的 IRLS 算法，目标函数在每次迭代求解中都是递减的。又因为 **W** 每次的迭代中都是一个闭合形式的解析解，IRLS 最终将收敛于一个固定值。

3.5 实验结果

为验证提出的 $L_{2,p}$ 正则化特征选择模型的性能，实验选取了 9 个公开数据集，

表 3-1 数据集信息总结

数据	样本数	特征数	类别数	关键词
orlraws10P	100	10304	10	图像，人脸
pixraw10P	100	10000	10	图像，人脸
warpAR10P	130	2400	10	图像，人脸
warPIE10P	210	2420	10	图像，人脸
TOX-171	171	5748	4	生物，微阵列
Carcinoma	174	9182	11	生物，微阵列
LUNG	203	3312	5	生物，微阵列
Prostate-GE	102	5966	2	生物，微阵列
GLIOMA	50	4434	4	生物，微阵列



a) orlaws10p　　b) pixraw10p　　c) wapAR10p

d) warpPIE10p　　e) TOX-171　　f) Carcinom

g) LUNG　　h) Prostate-GE　　i) GLIOMA

图 3-1 9 个数据集下选择不同维数特征的分类准确率

分别包括了 4 个人脸数据集 (orlraws10P, pixraw10P, warpAR10P, warPIE10P) 和 5

个微阵列数据集(TOX-171, Prostate-GE, Carcinoma, LUNG, GLIOMA)。 这些样本都是高维样本，特征数分布在 2400 到 11340 的范围内，表 3-1 总结了这些数据集的详细信息。我们将选择标准的 $L_{2,1}$ 正则化方法、Laplacian Score[43]、 UDFS[44]、SPEC [45]来与 $L_{2,p}$ 正则化方法进行比较。

表 3-2 不同方法的分类准确率(%)

数据	p=0.4	p=0.6	p=0.8	p=1.0	Laplacian	UDFS	SPEC
orlraws10P	81.40	80.20	**88.80**	**86.00**	27.40	79.80	54.60
pixraw10P	**95.80**	95.40	**97.40**	95.40	54.80	88.00	30.20
warpAR10P	**84.20**	**85.20**	83.80	74.10	52.60	71.40	63.30
warpPIE10P	**98.60**	97.83	**98.30**	96.23	72.67	91.63	54.50
TOX-171	57.31	57.03	**57.34**	55.08	52.98	**59.72**	46.62
Carcinom	**71.48**	**73.36**	69.68	66.09	45.49	62.75	65.28
LUNG	**89.80**	88.98	85.57	**89.02**	79.60	85.88	81.01
Prostate-GE	75.80	74.47	71.10	74.53	57.40	**79.03**	**76.33**
GLIOMA	42.87	**44.83**	44.14	42.37	40.89	**46.39**	36.55
Average	**77.47**	**77.48**	77.35	75.43	53.76	73.85	56.49

表 3-3 不同方法的聚类结果(NMI)

Data\Method	p=0.4	P=0.6	p=0.8	p=1.0	Laplacian	UDFS	SPEC
orlraws10P	69.14	64.15	**71.47**	**74.87**	39.46	62.40	42.95
pixraw10P	80.73	**84.10**	**84.17**	81.12	58.26	64.70	36.97
warpAR10P	42.74	42.43	47.63	**50.19**	16.90	**48.21**	44.38
warpPIE10P	**52.79**	51.14	48.81	43.14	18.31	**53.52**	24.23
TOX-171	21.06	**22.73**	**24.05**	16.73	10.14	10.86	10.03
Carcinom	**65.94**	**66.84**	63.96	57.58	42.37	46.51	57.76
LUNG	53.98	**55.89**	48.59	**57.26**	40.57	43.24	47.94
Prostate-GE	5.57	**6.26**	4.68	5.41	3.69	**7.08**	1.64
GLIOMA	17.59	**20.19**	12.54	13.90	**17.82**	17.06	15.36
Average	**45.50**	**45.97**	45.10	44.46	27.50	39.29	31.25

3.5.1 分类准确率比较

实验的预处理阶段将所有数据进行 0 均值 1 方差的标准化处理，最终分类时选用最近邻分类器，并记录 10 折交叉验证的平均结果。对于提出的模型，我们记录了 p 值等于 0.4，0.6 和 0.8 时的分类准确率，同时也使用上述提到的方法进行特征选择，并比较分类准确率。实验中，不同的方法需要设置不同的参数，为了公平比较，都尽可能的使每种方法都能取得最好的效果，对于方法 Laplacian Score, 和 UDFS，按照原论文的设置，近邻值 k 设为 5。对于方法 UDFS 和标准 $L_{2,1}$-RSR [41]

以及本文提出的方法，正则化参数需要进行调试。Laplacian Score 和 SPEC 方法中的高斯核参数同样需要进行合理选择。因此，实验中统一设置了带宽为{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100}的供选参数并记录不同参数下所对应的最好结果。选取特征数分别为{10, 15, 20, 25, 30}，并记录结果。

图3-1显示了所有特征选择方法所选择出的特征在9个数据集上进行分类的分类准确率。图中红色、粉色和蓝色对应 p 值为0.4、0.6和0.8时的分类准确率。从图中可以看出，本文提出的特征选择算法在实验中大部分数据集上都能取得最好的效果。表3-2显示了不同特征维数下对应分类结果的平均值。从表3-2可以看出，本章所提出的特征选择方法($0 < p < 1$)在分类效果上基本都优于标准的 $L_{2,1}$ 正则化方法以及其他流行的特征选择方法。

3.5.2 聚类效果比较

本节通过聚类结果来比较提出的方法和其他特征选择方法，使用最简单的K-Means算法进行聚类。为减少K-Means算法的聚类结果对初始化中心的依赖，初始化中心随机选择，最终的实验结果是20次随机实验的平均值。所有特征选择方法的参数设置参照分类部分的设置。我们使用 Normalized Mutual Information (NMI)[46]值来衡量聚类效果，其中，NMI值越大，聚类效果越好。表3-3给出了不同特征维数时，每种方法聚类结果的NMI平均值。从表中可以看到，当 $0 < p < 1$ 时，本文提出的方法在多数数据集上和总体平均结果上都能取得更好的结果。

3.6 小结

基于特征自表示的特性，本章提出了一种非凸的自表示正则化模型来进行特征选择。模型通过对表示矩阵进行行稀疏约束来达到特征选择的目的。由于 L_p 范数具有比 L_1 范数更加稀疏的特性，因此选用 $L_{2,p}$ 正则化项来约束线性表示矩阵。但是随着 p 值取值范围变为0到1之间，模型也由原来凸模型变为非凸模型。为了解决这一问题，本章提出了一种改进的迭代再加权最小二乘算法，同时对算法的收敛性进行了严格的证明。通过与标准 $L_{2,1}$ 特征选择方法和其他流行的特征选择方法进行对比，实验结果证明我们的方法所选出的特征有着更好的代表性，在分类和聚类上都取得不错的效果。

第4章 级联深度学习模型

4.1 引言

在研究了特征学习的基础上，通过对特征提取方法进行组合，本章将实现一个具体的级联深度模型。作为整个研究工作的指导，本章中首先构建了一个逐层级联深度学习模型框架。框架主要由两部分组成：多层的特征学习部分和最后的分类部分。由于构建模型的出发点是可泛化的、简单的深度学习，因此，特征学习部分中，特征提取、特征选择和分类策略中所使用的方法也应该具有较好的通用泛化能力。为成功实现该模型，我们对特征学习过程中的各阶段进行详细分析。同时提出了一种有效组合各层特征的方法，并介绍模型的分类算法。最后通过模型在 UCI 数据集上的实验结果来分析模型的有效性。

4.2 模型框架

深度学习是一个特征不断提取组合的过程，通过多层的非线性操作组合，模型可以抽象出数据的一些高阶的语义信息。实际操作中，对于经过预处理的原始数据，执行一个级联的多层操作，每层分别由：“特征提取”、“非线性变换”、“特征选择”三个子层组成，初始层的输入是原始数据经过预处理后的数据，其他层的输入是上一层的输出，最后一层输出最终的数据抽象表示。每层中，通过特征变换，抽取出数据的表示特性，该过程一般是维数升高的过程，相较于输入，经过变换的特征有了一定的表示特性，但是特征数却变多了，通过特征选择，我们进行一次降维，同时也是选择具有判别特性或表示特性特征的一步，一些特征选择方法是为了使模型拥有更好的区域自适应性，如卷积神经网络中的 max-pooling 和 mean-pooling 操作。非线性变换一般在特征变换和特征选择的中间执行，是该框架中重要的一环，它可以仿照神经元的激活和未激活状态，同时非线性变换的另一个重要原因是如果特征提取时使用线性变换，多层的线性变换仍然是一个线性变换，多层操作的作用就和直接学习一个线性变换的作用相同，不会起到逐层抽象的作用。

本文中，为消除不同特征可能存在的测量尺度的不同，首先对原始数据按最大最小方法进行归一化。在特征提取阶段，为获取数据的结构信息，选取了一种与领域关系小的特征变换方法，即提取数据多项式特性的消失成分分析（VCA）方法，由于 VCA 方法自身就是一种非线性变换，所以框架中可以不再进行非线性操作。特征选择的方法我们使用第二章提出的基于自表示正则化的非监督特征选择。由于 VCA 需要输入数据的维度不能太高，因此，我们对 VCA 变换的输入数

据进行 PCA 降维（由于 PCA 降维时，损失的信息比较少，而且可以降低特征维数，因此，也可以看作是一种特殊的特征选择方法）。在最后分类时，我们使用 Boosting 算法，一种既能分类又可以进行特征选择的方法，而在分类时的特征使用问题上，我们可以使用各层学到的全部特征，也可以仅使用最后一层特征。至此，我们的深度特征学习框架已经成型，见图 4-1。

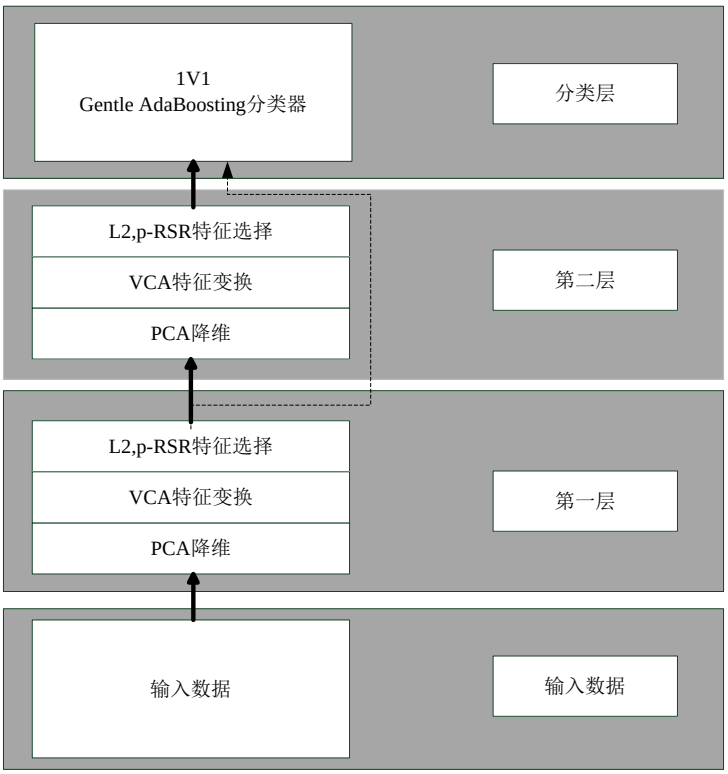


图 4-1 级联深度学习框架

4.3 特征学习

根据已有的级联深度学习框架，我们实现一个具体的深度模型，因此，我们需要对模型的训练过程进行研究与说明。本文提出的级联深度模型是一个多层结构，不同层之间，上一层的输出是下一层的输入，同一层之内，包括 PCA 降维阶段、VCA 特征变换阶段和 $L_{2,p}$-RSR 特征选择阶段。模型各层都能学到当前层的输出特征，不同层抽象的信息不同，最低层的特征最接近于原始特征空间，层数越高，特征抽象程度越高，高层特征可以为低层特征提供互补的信息。我们将充分利用各层学习到的特征，提出一种有效的特征组合方式。最后，我们使用基于二值分类问题的 Boosting 分类器，并将其成功扩展到多分类问题中。下面将一一解决这些问题。

4.3.1 PCA 降维阶段

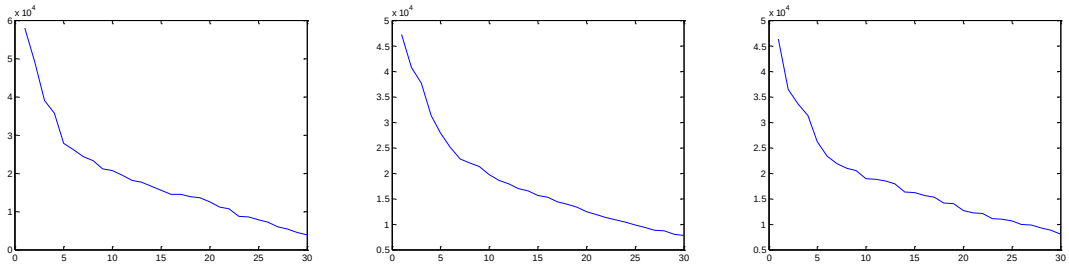
VCA 特征变换需要较严格的数据空间维数控制，因此我们需要在进行 VCA 特征变换前降维。PCA 降维的方式一般有两种，一种是直接指定保留的维数，另一种是通过设定特征值加和占总特征值和的比例来设置。按比例设置理论上可以控制数据信息的保留百分比，但是这种方法不便于控制保留后的特征维数，尤其是当特征维数多的时候，更加难以控制。而本模型中 VCA 变换后会产生大量的特征，其输入空间又需要严格控制特征维数，因此，我们直接设定 PCA 变换的保留维数。由于设定不同维数保留信息不同，对模型性能和实验结果都有影响，且这种影响不是与保留维数成正相关，因此，具体实验中，需要对 PCA 维数进行调试。

4.3.2 VCA 特征变换阶段

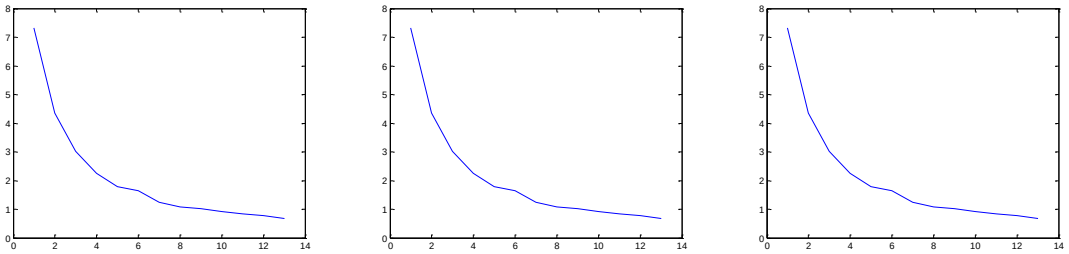
我们使用第二章介绍的 VCA 方法进行特征变换。VCA 方法可以将原始数据投影到零多项式空间内，从而起到特征提取的作用。使用该方法，我们可以不必知道数据的使用领域或其他先验知识，因为，只要输入空间是实数矩阵，就可以学习到其零空间多项式变换表示，这种变换方式不仅可以提取数据样本中的线性特征，还可以提取非线性特征，即如果一阶多项式包含数据的零空间，则多项式中包含数据的线性信息。其他不同次数的多项式可以提取数据的 2 阶甚至高阶的非线性特性。在本模型使用中，使用 VCA 进行特征变换涉及到两个具体问题：（1）多项式次数设定及初始特征维数设定；（2）算法求解时使用 SVD 分解时的极小值 ε 的设定。我们分别对这两个问题进行分析。

4.3.2.1 多项式次数及初始特征维数分析

本节所使用的符号同本文 2.3 节中介绍一致。在 VCA 求解算法中，对每一类数据 $\mathbf{X}$，样本维数为 m，特征维数为 n，初始化构造 C_1 的空间复杂度是空间复杂度是 O(n)，由 SVD 计算 F_1 和 V_1 的空间复杂度是 O(n^2)；构造 C_2 时，不妨设一阶多项式中不存在零空间，因此，C_2 的空间复杂度是 O(n^2)，同理求解 F_2 和 V_2 时的复杂度为 O(n^4)；构造 C_3 时，假设非零空间多项式的比例是 a（0<a<=1），则 C_3 的空间复杂度为 O(a*n^3)，求解 F^3 和 V^3 的复杂度为 O(a^2n^6)。为形象说明，假设初始样本的特征维数是 10^2，则当多项式次数为 2 和 3 时，空间复杂度为 10^8 和 10^{12}。样本用双精度浮点型表示，则二次多项式对应的内存空间为 0.8G 左右，3 次为 8000G 左右，8000G 的内存需求在实际应用中完全不可行，因此，本模型中只选用了一次和二次的多项式空间，初始数据的特征维数也通过 PCA 降低到 100 以下。VCA 方法的目标是为了更好的表示样本，获取数据样本的真实表示的零空间，因此，方法的初始输入特征数必须满足一个上界，这一上界和数据集单类的样本数

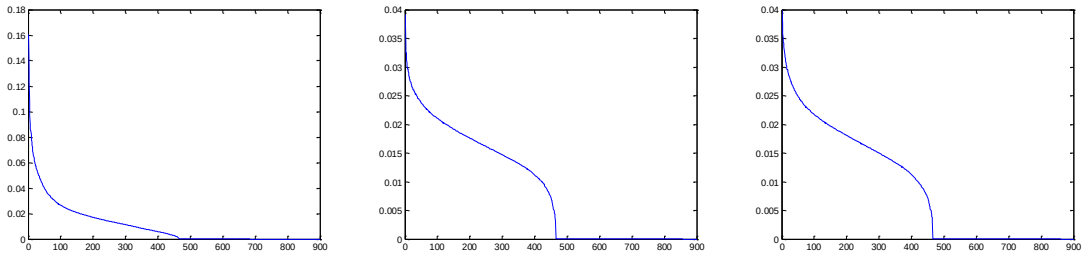


a) MNIST 数据 1　　b) MNIST 数据 3　　c) MNIST 数据 8

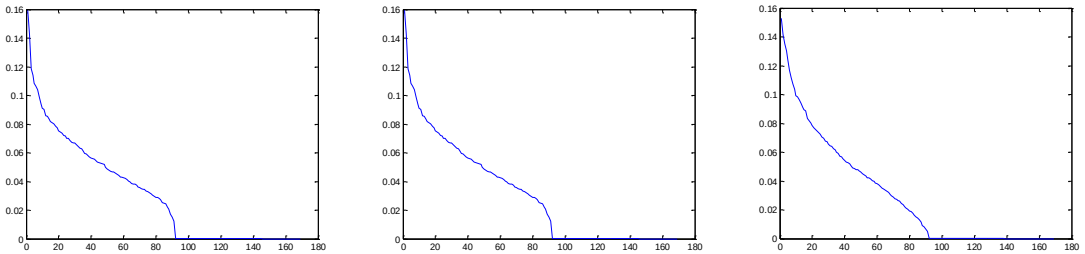


d) Letter 数据 A　　e) Letter 数据 B　　f) Letter 数据 H

图 4-2 不同数据的一阶 SVD 分解奇异值



a) MNIST 数据 1　　b) MNIST 数据 3　　c) MNIST 数据 8

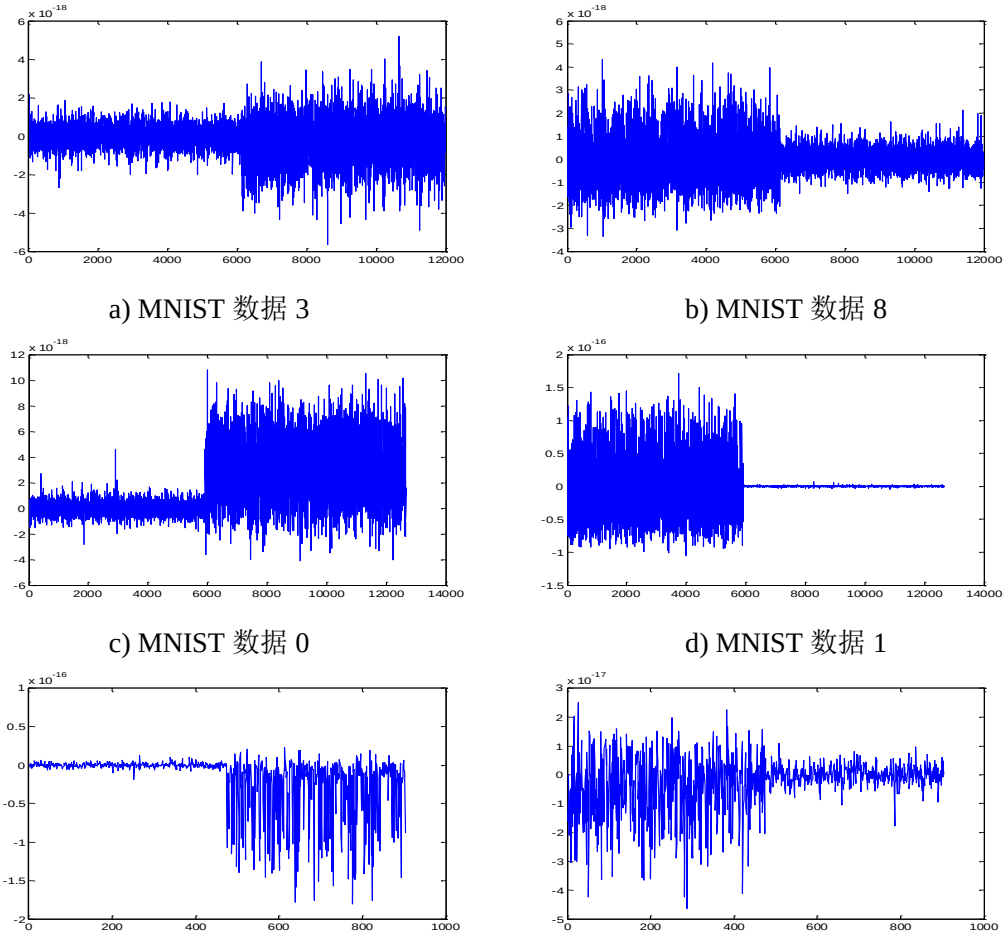


d) Letter 数据 A　　e) Letter 数据 B　　f) Letter 数据 H

图 4-3 不同数据的二阶 SVD 分解奇异值

有关。假设一类数据集的样本数只有 10 个，初始特征有 5 个，经过 VCA 变换后供选的 2 次多项式个数超过 25（大于 5^2）个，这是一个求解 10 个包含 25 个变量的方程组问题，根据方程组的知识，可以得到无穷组解，显然这些解不能真实的

反映数据的零空间信息，这就导致所求的多项式基本会反映样本错误的语义信息，从而在测试数据上，投影值不再准确。因此，综合考虑，VCA 算法的初始输入特征数应该小于数据样本数的开方，且小于 100。



a) MNIST 数据 3　　b) MNIST 数据 8

c) MNIST 数据 0　　d) MNIST 数据 1

e) Letter 数据 B　　f) Letter 数据 H

图 4-4 数据在零空间投影：a) 和 b) 分别是 3 和 8 在(3, 8)的零空间投影；c)和 d)分别是是数据 0 和 1 在(0, 1)的零空间投影；e)和 f)分别是 Letter 数据 B 和 H 在（B，H）的零空间投影

4.3.2.2 参数 ε 值的设定

VCA 算法中，为增加对数据集的鲁棒性，使用了极小值 ε 代替 0，为了确定 ε 的取值范围，本节在实际数据中进行了简单的测试，发现奇异值分解产生的奇异值有着明显的分界点，分界点处的值基本都是从 10^{-4}~10^{-2} 直接过度到 10^{-17} 的数量级，实验中这一经验可以直接使用。下面给出几个图示说明问题。

从手写体数据集 MNIST 和英文字母 Letter 数据集分别中选取三个数据 1、3、8 和 A、B、H，图 4-2 显示了选定数据在 1 阶时求解所得的全部奇异值，图 4-3 是

同样的数据 2 阶奇异值的求解曲线，从图中可以看出，一般数据集一阶多项式集合为空，2 阶多项式中大概有一半符合条件，通过查看具体数据，发现奇异值以 10^{-17} 数量级为分界点。

在确定了初始特征维数满足的条件和参数 ε 的设定后，我们在数据集 MNIST 和 Letter 上选取几对数据进行测试，验证分析的正确性。选取数据时，分别选取了 MNIST 数据集中相似的数据对（3,8），不相似的数据对（0,1）以及 Letter 数据中的数据对（B，H），变换的结果见图 4-4。从图中个可以看出，该方法能有效的将两类数据进行区分，同时，观察发现，变换后的数据有正有负，且数据量级都较低，为此，在使用这些数据时，可以先进行绝对值操作，这是因为 VCA 是查找零空间，即变换后的数据如果更靠近 0，说明数据属于该零空间对应的类别中，由于用的是绝对值距离度量，因此这里需要对变换后的数据进行取绝对值处理。绝对值变换后，我们再乘以一个较大的比例系数（一般是 10^{17} 或 10^{18}），以使数据在后面的变换中不至于损失精度。

4.3.3 $L_{2,p}$-RSR 特征选择阶段

VCA 特征变换后，如上一节分析，会产生大量的特征，我们需要对特征进行选择，特征选择的原因有很多，我们在绪论和第三章都有介绍，而且深度学习模型有效的一个原因也和特征选择有关，特征选择可以有效的选择与任务相关的特征。本文提出的 $L_{2,p}$-RSR 方法不仅可以有效的选择特征线性表示中起重要作用的特征，同时由于使用了 $L_{2,1}$ 范数约束损失项，还可以排除奇异样本的作用。该方法是基于特征空间的自表示特性，任何矩阵空间数据都具有这种特性，因此与领域无关，这也符合我们模型泛化能力的要求。本模型中，只需要设定方法中 $L_{2,p}$ 范数中的 p 值以及正则化参数值 λ 即可。$L_{2,p}$-RSR 特征选择操作的输入是 VCA 变换后的输出空间，输出是深度模型当前层的输出特征。

4.3.4 Boosting 分类与特征选择

学习到特征后，我们需要使用特征进行分类。通用的分类方法有很多，最简单的有最近邻分类器。SVM 分类器在研究和应用中也被广泛使用，基于核函数的 SVM 分类器还可以解决非线性问题。对于本模型，我们使用一种具有特征选择功能的分类方法：基于桩函数的 Gentle AdaBooting 分类器。基于桩函数的 Gentle Adaboosting 算法不仅能够进行分类，还可以起到特征选择的目的[11]，即每次从特征空间中选择一个特征并进行分类，通过控制弱分类器个数达到选择特征的目的，这与本文所提出的框架有很好的契合性。不仅因为其可以起到分类的效果，而且，还可以起到特征选择的作用。我们可以仅把其作为一种分类方法，也可以利用其特征选择的性能。下面我们将对基于桩函数的 Boosting 分类器进行分析和实现。

基于桩函数的离散 AdaBoost 算法

输入：训练样本 X，标签 Y

输出：分类器模型 $\sum_1^M f_m(x)$

1. 权重矩阵初始化：$w_i=1/m$, $i=1,\ldots,m$
2. 重复：$t=1,2,\ldots,T$
3. For $d=1,\ldots n$ do： $\left(err^d,\delta^d,a^d,b^d\right)=\min\sum_i^N w_i\left\|a^d\hbar\left(x_i^d>\delta^d\right)+b^d-y_i\right\|^2$
4. $feaId=\arg\min_d\left(err^d\right)$, $\left(feaId,\delta,a,b\right)=\left(feaId,\delta^{feaId},a^{feaId},b^{feaId}\right)$
5. $f_t(x)=a\hbar\left(x^{feaId}>\delta\right)+b$
6. 更新：$w_i\leftarrow w_ie^{-y_if(x_i)}$， $i=1,2,\ldots m$, 对 w 做标准化使得 $\sum_i w_i=1$
7. 重复结束

4.3.4.1 基于桩函数弱分类器的 Gentle Adaboost 算法

基于桩函数的 Gentle AdaBoosting 算法遵从本文 2.4 介绍的 Boosting 算法框架。其使用了一个简单的分类器模型：

$$F(x_i)=\sum_{m=1}^M f_m(x_i) \tag{4-1}$$

其中，弱分类器 f_m 的定义为：

$$f_m(x_i)=a\hbar\left(x_i^d>\delta\right)+b \tag{4-2}$$

$\hbar$ 函数是指示函数，x_i^d 代表第 i 个样本的第 d 维特征，δ 是阈值（即所谓的桩），a 和 b 是线性回归函数的参数。学习弱分类器时，对样本的每个特征都学习一个基于最小二乘的桩函数，得到并记录最小二乘的误差值，然后选取误差值最小时所对应的特征。因此，(d,δ,a,b) 可以通过如下的加权最小二乘法得到：

$$\min_{1\le d\le D}\sum_i^N w_i\left\|a^d\hbar\left(x_i^d>\delta^d\right)+b^d-y_i\right\|^2 \tag{4-3}$$

得到弱分类器 f_m 后，对 **W** 的权值进行更新：

$$w_i\leftarrow w_ie^{-y_if_m(x_i)} \tag{4-4}$$

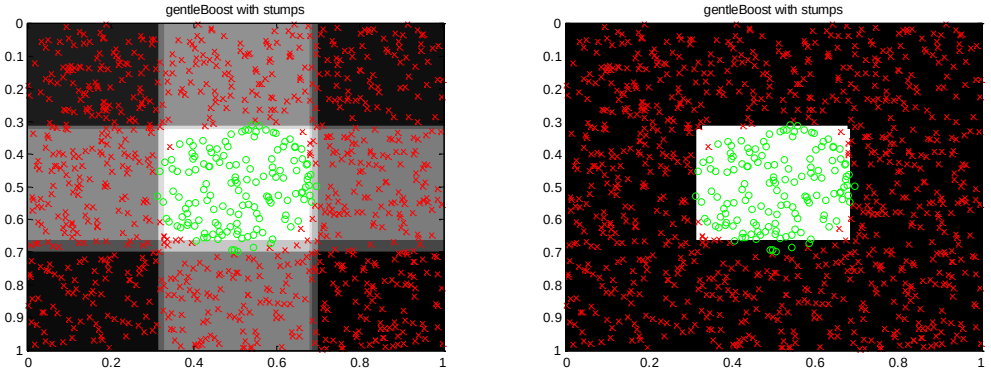
F 函数更新为 $F=F+f$。最终的分类结果为 $sign\left(F(x)\right)$。$F(x)$值的绝对值大小提供了分类的可信度。简单总结为基于桩函数的离散 AdaBoost 算法。

4.3.4.2 仿真结果

本节通过一个简单的模拟实验来验证基于桩函数的 Gentle AdaBoosting 分类器

算法的正确性。随机生成 1000 个 2 维的样本点，每个点的横纵坐标都在[0,1]之间。将以(0.5,0.5)为圆心，半径为 0.2cm 的圆内的点标记为正样本，圆外的点为负样本。测试集采用相同方法产生。

我们使用 Gentle AdaBoosting 分类器进行分类，弱分类器的个数设置为 15 个。图 4-5 是桩函数学习到的分类结果。图中，左边的灰度底色反映了分类器 $F(\mathbf{X})$的值，右边的黑白底色反映了 sign($F(\mathbf{X})$)的分类结果。图中的红色小叉是正样本，绿色小圆圈是负样本。从图中可以看出基于桩函数的 Gentle AdaBoosting 算法对非线性空间也具有较好的分类特性。



a) $F(\mathbf{X})$灰度值　　b) sign($F(\mathbf{X})$)灰度值

图 4-5 基于桩函数弱分类器的 Boosting 分类结果

4.4 特征组合

通过对级联深度模型中各阶段进行研究分析，我们成功构建了一个多层深度学习模型，并学习到多层特征。不同层特征提供的信息不同，底层特征是最接近于数据的真实信息，包含了最多的数据信息，但是却含有较少的语义信息，不够抽象。当模型为两层或 3 层时，就分别是对特征进行了不同层次的抽象，包含了一定的语义信息，当层数更高时，包含了更高级别的抽象信息。我们认为每层特征都是有用的，底层特征保证了数据信息失真量不至过多，高层特征又能为底层特征提供底层所不具有的互补性结构信息，如果能有效的组合各层特征，我们将得到一个很好机器学习模型。使用 Gentle AdaBoosting 特征选择和分类功能二合一的特性，本节将提出一种有效的组合各层特征方法，并使用组合特征进行分类。

级联深度模型所使用的分类器是有特征选择功能的 Gentle AdaBoosting 分类器，因此，我们将不同层的特征输入到分类器算法中进行分类，从而达到组合各层特征的目的。Gentle AdaBoosting 分类器的形式为 $\sum_1^M f_m(x)$，假设第一层的特征空间为 X_1，第二层的特征空间为 X_2，则通过第一层的特征学习一个强分类器

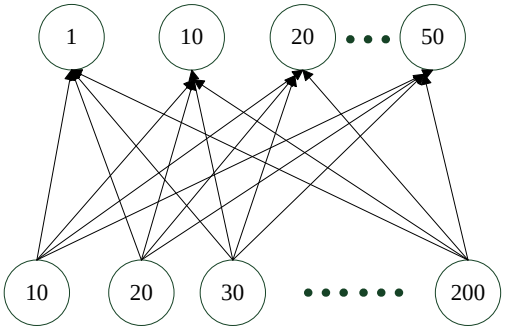


图 4-6 两层特征的组合方式

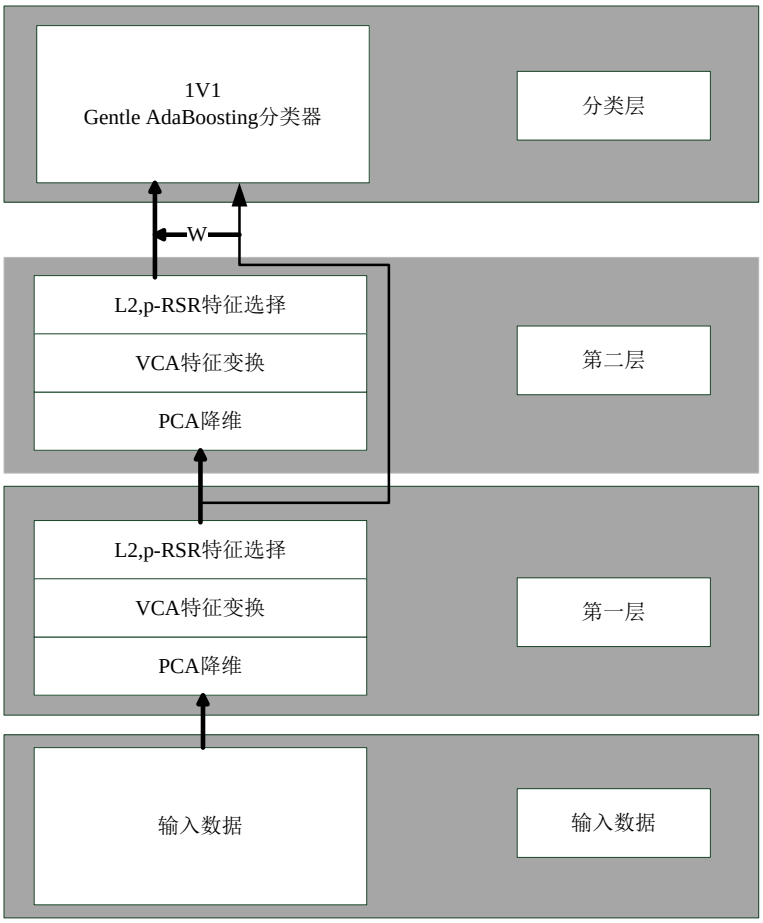


图 4-7 组合各层特征的级联深度学习模型

$F_1 = \sum_1^M f_m(\mathbf{X}_1)$，由于分类算法在每次弱分类结束后都会给出各样本当前的权值 $\mathbf{W}$，在已经产生了 F_1 的基础上，再单独使用第二层特征进行分类，就可以使用 F_1 结束时样本的权重分布 $\mathbf{W}$ 来初始化第二层分类 F_2 训练前的样本权重，进而学习到

分类器 $F_2=\sum_1^M f_m(\mathbf{X}_2)$，最后的分类结果为 F_1+F_2。其中 F_1 只用了第一层的特征，F_2 只用了第二层特征，但却用了 F_1 过程中的样本权值 **W**。

一个现实的问题是未分类前我们不能确定 Gentle AdaBoosting 使用多少个弱分类才能达到最佳的分类效果，同时，我们也不能确定第一层和第二层分别使用多少个特征进行组合才时能达到最优的分类效果，针对这两个问题，我们提出一种有效的组合特征方式。弱分类个数对应于特征选择个数，故在本节中二者的说法等价。经过实验验证，第一层分类器个数选择 200 以下，分类准确率基本就达到平衡，因此设定使用第一层特征学习时，弱类器个数选取带宽值为 weakClaNumber1=[10 20 … 200]，第二层为 weakClaNumber2=[1 10 20 30 40 50]，带宽值表示弱分类个数。训练分类器时，第一层直接训练 200 个弱分类器的训练模型即可，测试时通过控制弱分类器个数来记录不同数目弱分类器下对应的测试结果。

第二层的训练时，使用第一层不同的特征数和第二层不同的特征数进行组合。即第一层选择特征数为 10 个时，第二层可以选择的特征数为[1 10 20 30 40 50]，同理，第一层选择 20 个特征时，第二层为同样选取上述 6 组不同的弱分类器数。以此类推，直到第一层选择 200 个特征。这种方法看似会进行大量的组合，但是使用 Gentle AdaBoosting 算法的一个好处是，我们只需要设定第一层弱分类器数为 200，运行过程中分别记录弱分类器数为 10、20……200 的样本的权重值即可，同理第二层也是如此。按上述设置，第一层运行一次，会得到 20 个测试的分类结果，我们记录最大值；对于第二层，需要运行 20 次，得到 20*6 共 120 个结果，同时记录最大值时所分别对应的第一层和第二层特征数。3 层的特征组合方法与此类似。两层特征的组合方式见图 4-6。由此，得到最终包含分类方式的深度学习模型见图 4-7。第二层设定弱分类器个数时，通过加入特征选择数为 1 这个设定，使得两层的的选择效果总是不低于一层，这是因为如果第二层特征的加入会降低分类结果，则第二层特征数可以只选一个，这样第二层的弱分类器就不会降低分类效果。相反，如果第二层特征有效，则组合中会用到部分第二层的特征，从而提高最终的分类性能。这也是设置这种组合方法的一个优点。

4.5 基于二值分类器的多分类问题

基于桩函数的 Gentle AdaBoosting 分类器在分类过程中，每次学习弱分类器时只用一个特征进行分类，因此，我们不仅可以使用其作为最后的分类器，而且在组合模型各层特征中有着有效的作用。但是 Gentle AdaBoosting 分类器都是基于两类问题进行分类的，而我们的目标是多类问题的分类。因此本节考虑使用原始的二值分类器来解决多类问题。不妨设原始问题是一个 N 类问题，将二分类器扩张

到多类问题上常见的有三种方法：

（1）一对多法（1-v-all），共需构建 N 个 2 分类的 boosting 分类器。对于第 i 个分类器，训练集中正样本为第 i 类样本，负样本为除 i 类的所有样本；

（2）1 对 1 法（1-v-1）。共需构建 $N(N\text{-}1)/2$ 个弱分类器。将第 i 类样本设置为正样本，第 j 类设置为负样本，构造分类器 f_{ij}，由于 $f_{ij}=-f_{ji}$，因此，我们只需要构造 $N(N\text{-}1)/2$ 个分类器即可；

（3）层次分类器，构造一个树形结构的多类问题分类器。首先将所有类别分为两个子类，然后再将每个子类按相同思想划分为两个子类，直到问题只剩一类。

对于方法（3），从分类器的根节点出发，直到叶节点即可得到最后的分类结果，但是该方法的一个不足之处是：如果某个节点分类出现错误，那么后来的工作就都没有意义，也即该方法对分类路径上的各个分类器的精度要求都比较高，分类的准确率是路径上各分类器准确率的乘积，这样就导致了最后的分类准确率较低。方法（2）分类时，对于同一个测试样本，每个分类器都会给出一个分类结果，可以通过各分类器投票，票数最多的被选为分类结果。对于方法（1），训练分类器时，正负样本会出现严重不平衡的情况，这也导致了训练所得的分类器的置信度严重下降，而在每次训练时，相对与方法（2）每次训练集只是两类样本，方法 1 训练样本是全部样本，因此训练速度也会大幅度降低，在预测时，方法（1）还可能出现没有被分类的情况，即预测结果全部为负，这也是实际应用中所不能容忍的。因此综上考虑，本文最终选取方法（2）来构建多类问题分类器。

在使用方法 2 进行预测时，首先对测试样本使用全部分类器进行分类，由于 2 分类结果给出的是{+1,-1}，我们将其映射成对应的分类标签，然后进行投票，即统计单个样本在所有分类下被分到各类的票数，选取数目最多的类别标签作为我们预测的结果。对于有多个类别标签得票数一样的情况， 我们使用 Boosting 分类器的另一特性来进行筛选：即 Boosting 分类器返回值不仅有类别标签（正负值），而且还会返回一个可以代表分类置信度的实数值，将投票结果为相同类别的 $F(x)$ 值进行绝对化相加并选取最大值所对应的类别作为最后的标签。通过这种策略可以得到一个有依据的确定的分类结果。

4.6 实验结果及分析

为验证本文提出的级联深度模型在特征学习方面的有效性，本节将在 UCI 数据集上进行具体的实验。我们将构建一个 3 层的级联深度模型，实验中，记录层数分别为 1、2 和 3 的实验结果并对结果进行分析。

4.6.1 实验数据

为了便于与同类方法进行比较，本节选取了 UCI 上的四个数据集，Letter 、

Pendigit、USPS 和 MNIST，并将实验数据进行统一处理，即将原始训练样本按随机的 80%/20%重新分配为训练集和测试集。处理后数据集信息汇总情况见表 4-1。

表 4-1 训练样本信息总结

数据集	样本数	特征数	特征范围	类别数
Letter	12000	16	(-1,1)	26
Pendigit	5996	16	(0,100)	10
USPS	5833	256	(-1,1)	10
MNIST	48000	784	(0,255)	10

4.6.2 参数设置

为减小各特征度量尺度的不同，实验数据的预处理阶段我们对各维特征进行了最大最小归一化，将各维特征归一化到[0,1]的范围内。进行特征的 VCA 变换前需对数据进行降维。这是由于 VCA 特征变换对输入的特征空间维数有着较严格的上界要求。在进行 PCA 操作时，保留的维数越多，原数据损失的信息就会越少，但保留信息维数却和最后的分类准确率没有严格的正相关关系，这可能是由于如果保留维数较多，使得 PCA 变换时特征值较小的新维度投影后的特坐标也较小，这些特征比较符合零空间的特性，容易被 VCA 变换放大，但是从测试集的角度来看，较小的特征值意味着信息的区分性差，正负样本更容易混淆，从而导致分类准确率下降，因此，在实际模型中，PCA 降低的维数需要进行调试并选取合适的值。

消失成分分析在具体使用中只和输入特征空间以及容忍度参数 ε 有关，其中输入的特征空间通过 PCA 降维进行调节，容忍度参数 ε 的设置之前已经进行了详细的分析，在实际使用中，遵循分析结果，将其设定为常数 e^{-10}。模型中特征选择算法使用本文提出的 $L_{2,p}$-RSR 方法。在给定输入特征空间的情况下，该方法需要设置正则化参数 λ 以及需要选择的特征数 K。由于方法是是非监督的，并且最后分类时用到 Boosting 的监督选择分类算法，因此，此处尽可能在保留原始信息多的情况下选择特征，这里选择的特征数 K 设置为输入特征数的 0.8 倍，正则化参数 λ 设置为 0.1，即损失项对目标函数的贡献是正则化项的 10 倍。对不同的数据集，弱分类器个数设置也统一且固定，分别为 weakClaNumber1=[10 20 … 200]，weakClaNumber2=[1 10 20 30 40 50]，weakClaNumber3=[1 10 20 30 40 50]，这样设置的原因是第一层的特征保留了大部分的数据集信息，因此，选择尺度应该较大，这里设置最大为 200；高层特征提供低层特征的互补信息，相对用到的特征数较少，因此最大值设置为 50。

由于算法是基于二值分类的，我们认为如果两类问题的分类准确率高，则最

终的投票结果也会有着较高的准确率。因此分类准确率低的数据对就会影响整体的分类效率，我们将通过提升这些数据对的分类准确率来确定模型中的参数值，即 PCA 保留维数的设置。实验中使用如下方法进行操作：

（1）对数据样本直接使用 Gentle AdaBoosting 算法分类，选出若干对分类准确率明显低于平均值的数据对；

（2）设置 PCA 带宽值，在（1）中选取的几对数据对上进行测试，选出最优效果时所对应的 PCA 值。其中对于 PCA 的带宽设置，如果数据样本数较少，带宽的步长可以适当小，供选参数可以适当的增多；相反如果数据样本数大，带宽的步长可以适当地放大，供选参数可以适当地减少；

（3）使用（2）中选定的 PCA 值，进行全部类别的训练；

（4）多层训练时，训练当前层使用的 PCA 值，需要确定之前层的 PCA 值。我们将这些值设置为已经选定的 PCA 最优值，即如果只有一层时，PCA 的最优设定是 30，进行两层训练时，第一层的 PCA 设置固定在 30。当前层 PCA 的设定和选择遵从方法（1）和（2）。

4.6.3 实验结果

（1）Letter 数据集

我们直接在 Letter 数据集的原始特征空间使用 Gentle AdaBoosting 进行分类，得到图 4-8 所示的两两分类结果。从图中我们可以看出有较多的数据对分类准确率较低，我们挑选出最低的几个，其对应序号为：[31 41 45 125 128 157 164 173 304]。第一层 PCA 设置的供选带宽为：[4 8 10 13 16]，第二层为：[3 5 8 11 13 15 20]，第三层为：[3 5 8 11 13 15]。选中的部分数据在一层 PCA 值下的分类准确率见表 4-2。从表中可以看到，最优值为 PCA1=13。固定 PCA1=13，第二层在不同 PCA2 值下的分类准确率见表 4-3，最优值为 PCA2=11，同理，第三层结果见表 4-4，对应的 PCA3=13。实际带宽设置中，由于 Letter 数据每类的样本数都较小，PCA 保留值可以多设置几组。最终，各层使用 PCA 保留维数为[13 11 13]设置，得到如表 4-5 各层下 Letter 数据的总分类准确率。对比表 4-2， 4-3 和 4-4，发现随着层数的

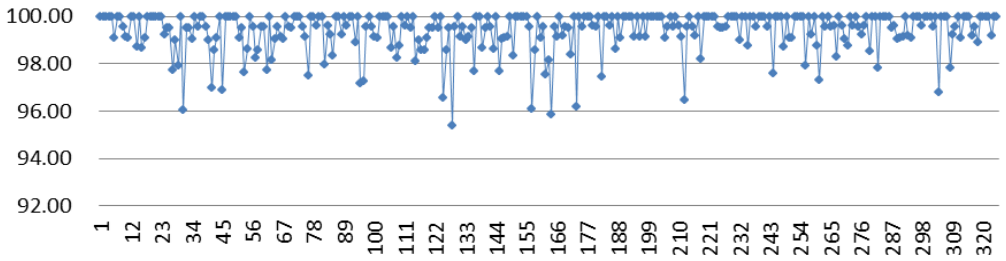


图 4-8 Letter 数据两两分类准确率(%)

表 4-2 Letter 部分数据第一层不同 PCA 保留值下分类准确率(%)

数据序号	PCA =4	PCA =8	PCA =10	PCA =13	PCA =16
31	91.09	97.52	98.51	98.51	98.02
41	84.92	93.47	93.47	96.48	97.49
45	96.02	97.35	97.79	97.79	96.90
125	82.76	93.10	95.69	97.41	97.41
128	89.91	94.95	95.87	99.54	99.54
157	83.48	95.22	93.91	96.96	94.78
164	93.09	94.01	96.31	97.70	98.16
173	90.72	94.09	95.78	97.47	96.20
304	84.55	95.91	98.18	100.00	100.00
均值	88.50	95.07	96.17	**97.98**	97.61

表 4-3 Letter 部分数据第 2 层不同 PCA 保留值下分类准确率(%)

数据序号	PCA =3	PCA =5	PCA =8	PCA =11	PCA =13	PCA =15
31	99.01	99.01	99.01	99.01	99.01	99.01
41	96.48	96.48	96.98	96.48	96.98	96.48
45	98.23	97.79	98.23	98.67	98.67	98.23
125	98.71	97.84	97.41	98.28	98.28	97.84
128	99.08	99.54	99.08	99.54	99.54	99.08
157	97.83	97.83	97.83	97.83	97.39	97.39
164	98.16	98.16	98.62	98.62	97.70	98.16
173	97.47	97.47	97.89	97.89	97.89	97.89
304	100.00	100.00	100.00	100.00	100.00	100.00
均值	98.33	98.24	98.34	**98.48**	98.38	98.23

表 4-4 Letter 部分数据第 3 层不同 PCA 保留值下分类准确率(%)

数据序号	PCA =3	PCA= 5	PCA =8	PCA =11	PCA =13	PCA =15
31	98.51	98.51	98.51	98.02	98.51	98.51
41	96.48	96.98	96.48	97.49	97.49	96.98
45	98.67	98.67	99.12	98.67	99.12	98.67
125	98.28	98.28	98.28	98.28	98.71	99.14
128	99.54	100.00	99.54	99.54	99.54	99.54
157	97.83	97.83	97.83	97.83	97.39	97.83
164	99.08	99.08	98.62	98.62	99.08	99.08
173	97.89	97.89	97.89	97.89	97.89	97.47
304	100.00	100.00	100.00	100.00	100.00	100.00
均值	98.48	98.58	98.47	98.48	**98.64**	98.58

表 4-5 Letter 数据在不同层的分类准确率(%)

0 层	1 层	2 层	3 层
94.37	96.83	97.37	97.60

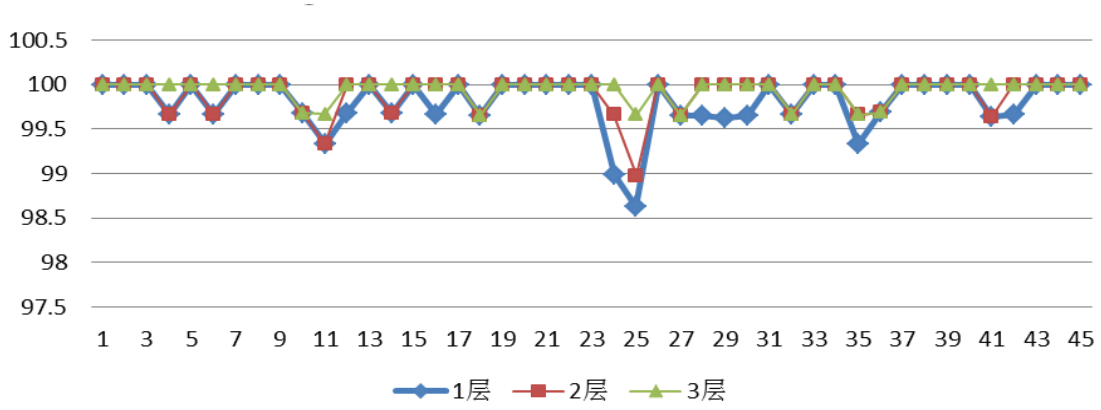


图 4-9 Pendigit 数据集在不同层的两两分类准确率(%)

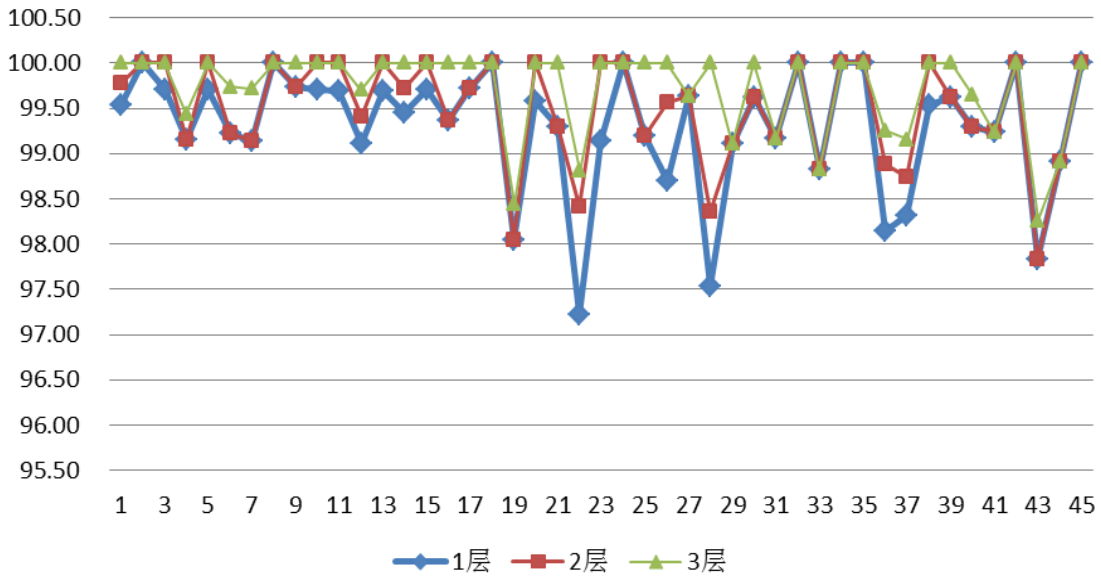


图 4-10 USPS 数据集在不同层的两两分类准确率(%)

在局部数据对上的分类准确率显著提升。当固定 3 层 PCA 值，表 4-5 也显示了模型在进行特征为属性值的 Letter 数据分类时，随着层数的增加，分类准确率增加。

（2）其他数据集

同在 Letter 数据上进行实验时的方法一样，本文在 Pendigit、USPS、MNIST 也进行相同的实验，首先通过部分数据对选取各层最优的 PCA 值，然后在整体数

据集上进行实验，并分别记录了结果。图4-9，4-10,4-11分别显示了数据集Pendigit、USPS、MNIST在不同层下的所有2值分类准确率。

为更加公平对比，我们使用本文的方法得到的分类结果同使用VCA特征变换后再直接分类的实验结果[33]进行对比，其中标准VCA方法使用线性SVM分类器，KSVM使用的是多项式核。结果见表4-6。从该对比结果中我们能够得到如下两个结论：（1）本文提出的特征学习模型是有效的，高层的特征可以为低层特征提供有效的互补信息，使得组合后的特征有着更高的分类准确率；（2）本文基于VCA特征变换所构建的模型在单层的情况下有着与原著中可比拟的分类准确率，2层和3层的结果较原模型相比都有着明显的提升。

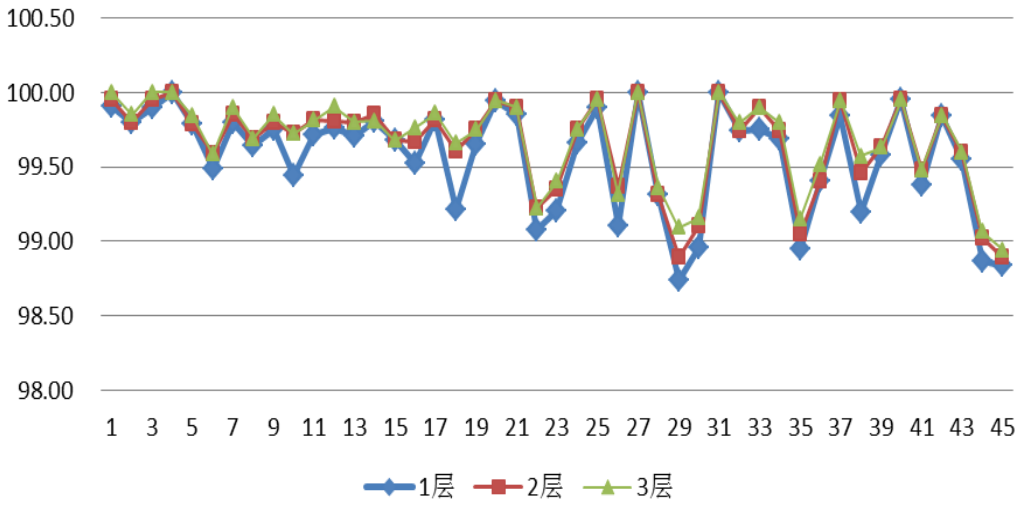


图 4-11 MNIST 数据集在不同层的两两分类准确率(%)

表 4-6 不同方法的分类准确率(%)对比

数据	标准 VCA	KSVM	本文提出的模型		
			1 层	2 层	3 层
Letter	95.20	95.70	96.83	97.37	**97.60**
Pendigit	99.58	99.58	99.53	99.67	**99.73**
USPS	98.50	98.60	99.78	98.35	**98.90**
MNIST	97.80	98.00	97.76	98.15	**98.26**

4.7 小结

本章构建了一个级联的深度学习框架，通过对框架中使用到的具体特征变换和特征选择及分类器算法进行研究与分析，成功实现了一个具体的级联深度模型。

针对模型各层学习到的不同特征，提出了一种基于 Boosting 特征选择和分类的特征组合方法。在 UCI 上的实验结果显示了随着模型层数的增加，我们的方法在分类准确率这个度量上可以取得明显的提升，这一结果也验证了本文提出的模型可以学习到数据更加有效的表示特征，同时也验证了本文提出的特征组合方式的有效性。

第5章 扩展数据对模型的影响

5.1 引言

深度模型可以逐层进行特征抽象，使得抽象出的特征具有丰富的语义信息。这其中除了模型结构可以有效的学习数据的抽象特征外，还有另外一个重要的限制因素——样本的数量。深度学习的发展离不开大数据的提出和分布式计算的应用。大数据的出现使得模型从数据中学习真实的统计信息有了物质基础，分布式计算为模型的有效计算提供了必要的工具。本文前面章节已经成功实现了可处理一般向量数据的级联深度模型，本章我们探讨数据量对模型性能是否有影响进行分析。

由于本文提出的模型适应于普通的向量数据而不需要任何先验知识，当知道数据的一些先验知识时，我们可以充分利用先验信息进行数据增强，对训练数据进行有效的扩展，并以此增强模型的学习性能。前文的实验中，使用到了 USPS 和 MNIST 数据，在模型构建和数据使用中，我们没有使用任何先验知识，但实际上由于这两个数据集都是数据手写体图像数据，因此，其满足一定的先验知识，即样本间可能存在着微小的变换，如平移、旋转、放缩等。我们将使用这些先验知识对数据进行扩展，从而提升模型的性能。本章首先提出一种基于手写体数字图像的扩展方式，并对扩展中使用的参数对模型的影响进行分析，最后达到提高级联模型特征学习性能的目的。

5.2 基于切向量的图像数据扩展方法简介

多数机器学习任务中，不仅需要训练数据，还需要一些关于任务的高维先验知识。比如图像字符识别领域，需要应对数据样本在较小的局部区域有着平移旋转等不变性。通常这种不变指的是基于流形（模型的可能变换的集合）的形式不变性。因此我们可以生成数据的流形空间，然后根据流形空间学习分类器。然而，基于流型的变换通常是高阶的非线性变换，这种操作常常伴随着较高的代价和不可信性。为此，Simard 等提出了使用切向量表示法将非线性变换转化为线性变换的趋近[47]。假设流形函数是对应于模式和变换参数的连续函数，定义流形变换 s，原始模式 P，变换参数 α。根据泰勒公式可得：

$$s(P,\alpha)=s(P,0)+\sum_{i=1}^{L}\alpha\frac{\partial s(P,\alpha_i)}{\partial\alpha_i}+\sum_{i=1}^{L}O\left(\alpha_i^2\right)\approx P+\alpha T \tag{5-1}$$

如式（5-1），当参数 α 较小时，可以忽略泰勒展开式的二次项和更高次项，因此，非线性的流形变换 s 就可以写为原始模式与特征向量的线性组合，其中 T 为流形

变换的切向量。更形象一点，如图 5-1，显示了像素空间的旋转变换，当α值较小时，通过切向量的线性表示可以趋近实际旋转变换。

通过泰勒展开式，可以看出使用切向量可以表示将非线性流形变换变为线性表示。然而图像数据却不能直接使用这种方式，因为图像数据都是离散数据，流形函数也因此变为离散函数。对此一个解决方案是通过使用二维高斯函数的卷积形式，产生可微的插值函数。并最终通过高斯微分与高斯卷积的关系，得到图像的基于切向量的流形扩展方式。

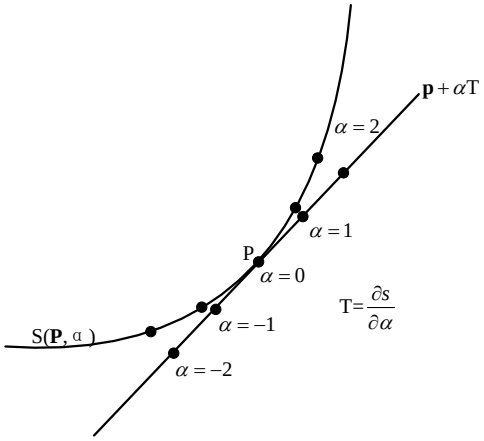


图 5-1 像素空间的旋转表示

5.3 手写体图像数据直接扩展法

切向量生成法从流行的角度，使用高斯卷积函数对模式进行变换，顺利的用线性变换趋近于非线性变换。给定变换函数，可以直接求得切向量的形式，只需再给定变换参数 α，既可以完成数据集的扩展。仔细分析，图像数据中使用的变换方式主要涉及到 x 和 y 平移变换，旋转变换，剪切变换以及缩放变换等。虽然除平移变换为其他均是非线性变换，然而我们却可以直接对图像进行这些操作。由于图像是离散空间的数据，因此在变换过程中需要进行插值。本节中我们不使用切向量法而直接对图像进行这些变换，设定变换尺度以及随机生成变换顺序，观察生成的数据集对实验结果的影响，同时讨论变换尺度对模型有效性的影响。

我们的模型是基于点值运算的，因此，同类数据模式可以看作一个流行。图像数据流形中，我们考虑的有效变换有以下几种：x 和 y 方向平移、旋转、剪切放缩以及平行四边形变换。首先将图像归一化到[0,1]之间，并且背影颜色为黑色，即数值 0，图像变换前后大小不变。为下文说明清楚，定义两个空间集合 S 和 C，其中原始空间表示为 S，变换后的空间表示为 C，则 $S \wedge C$ 表示变换后两幅图像空间重合区域，同理 $S - C$ 表示变换后造成的空白区域，$C - S$ 表示变换后超出原图像

的区域。下面我们对使用的变换进行说明。

（1）平移变换

平移变换包括横向平移和纵向平移，假定 X 轴为横向，则其对应为沿 x 轴方向平移和沿 y 轴方向平移。设定平移变换的参数为α，α是以 0 为对称轴，对称分布在坐标轴上的整数值，如果α为负，代表图像沿 x 轴负方向或 y 轴负方向平移，相反，则向 x 轴正方向或 y 轴正方向平移。图像的平移遵从以下函数：

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x+\alpha \\ y \end{pmatrix} \tag{5-2}$$

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ y+\alpha \end{pmatrix} \tag{5-3}$$

平移变换前后，图像大小不变，因此在平移后需要对图像进行切割和补充。设定平移造成的空缺$S-C$区域为背景区域，因此直接以 0 值填充，平移导致超出原图的区域$C-S$区域直接切割。

（2）旋转变换

图像的旋转是指以图形的中心为圆心进行的顺时针或逆时针旋转（α为正时，表示逆时针旋转，相反为顺时针旋转）。设定旋转变换参数为α，旋转操作遵从以下函数：

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x\cos\alpha - y\sin\alpha \\ x\sin\alpha + y\cos\alpha \end{pmatrix} \tag{5-4}$$

从式（5-4）我们可以看出，旋转后的图像坐标不再是整数，因此，旋转后必须对新的像素点灰度值进行插值运算。本文中选用双线性插值算法。和平移变换相同，旋转后产生的空白以 0 值填充，超出的区域直接切割。

（3）放缩变换

图像的放缩是指图像大小的变换，变换过程中，图像中心仍然为坐标轴原点。设定放缩参数α，则α分布在 1 周围表示图像变化大或缩小的倍数，当α等于 1，图像不变。图像的放缩变换遵循如（5-5）表达式：

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x+\alpha x \\ y+\alpha y \end{pmatrix} \tag{5-5}$$

同前几种变换一样，放缩变换也产生小数坐标，使用双线性法进行插值，对于图像变换中参数的空白和多余部分，分别用 0 值填充和切割操作。

（4）剪切变换

图像的剪切变换也称错切变换，是一种类似于四边形不稳定性的变换。假设变换参数为 α，变换过程中以图像中心为原点，进行类似平行四边形变换，变换方向分为 x 方向剪切变换（如式（5-6））和 y 方向剪切（如式（(5-7））。

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{pmatrix} x - y\tan\alpha \\ y \end{pmatrix} \tag{5-6}$$

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{pmatrix} x \\ y - x\tan\alpha \end{pmatrix} \tag{5-7}$$

同样变换过程中产生了小数的坐标情况，我们使用双线性插值方法进行插值。变换过程产生空白和超出的区域，分别进行 0 值填充和切割。

（5）综合变换

上述变换中，着重分析了图像的各种单独变换，本节考虑如何对这些变换进行综合。这里涉及两个主要问题：（1）各变换的顺序问题。我们可以直接按先后顺序对图像进行平移、旋转等变换。然而，我们需要为这些变换指定一个顺序，因为，顺序不同，有些变换的结果不同。我们通过简单的 x 平移和放缩变换说明这一问题。假定变换前像素坐标为(x,y)，平移变换参数为α_1，放缩变换参数为α_2。则先平移后放缩的坐标变换为$(x,y)\rightarrow(x+\alpha_1,y)\rightarrow((1+\alpha_2)(x+\alpha_1),(1+\alpha_2)y)$，而先放缩后平移的坐标变换为$(x,y)\rightarrow((1+\alpha_2)x,(1+\alpha_2)y)\rightarrow((1+\alpha_2)x+\alpha_1,(1+\alpha_2)y)$，显然二者不一样。如果直接指定了变换的顺序，那么就限定了数据样本的变换趋势。但真实数据的变换趋势却是随机的，因此，为解决这一问题，每次生成新样本时，我们随机生成变换的顺序，并按序逐一变换。（2）变换的参数问题。实际手写体

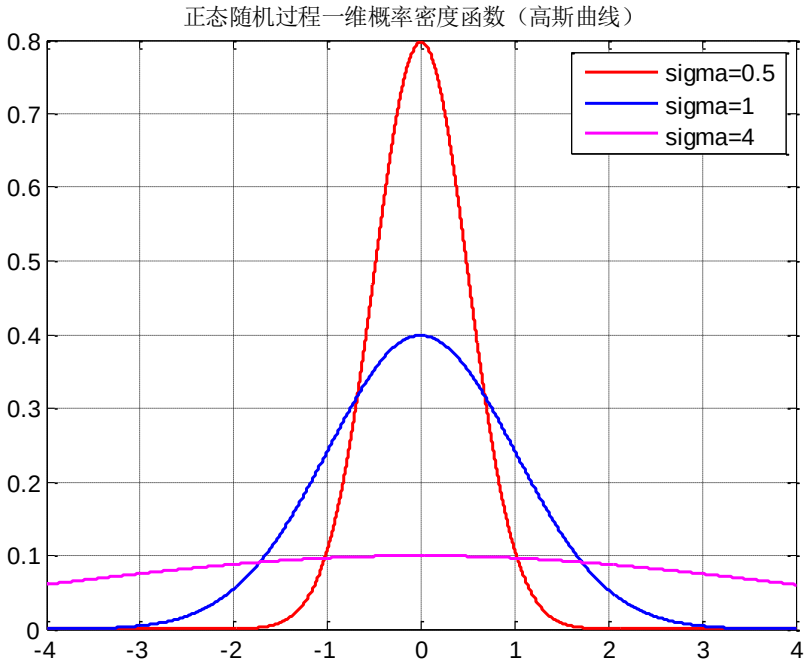


图 5-2 不同标准差的高斯分布函数

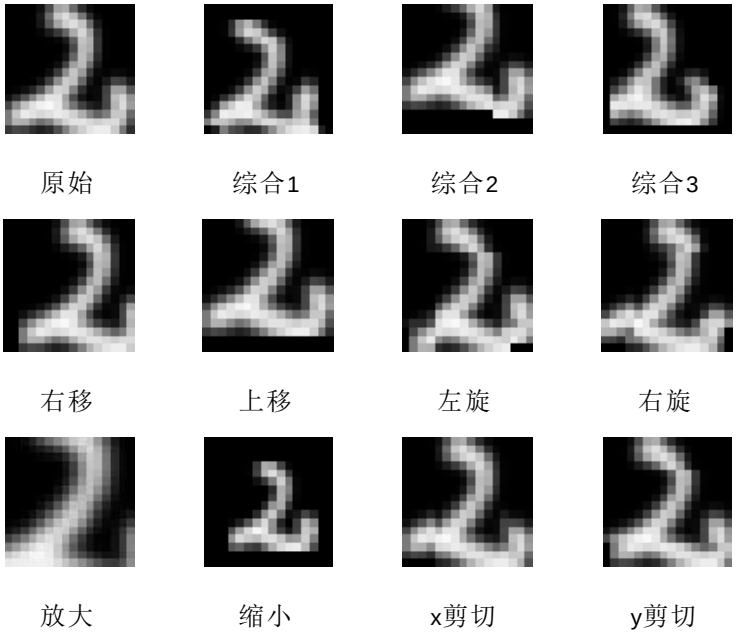


图 5-3 USPS 数据变换

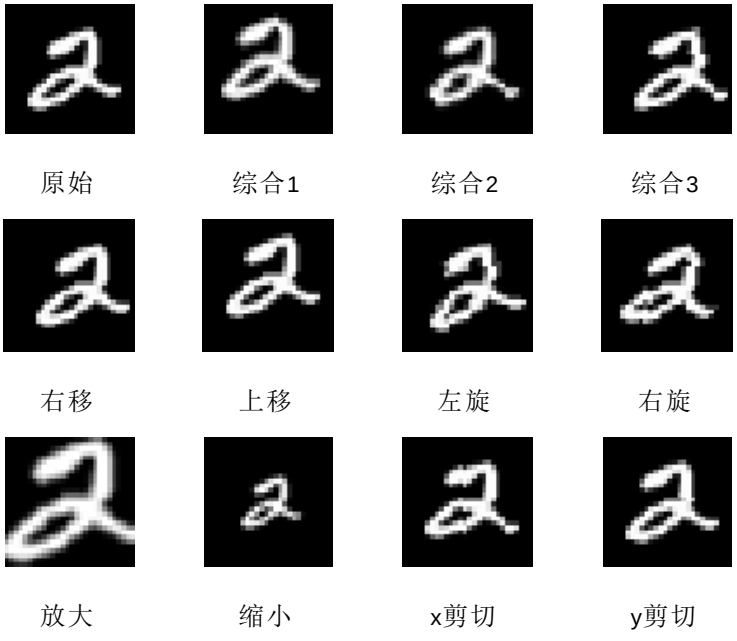


图 5-4 MNIST 数据变换

图像可能会因为有上述的变换而不同，然而，由于 USPS 和 MNIST 已经经过了对齐等预处理，因此数据之间变换不会太大。所以我们需要确定变换参数的大小范围，并从范围中选择具体参数值。由于同一变换的不同尺度值发生的概率不同，比如平移变换，平移一个像素值的概率肯定要大于平移两个像素值的概率。因此，我们可以使用高斯采样从参数范围中选取具体的值。高斯分布的概率密度函数见图 5-2，具体的参数设置我们在下节研究。图 5-3 和图 5-4 分别显示了 USPS 和 MNIST 图像经过本文提出的方法的变换结果。其中左上角的图像为原始图像，第一行的图像为所有变换的综合结果。第二和第三行是只用一种方式进行变换的结果。

5.4 扩展数据集对模型性能的影响

为提升模型学习性能，我们对数据集进行了扩展。然而，一个现实问题是只有有效训练数据的增加才可能提升模型的特征学习性能及最后的分类结果。 因此本节主要研究以下几个问题：（1）扩展变换的参数对特征学习的影响；（2）分别使用提出的手写数字图像变换方法和基于切向量变换方法产生的 MNIST 数据集进行分类，对比检验提出的图像变换方式的有效性。（3）研究扩展数据量对模型的影响。下面将分别对这几个问题进行研究和分析。

5.4.1 扩展参数对模型性能的影响

通过对数据集按本文提出的方法进行扩展，我们讨论什么尺度的数据变换更合理。由于 MNIST 原始数据量巨大，扩展后调试时间较长，而 MNIST 和 USPS 的性质相似，二者应该有相似的结论。因此本节只选取 USPS 数据集进行验证说明。USPS 数据变换过程中，参数较多，如果枚举所有变换的影响，需要进行各种组合，且组合方式不同，可能得到的同一种变换参数的有效区间也不同。因此我们同时综合各变换参数进行探讨。USPS 初始数据的大小为 16*16，在可考虑范围内，参数的变换范围应该较小。我们设定如表 5-1 的基准参数区间。

表 5-1 为各变换提供了基准尺度。各变换的原始变换参数都分布于基准参数范围内。由于变换的范围已经给定，我们需要从这些范围内选择具体的变换值来进行各变换操作。在实际值的选取中，如 5.3.5 节分析，我们使用高斯分布对变换参数进行采样。对于高斯分布函数，方差小则采样分布高而瘦，方差大则矮而胖，如图 5-2。均值为μ，方差为σ的高斯分布中，横轴区间(μ-1.96σ，μ+1.96σ)内的面积是 95.5%。因此当高斯分布是参数为(0,1)的标准正态分布时，位于区间(-2,2)之间的面积是 95.5%。对于基准变换参数范围，放大 1.1 倍的概率和旋转 10°的概率相同，故不能直接对这两个尺度范围内的数值进行高斯采样。我们统一在(-2,2)的区间范围内进行高斯采样，然后乘以各自基变换最大值作为实际变换参数。

表 5-1 USPS 基变换尺度

变换名称	变换范围	类型
X 平移变换	{-2,-1 0,1,2}	整型
Y 平移变换	{-2,-1 0,1,2}	整型
旋转变换	[-10°,10°]	实数
放缩变换	[0.9,1.1]	实数
剪切变换	[-10°,10°]	实数

表 5-2 不同高斯采样参数 USPS 分类准确率(%)

高斯分布标准差	1 层	2 层	3 层
原始结果	97.78	98.35	98.90
σ =0.5	98.01	98.49	98.90
σ =1.0	98.49	**99.04**	**99.11**
σ =4.0	98.08	98.49	98.77

表 5-3 不同尺度倍数 USPS 分类准确率(%)

尺度比例系数	1 层	2 层	3 层
原始结果	97.78	98.35	98.90
p =0.5	98.15	98.77	98.77
p =1.0	**98.49**	**99.04**	**99.11**
p =1.5	97.72	98.31	98.77

高斯分布中不同的标准差对应了采样对各尺度参数选择的概率不同，我们研究不同标标准差的高斯采样对性能的影响。同时，确定标准差，研究变换参数值放大和缩小对模型性能的影响。实验中，我们选取标准差分别为 0.5、1 和 4 的高斯采样函数对变换参数进行采样，通过这些参数设定进行 USPS 数据集扩展，扩展样本数为原始样本数的 4 倍。实验参数设定方法同本文 4.6 节所介绍的相同。我们得到不同标准差的高斯采样所对应的分类结果如表 5-2。从表 5-2 可以看出，3 层时，σ =1.0 所对应的分类结果最好。σ =0.5 的结果较原始结果稍好，而σ =4.0 的实验结果甚至有少许下降。忽略 PCA 参数设置的影响，这种现象的出现验证了扩展方法的正确性。由于σ =0.5 时，高斯采样基本都集中在各变换参数的中心附近，而参数的中心对应于数据样本不变，因此此时扩展的样本相对于原始样本，并不能提供太多的额外有效信息，不能起到较好的扩展作用。σ =4.0 时，对于基准尺度区间内值的选取比较平均，造成变换尺度太大，使得有效数据减少，并且混入

表 5-4 直接扩展 USPS 数据的分类准确率(%)

扩展倍数	1 层	2 层	3 层
原始数据	97.78	98.35	98.90
1 倍扩展	98.01	98.53	98.97
2 倍扩展	98.56	98.97	99.04
4 倍扩展	98.42	98.90	**99.11**

表 5-5 直接扩展 MINST 数据的分类准确率(%)

扩展倍数	1 层	2 层	3 层
原始数据	97.76	98.15	98.26
1 倍扩展	97.97	98.19	98.32
2 倍扩展	97.97	98.24	98.33
4 倍扩展	**98.05**	**98.27**	**98.39**

表 5-6 切向量扩展 MNIST 数据的分类准确率(%)

扩展倍数	1 层	2 层	3 层
原始数据	97.76	98.15	98.26
1 倍扩展	97.97	98.27	98.40
2 倍扩展	97.97	98.29	98.45
4 倍扩展	**98.19**	**98.36**	**98.50**

了很多对模型有干扰的数据，因此，分类准确率可能会低。

固定高斯采样的标准差σ值为 1，我们对高斯采样得到的尺度进行放缩，即乘以一个比例因子 p，p 值取 0.5,1 和 1.5。则新变换参数值为原变换参数值的 0.5 倍，1 倍和 1.5 倍得到的结果如表 5-3。从表 5-3 可以看出，当 p 值为 1 时，扩展后的结果在 1 层、2 层、3 层下最优，且较原始数据得到的结果都有提升。而其他两组则不然。当 p 值为 0.5 时，变换的尺度较小，不能起到很好的扩充效果，提升效率有限，而且在选择 PCA 参数时，只是给出了一种合理的选择值，并不能保证最优，所以可能造成 p =0.5 时 3 层的结果较原始结果差。当 p =1.5 时，变换尺度太大，引入了较多的无效样本，使得分类结果降低。

5.4.2 扩展样本量对模型性能的影响

在不考虑时间的前提下，有效训练样本越多，模型在特征的抽象中应该表现的越好，这是因为当训练样本充满了整个模式的流形时，每一层特征学习得到的特征才更加准确有效。为验证模型的这一特性，本节将通过实验分析样本数据量

对模型学习特征效果的影响。

选取基于切向量扩展生成的 MNIST 数据集和本文提出的方法生成的 MNIST 和 USPS 数据集。直接图像变换扩展方式的参数使用表 5-1 的基准参数范围内的值，具体变换值通过高斯采样产生，高斯变换的标准差为 1，变换顺序随机生成。扩展数量是在原始数据集基础上再扩展 1 倍、2 倍和 4 倍。实验中参数设置和调试方法同本文 4.6 节介绍的一致，唯一需要变动的参数仍然是各层 PCA 值的设定。

表 5-4 显示了 USPS 数据集扩展不同倍数时分类准确率的变化。从表中可以看出，增加样本量后，2 层模型达到的分类结果明显提升。当数据量扩充 4 倍后，相比于原始数据，3 层模型所得到的分类错误率由 1.1%降低到 0.9%。结果的总体趋势仍旧可以说明样本扩充的有效性。

表 5-5 显示的是使用本文提出的扩展方式扩展 MNIST 数据集得到的结果，随着样本量的增加，分类准确总体呈提升趋势。表 5-6 显示的是使用切向量方法得到的结果，我们发现，扩充一倍数据集时，效果提升明显，随着扩充数据量的增多，分类准确率仍然提升，但提升度降低。对比表 5-5 和 5-6 中的实验结果，虽然使用本文提出的方法进行样本扩充后提升效果没有基于切向量的方法提升明显，但仍旧可以有效的提升模型的性能，这也验证了直接扩展手写体图像数据的有效性。

通过上述实验结果可以看出，本文提出的手写体图像数据扩充方法可以有效的对 MNIST 和 USPS 数据集进行扩展。当采用合适的高斯采样进行采样，并选取合适的采样范围，扩展生成的数据可以有效的的提升模型的学习性能，提高 USPS 和 MNIST 数据集的分类结果。

5.5 小结

本章提出了一种基于手写体图像数据的直接扩展方式，并研究了扩展参数对本文提出的级联深度学习模型性能的影响。通过对手写体图像数据的分析，实现了基于平移、旋转、剪切、放缩随机组合的数据扩展方式，采用高斯采样选取扩展参数。通过对高斯采样的标准差和各变换的尺度范围进行研究，得到了更合适的采样参数和尺度范围。通过对比本文提出的扩展方式和基于切向量的扩展方式的实验结果，验证了本文提出的手写体数字图像扩展方式的有效性，同时也通过使用数据的先验知识对数据样本进行扩展，成功提高了模型的学习性能。

结　论

深度学习是目前最主流的特征学习方法之一，已经在图像、语音等研究中取得了突破性的进展。然而，多数深度学习模型仍旧只针对具有二维空间语义信息的问题，对于一般的向量特征普适性较弱。因此，本文提出一种面向普通向量特征的结构简单的通用级联深度学习框架，并对框架中使用到的特征提取和特征选择方法进行了研究与分析，最终实现了一个有效的级联深度学习模型。本文的主要工作及结论如下：

（1）提出了一种非凸正则化的非监督选择模型，有效解决了特征空间相互冗余的问题。针对模型中出现的非凸问题，提出了有效的迭代求解算法。在大量人脸和微阵列数据集上的分类和聚类实验结果表明，本模型可以有效的进行特征选择，显著提升分类和聚类效果。

（2）实现了一个通用性较强的级联深度学习模型，有效解决普通向量数据的特征学习问题。基于深度结构的不同层特征提供信息抽象程度不同，提出一种有效的组合模型各层特征的方法。在普通属性值数据和图像数据集上的实验结果都表明本模型可以学习到更有效的特征信息，显著提高分类准确率。

（3）为使用数据的先验知识，提出了一种简单却有效的手写体数字图像扩展方式。成功的实现了数据增强的作用，提升了模型的学习性能。同时也为其他领域的图像数据增强提供了一定的借鉴意义。

本文构建了一种成功的级联深度学习模型，模型中特征提取和特征选择方法相独立，因此，未来工作中，我们可以考虑分析不同的特征提取方法和特征选择方法对模型的影响，从而进一步增强模型的学习能力。同时，我们也可以完善提出的级联深度框架，为不同领域的学习任务提供统一的接口，方便大家使用不同的特征变换或特征选择算法对模型进行扩展。

参考文献

[1] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[2] Arnold L, Rebecchi S, Chevallier S, et al. An Introduction to Deep Learning[C]. ESANN, 2011.

[3] Hubel D H, Wiesel T N. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex[J]. The Journal of physiology, 1962, 160(1): 106-116.

[4] Hinton G E. Learning Distributed Representations of Concepts[C]. Proceedings of the eighth annual conference of the cognitive science society, 1986, 1: 12-20.

[5] Bengio Y, LeCun Y. Scaling Learning Algorithms Towards Ai[J]. Large-scale kernel machines, 2007, 34(5): 203-218.

[6] Bengio Y, Delalleau O, Roux N L. The Curse of Highly Variable Functions for Local Kernel Machines[C]. Advances in neural information processing systems, 2005: 107-114.

[7] Bengio Y, Lamblin P, Popovici D, et al. Greedy Layer-Wise Training of Deep Networks[J]. Advances in neural information processing systems, 2007, 19: 153-167.

[8] Hinton G E, Osindero S, Teh Y-W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[9] LeCun Y, Bengio Y. Convolutional Networks for Images, Speech, and Time Series[J]. The handbook of brain theory and neural networks, 1995, 3361(10): 257-269.

[10] Lange S, Riedmiller M. Deep Auto-Encoder Neural Networks in Reinforcement Learning[C]. The 2010 International Joint Conference on Neural Networks (IJCNN), 2010: 1-8.

[11] Peng Z, Lin L, Zhang R, et al. Deep Boosting: Layered Feature Mining for General Image Classification[C]. 2014 IEEE International Conference on Multimedia and Expo (ICME), 2014: 1-6.

[12] Simonyan K, Vedaldi A, Zisserman A. Deep Fisher Networks for Large-Scale Image Classification[C]. Advances in neural information processing systems, 2013: 163-171.

[13] Chan T-H, Jia K, Gao S, et al. Pcanet: A Simple Deep Learning Baseline for Image Classification?[J]. arXiv preprint arXiv:14043606, 2014.

[14] Scholkopft B, Mullert K-R. Fisher Discriminant Analysis with Kernels[J]. Neural networks for signal processing IX, 1999: 71-83.

[15] Weldon T P, Higgins W E, Dunn D F. Efficient Gabor Filter Design for Texture Segmentation[J]. Pattern Recognition, 1996, 29(12): 2005-2015.

[16] Ke Y, Sukthankar R. Pca-Sift: A More Distinctive Representation for Local Image Descriptors[C]. Computer Vision and Pattern Recognition, 2004, 502(2): 506-513.

[17] Zhu Q, Yeh M-C, Cheng K-T, et al. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients[C]. Computer Vision and Pattern Recognition, 2006, 2: 1491-1498.

[18] Langley P. Selection of Relevant Features in Machine Learning[M]. Defense Technical Information Center, 1994: 448-456.
[19] Kohavi R, John G H. Wrappers for Feature Subset Selection[J]. Artificial intelligence, 1997, 97(1): 273-324.
[20] Vapnik V. The Nature of Statistical Learning Theory[M]. Springer Science & Business Media, 2013: 336-341.
[21] Zhu J, Rosset S, Hastie T, et al. 1-Norm Support Vector Machines[J]. Advances in neural information processing systems, 2004, 16(1): 49-56.
[22] Hou C, Nie F, Yi D, et al. Feature Selection Via Joint Embedding Learning and Sparse Regression[C]. IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 2011, 22(1): 1324-1340.
[23] Dahl G, Mohamed A-r, Hinton G E. Phone Recognition with the Mean-Covariance Restricted Boltzmann Machine[C]. Advances in neural information processing systems, 2010: 469-477.
[24] Deng L, Seltzer M L, Yu D, et al. Binary Coding of Speech Spectrograms Using a Deep Auto-Encoder[C]. Interspeech, 2010: 1692-1695.
[25] Seide F, Li G, Yu D. Conversational Speech Transcription Using Context-Dependent Deep Neural Networks[C]. Interspeech, 2011: 437-440.
[26] Mohamed A-r, Dahl G E, Hinton G. Acoustic Modeling Using Deep Belief Networks[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.
[27] Dahl G E, Yu D, Deng L, et al. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42.
[28] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. Signal Processing Magazine, IEEE, 2012, 29(6): 82-97.
[29] Ciresan D, Meier U, Schmidhuber J. Multi-Column Deep Neural Networks for Image Classification[C]. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012: 3642-3649.
[30] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.
[31] Mikolov T, Deoras A, Kombrink S, et al. Empirical Evaluation and Combination of Advanced Language Modeling Techniques[C]. INTERSPEECH, 2011: 605-608.
[32] Wold S, Esbensen K, Geladi P. Principal Component Analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1): 37-52.
[33] Livni R, Lehavi D, Schein S, et al. Vanishing Component Analysis[C]. Proceedings of The 30th International Conference on Machine Learning, 2013: 597-605.
[34] Friedman J, Hastie T, Tibshirani R. Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors)[J]. The annals of statistics, 2000, 28(2): 337-407.
[35] Freund Y, Schapire R E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting[J]. Journal of computer and system sciences, 1997, 55(1): 119-139.
[36] Breiman L. Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting: Discussion[J]. Annals of Statistics, 2000: 374-377.
[37] Bengio Y. Deep Learning of Representations for Unsupervised and Transfer

Learning[J]. Unsupervised and Transfer Learning Challenges in Machine Learning, 2012, 7: 19-36.
[38] Hastie T, Buja A, Tibshirani R. Penalized Discriminant Analysis[J]. The Annals of Statistics, 1995: 73-102.
[39] Schapire R E, Singer Y. Improved Boosting Algorithms Using Confidence-Rated Predictions[J]. Machine learning, 1999, 37(3): 297-336.
[40] Zhao Z, Wang L, Liu H. Efficient Spectral Feature Selection with Minimum Redundancy[C]. AAAI, 2010.
[41] Zhu P, Zuo W, Zhang L, et al. Unsupervised Feature Selection by Regularized Self-Representation[J]. Pattern Recognition, 2015, 48(2): 438-446.
[42] El-Shaarawi A H, Piegorsch W W. Encyclopedia of Environmetrics[M]. John Wiley & Sons, 2002:337-341.
[43] He X, Cai D, Niyogi P. Laplacian Score for Feature Selection[C]. Advances in neural information processing systems, 2005: 507-514.
[44] Yang Y, Shen H T, Ma Z, et al. L2, 1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning[C]. IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 2011, 22(1): 1589-1596.
[45] Zhao Z, Liu H. Spectral Feature Selection for Supervised and Unsupervised Learning[C]. Proceedings of the 24th international conference on Machine learning, 2007: 1151-1157.
[46] Estévez P, Tesmer M, Perez C, et al. Normalized Mutual Information Feature Selection[J]. Neural Networks, IEEE Transactions on, 2009, 20(2): 189-201.
[47] Simard P, Victorri B, LeCun Y, et al. Tangent Prop-a Formalism for Specifying Selected Invariances in an Adaptive Network[C]. Advances in neural information processing systems, 1992: 895-903.

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

[1] **Weizhi Wang**, Hongzhi Zhang, Pengfei Zhu, David Zhang, Wangmeng Zuo. Non-convex Regularized Self-representation for Unsupervised Feature Selection[C]. International Conference on Intelligence Science and Big Data Engineering, 2015. 已录用（Oral，比例：18/416）。

（三）参与的科研项目及获奖情况

[1] 左旺孟. 舌脉合参中信号和特征的约简与协同分析方法研究, 国家自然科学基金研究项目. 课题编号：61271093。

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于三元组约束的距离度量学习方法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：王缉智　　日期：2015年 6月 30日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：王缉智　　日期：2015年 6月 30日

导师签名：张利彪　　日期：2015年 6月 30日

致　谢

硕士研究生活即将结束，也宣告着自己的学生生涯的结束。短暂的硕士学习与生活使我成长很多，毕业设计这段时间也受到了很多人的帮助，至此，谨以此毕设论文献给所有帮助我的人，并向他们表示真挚的感谢。

首先感谢我的导师张大鹏教授，对我的研究工作提供了宝贵的意见与建议。张老师也为我提供了到香港理工学习与交流的机会，使我接触到了国际前沿的研究内容和学术氛围，开阔了我的视野，锻炼了能力。感谢王宽全教授在此期间对我的帮助，在做毕设的过程中，王老师对我的研究内容和研究思路提出了宝贵的意见与建议。

特别感谢左旺孟老师三年来对我悉心的教导和无私的帮助。学业上，左老师从研究方向、方法上都给予了我极大的指导。生活中，左老师为我考虑了很多，并还忍受着我各种各样的缺点，一直不离不弃的帮我改正。在工作态度和方法上，左老师一直是我的榜样，并将在以后的日子也深深的影响我。特别感谢张宏志老师，张老师为我的研究内容提出了中肯的建议和意见，也帮我改掉了科研学习中的很多缺点，并教会我要要注意细节。生活中，三年来张老师也一直对我关心有加，第一时间帮我解决困难。

感谢实验中心的邬向前老师、马琳老师和袁永峰老师，感谢他们在硕士期间对我的照顾，同时也感谢各位老师在开题、中期时的意见和建议。感谢同窗罗长春六年来的帮助和鼓励，感谢言宏亮、任东伟、张玮琦、潘峰、朱园园、党芸琪、邓红、李峰、杨伟、王鹏、崔振超、张凯等师兄弟姐妹们在我完成论文和研究的过程中提供的无私的帮助，同时也感谢实验室各位兄弟姐妹们两年来的照顾和帮助。感谢同级的陈丽、武小荷、王保全两年来的共同成长。感谢各位审稿老师和评委会老师。

最后感谢我的父亲母亲，感谢其他家人，是他们的支持与鼓励让我一直坚定的做着自己喜欢的事。