中图分类号：TP391 论文编号：1028716 19-S020
学科分类号：083500

硕士学位论文

标号噪声下图像数据的清洗和特征学习

研究生姓名 张魏宁
学科、专业 计算机科学与技术
研 究 方 向 计算机视觉
指 导 教 师 谭晓阳 教授

南京航空航天大学

研究生院 计算机科学与技术学院

二 O 一九年三月

Nanjing University of Aeronautics and Astronautics
The Graduate School
College of Computer Science and Technology

Image Data Cleaning and Feature Learning in the Presence of Label Noise

A Thesis in

Computer Science and Technology

By

Weining Zhang

Advised by

Prof. Xiaoyang Tan

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Engineering

March, 2019

承诺书

本人声明所呈交的硕士学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京航空航天大学或其他教育机构的学位或证书而使用过的材料。

本人授权南京航空航天大学可以将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、汇编学位论文。

（保密的学位论文在解密后适用本承诺书）

作者签名：_____

日　　期：_____

摘 要

在机器学习和深度学习等大数据技术在诸多应用领域中广泛应用的背景下，一个拥有可靠标记的大规模数据集是进行有监督学习任务的基础和必备条件。然而，在实际应用场景中收集到的数据往往存在一定程度的标号噪声现象，即数据的标号被错误地标记。针对计算机视觉领域中图像数据的标号噪声问题，本文对图像数据的清洗和鲁棒的特征学习方法展开研究，具体的工作和创新点如下：

（1）详细介绍了标号噪声的相关概念及多种常见的处理技术，并介绍了深度自编码网络和生成对抗网络的基本理论方法，为后续模型框架的设计提供参考。

（2）针对传统的标号噪声检测方法导致较高的假阳性率而降低了预测性能，提出了基于异常检测技术和重建误差最小化的数据清洗模型。通过异常检测技术获得候选的标号噪声数据，并根据重建误差最小化准则来进一步选出真正的标号噪声数据。

（3）针对因标号噪声的存在影响特征学习过程，提出了鲁棒的类专属自编码网络特征学习框架。具体地，该框架包含三个模块，分别是基于生成对抗网络的数据增广策略、基于重要性加权的优化策略以及基于最小重建误差的重标记迭代策略。通过大量的验证性实验表明三个策略均可以在一定程度上降低标号噪声对特征学习的影响。

（4）在 MNIST 手写数字数据集以及 Caltech-10 图像数据集中，将所提出的相关模型与最先进的数据清洗模型和标号噪声鲁棒的模型进行对比分析，分别在训练集上的数据清洗任务和测试集上的分类任务验证了所提出模型方法的有效性。

关键词：标号噪声，数据清洗，鲁棒特征学习，深度自编码网络，重建误差

ABSTRACT

Under the background that big data technologies such as machine learning and deep learning are widely used in many application fields, a large data set with reliable label is the basic and necessary condition for the supervised learning task. However, data collected in practical application scenarios often have a certain degree of label noise, that is, the label of data is wrongly labeled. Aiming at the problem of label noise of image data in the field of computer vision, this paper conducts research on image data cleaning and robust feature learning methods, and the specific work and innovation points are as follows:

(1) Related concepts of label noise and various common processing techniques are introduced and analyzed in detail, and the basic theoretical methods of deep autoencoder network and generative adversarial networks are introduced.

(2) A data cleaning model based on anomaly detection technology and reconstruction error minimization is proposed to reduce the prediction performance due to high false positive rate caused by the traditional label noise detection method. The candidate label noise data are obtained by using the anomaly detection technique, and the true label noise data is further selected according to the reconstruction error minimization criterion.

(3) In view of the influence of label noise on feature learning, a robust class-specific autoencoder based feature learning framework is proposed. Specifically, the framework contains three modules, which are respectively the data augmentation strategy based on generative adversarial networks, the optimization strategy based on importance weighting and the iteration strategy based on the minimum reconstruction error. A large number of verification experiments show that all three strategies can reduce the influence of label noise on feature learning to some extent.

(4) In the MNIST handwritten digital dataset and caltech-10 image dataset, the proposed correlation models were compared and analyzed with the state-of-the-art data cleaning model and the label noise robust model. Moreover, the proposed models were validated by the data cleaning task on the training set and the classification task on the test set, respectively.

Keywords: Label noise, Data cleaning, Robust feature learning, Deep autoencoder, Reconstruction error

目 录

图表清单

注释表

N	数据集中训练样本的个数	X	数据样本集合
x_{i}	数据集中第 i 个样本	d	样本的维度
$\hat{x}_{\mathrm{i}}$	样本 x_{i} 的重建结果	W	模型的权重项
b	模型的偏置项	Z	随机噪声变量
C	样本的总类别数	λ	损失函数的平衡因子
θ	模型参数集合	α	梯度下降学习率
h	隐含层的特征编码	k	隐含层节点数
$E_\theta(\cdot)$	编码函数	$D_\theta(\cdot)$	解码函数
$L(\cdot)$	目标函数	$f(\cdot)$	激活函数
S	异常数据样本集	ε	优化算法收敛阈值

缩略词

缩略词	英文全称
NCAR	Noisy Completely At Random
NAR	Noisy At Random
NNAR	Noisy Not At Random
AE	Autoencoder
dAE	denosing Autoencoder
DAE	Deep Autoencoder
SAE	Sparse Autoencoder
VAE	Variational Autoencoder
RDA	Robust Deep Autoencoder
GAN	Generative Adversarial Networks
CGAN	Conditional Generative Adversarial Networks
AC-GAN	Auxiliary Classifier Generative Adversarial Networks
InfoGAN	Information Maximizing Generative Adversarial Networks
LOF	Local Outlier Factor
SVM	Support Vector Machine
OC-SVM	One Class Support Vector Machine
TC-SVM	Two Class Support Vector Machine
FSVM	Fuzzy Support Vector Machine
REM	Reconstruction Error Minimization
ADMM	Alternating Direction Method of Multipliers
PG	Proximal Gradient
MNIST	Modified National Institute of Standards and Technology
iForest	Isolation forest
NN	Neural Network
CNN	Convolutional Neural Network
BP	Back Propagation
DT	Decision Tree
KNN	K Nearest Neighbor
OC-NN	One Class Neural Network

TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
LN	Label Noise
RBF	Radial Basis Function

第一章 绪论

1.1 研究背景

互联网的普及和计算机技术的快速发展，使得网络上和用户设备中产生了呈爆炸式增长的数据量、信息量。在当今大数据时代，掌握某领域中的海量数据成为了解该领域最基础的一个环节。基于所获得的大规模领域数据，机器学习和深度学习等数据挖掘技术在诸多研究问题中显现出相当优良的效果，其中包括图像分类[1]、文本检测[2]、语音识别[3]等。特别地，对于监督学习任务而言，数据集中的数据所对应的标号准确与否是能否较好完成这些任务的关键。只有确保有足够数量、标记可靠的训练数据集，才能学得具有高精度的训练模型，并在实际应用场景中给出准确的决策和预测结果。

图 1.1 搜图过程中的标号噪声现象

一般而言，收集大规模、标记可靠的训练数据需要付出非常昂贵的人力、物力和时间代价。因此，当今很多研究机构选择一些替代的、简便的方法去收集数据，如众包[4]、网络爬虫[5]等方法。然而，通过这些方法收集到的数据集往往会导致一定程度的标号噪声，也就是一些标号被错误地标记。图 1.1 显示了通过 Google 浏览器搜索关键词“horse face”获取图片时产生的标号噪声现象，其中含标号噪声的图像由红色矩形框框出。另一方面，标注过程本身具有主观性，在很多具体领域中时不可避免的，如医学诊断[6]等。可以看出，标号噪声是收集数据时普遍存在的现象。目前对数据集错误标号的检测主要还是借助人力的方式，通过领域专家或众包的方式来一次次地更新数据集。该方式不仅效率低下、耗费昂贵，而且极易产生新的标号错误。在具体的算法和应用当中，错误的标注信息不仅会带来预测性能的损失，而且可能会由于不安全操作带来的安全隐患问题。由此可见，如何通过有效的方法对数据集中标号噪声进行清洗是一个基础且亟需解决的问题。

1.2 研究的目的和意义

特别地，对于图像分类任务来说，要想得到较高分类精度的模型往往需要学习到准确的图像特征表示，这就需要大规模、多样化和标记正确的图像数据集作为基础。一个有效的图像数据集往往由数万张乃至百万张具有可靠标记的图像组成。例如斯坦福大学李飞飞团队耗时三年收集的 ImageNet[7]图像数据集，共包含 1400 万张图片并涵盖 2 万多个物体类别。大量的物体分类、检测算法均视其为标准集并在该数据集上进行实验来验证算法的有效性。这不仅反应了研究人员对可靠标记、大规模的数据集的需求，另一方面也促进了深度学习技术乃至计算机视觉领域研究的发展。该数据集的诞生以及产生的长远影响，让越来越多的学术研究者和工业界的从业人员意识到，一个标记良好的大规模数据集与一个有效的算法同样重要。在可靠的图像数据集下训练的机器学习、深度学习等模型，可以给出更加准确的判断和决策，从而进一步解决实际应用中的问题。

本课题的研究目的是解决当大规模图像数据集中因存在标号噪声现象而对训练模型带来一系列问题时，如何在不借助领域专家检查的条件下，通过设计、改进学习算法，利用软件的方式来自动地减少标号噪声带来的影响。具体地，一方面从异常检测技术的角度，找出并纠正标号噪声的图像；另一方面，从模型本身的角度，通过改进模型的学习算法、设计鲁棒的目标函数来缓解标号噪声对特征学习过程的影响。

本课题的研究意义在于：

（1）通过设计图像数据过滤、纠正算法，降低人工手动对标号修正的依赖，用软件的方式自动地清洗数据集中的标号噪声数据。

（2）通过设计噪声鲁棒、容忍的学习算法，自适应、高效的缓解标号噪声对特征学习过程带来的影响，提高模型的训练精度。

（3）利用所设计的新型算法解决图像数据集的质量问题，并使数据集更好的服务于计算机视觉等领域的研究，进一步提高智能决策系统的广泛应用。

1.3 标号噪声问题

标号噪声问题一直以来都是研究的基础且热点问题，本节从标号噪声问题提出及其相关概念的角度来简洁阐述标号噪声的基本思想。

1.3.1 问题的提出

一个大规模真实世界的数据集的质量取决于很多方面[8]，但是数据的来源是最重要的因素。数据的输入和获取是很容易出错的，研究人员关于减少数据输入过程中的错误这一问题进行了很多努力。然而，在大规模数据集中出现错误仍是普遍且严重的，除非有很严格的检查和控制，

数据集中的错误率通常在 5%以上[9]。其中，在对数据集进行标注过程中出现的错误是最常见、也是最重要的问题[10]。

1.3.2 标号噪声的相关概念

在标号噪声的相关概念这一方面，本节重点围绕标号噪声的定义、来源、分类、生成、影响以及性能评价测试方面展开详细介绍。

1.3.2.1 标号噪声的定义

通常，数据集中每个样本都关联一个观察标号 y_{obs}，我们假设这个标号对应样本真正的类别 y_{true}。但是，由于某些原因它可能在输入到算法之前就受到噪声的影响[4]，使得 $y_{\text{obs}} \neq y_{\text{true}}$。我们称标号被污染的过程为标号噪声。标号噪声是一个常见却复杂的现象，它广泛存在我们收集到的数据集中，因此，根据观察到的标号推测出真正的标号是非常重要的。

1.3.2.2 标号噪声的来源

数据集中的标号错误通常发生在有人类专家涉及的时候。在这种情况下，标号噪声可能的原因包括不完美的证据、混淆了感兴趣的模式、感知错误等[11]。更一般的，标号噪声的来源主要分为以下四个方面：

（1）在标记过程中，提供给标记标号者的信息是无效的[12]。比如，在医学领域中，一些测试的结果是不知道的[6]；标号描述语言的限制导致减少了可利用的信息[13]；信息质量的参差不齐等[14]。

（2）标号错误可能出现在标记过程本身。由于收集具有可靠标号的大量数据集是一个耗时费力的任务，有很多数据集是通过非专业的方式获得的，例如借助 crowdsourcing[15]、Amazon Mechanical Turk[16]和网络爬虫等方法。在这些方法的标记过程中，往往会存在一些标号出现错误。

（3）当标记任务是主观的时候，往往会出现标号错误。比如，医学专家对疾病做出诊断，这个诊断通常是根据部分检测数据和医生自身的知识得出的，并不是 100% 准确的。所以，有部分标号的错误是因为这个标号的推断过程并不是完全准确和可信的。而且在标注的过程中会出现由于标注者操作失误或知识的片面性带来标注错误的现象。有调查研究表明，大规模数据集中的标注错误问题不仅普遍而且严重，错误率一般都在 5% 左右甚至更高。

（4）标号噪声还可能来自于简单的数据编码或沟通错误。例如，在垃圾邮件过滤中，标号噪声的来源包括误解了反馈机制和意外点击[17]。

1.3.2.3 标号噪声的分类

从统计的角度来分类，标号噪声可以分为三种，如图 1.2 所示，箭头反映变量与变量之间的依赖关系。对标号噪声过程的建模，使用了四个随机变量进行描述。X 是输入特征向量，Y 是真正的类标号，$\tilde{Y}$ 是观察到的类标号，E 是一个二元变量反应是否有标号错误发生。其中 $\tilde{Y}$ 总是依赖于 Y （即观察到的标号总是依赖于真实的标号）。

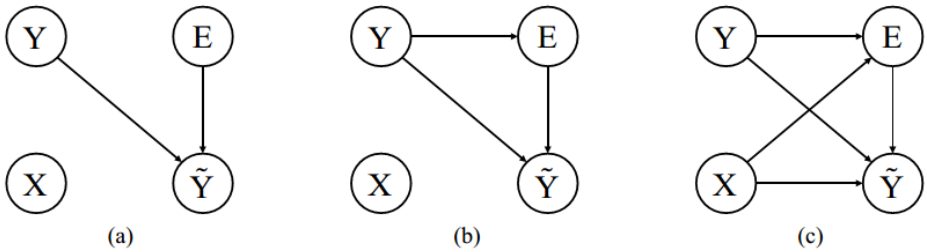


图 1.2 常见的三种标号噪声

（1）完全随机噪声（noisy completely at random, NCAR）

如图 1.2(a)所示，在这种情况下，标号噪声独立于输入数据和真实的标号，即完全随机的噪声。在很多相关工作中，都是使用人工设计的完全随机噪声来进行实验的。

（2）随机噪声（noisy at random, NAR）

如图 1.2(b)所示，在随机噪声中，标号的错误情况依赖于真正的类标号。虽然 E 仍然独立于 X，但是这种方式可以建立不对称的标号噪声，也就是说适用于当某些类的实例更容易出现时贴错了标号的情况。例如，在医学病例对照研究中，控制对象可能更容易被贴上标号。事实上, 用于标记案例主题的测试可能过于具有侵略性（例如，活检）或代价昂贵以至于不能用于控制对象，因此，由最优诊断测试代替控制主题[18]。

（3）非随机噪声（noisy not at random, NNAR）

很多在标号噪声的研究中认为标号噪声在所有数据中是没有区别的，然而，上述两类标号噪声是不现实的[19]。如图 1.2(c)所示，标号的错误情况也取决于数据本身的情况。例如，当某类样本与其他类样本更相似的时候（比如，MNIST 手写数字数据集中的数字 7 和数字 9），样本是更容易被标记错误的[20]。在非随机噪声中， X 会影响标号错误 E 和观测标号 $\tilde{Y}$ 。

1.3.2.4 标号噪声的生成

有很少的数据集中包含识别到的标号噪声，因此人工标号噪声在研究工作中更为常见。很多研究工作是建立在完全随机噪声（NCAR）上，通常，按照下述步骤为真实数据集添加完全随机类噪声：（1）随机选择一些数据（2）随机选择不同于当前标号的其他标号作为它的新的标号。利用上述方式，标号噪声是独立于真实标号 Y 。还有一些研究引入随机噪声（NCR），例如，Nettleton 等人[21]随机改变主要类别的标号，Pechenizkiy 等人[6]针对手写数据集中的视觉混

淆的类别之间添加随机噪声。对于非随机噪声而言，相对前两者的工作较少，Hanner 等人[22]通过实验分析得出噪声标号的概率分布情况往往取决于数据与分类器边界的距离长度。

在实践中，获得包含能够清楚识别错误实例的真实数据集是很有趣的。另一个重要的开放性研究问题是寻找真实标号噪声的特征。事实上，在目前的研究中，我们不知道各类标号噪声到底哪个是更加贴近现实的。

1.3.2.5 标号噪声的影响

标号噪声潜在后果的分析是是否有必要去考虑标号噪声现象的前提。一些理论和实验证据反映了标号噪声直接影响分类性能，这是最常被报道的问题，也是研究得最多的问题。Nettleton 等人[21]分别在四个分类器比较标号噪声的影响，分别是朴素贝叶斯、决策树、kNN 以及 SVM 分类器。其中基于贝叶斯的分类模型取得了最好的分类精度，取决于条件独立的假设和条件概率的使用。Freund 等人[23]表明 boosting 方法也受标号噪声影响，特别是适应性 boosting 算法（AdaBoost）倾向于花费过多的精力学习错误的实例。

也有研究表明，标号噪声的存在增加了学习算法所需要样本的数量，也将增加模型的复杂性。Quinlan 等人[24]验证了标号噪声的引入会使得决策树的大小增加。Brodley 等人[25][26]通过去除含标号噪声的样本，使得 SVM 的复杂度降低，即支持向量的数目。

另一方面，标号噪声也可能对其他相关任务造成威胁，如类频率估计和特征选择等相关问题。在类频率估计问题上，Bross 等人[27]在二类测试中发现错误的选择可能会造成严重的威胁：观察到的测试答案的平均值和方差会因为标号噪声的存在而受到很严重的影响。类似的问题也存在多类别情况中[28]。在特征选择方面上，Zhang 等人[29]在实验中发现标号噪声的引入严重影响微阵列数据的特征选择。相似地，Shanab 等人[30]发现了标号噪声可以降低特征排名的稳定性。

尽管标号噪声的存在导致很多负面后果，但人工标号噪声也有潜在的优势。例如，可以在统计研究中加入标号噪声来保护人们的隐私。其中，Hout 等人[28]在获取统计数据的问卷调查中，通过加入标号噪声使得个人的答案不会得到回复。

上述研究说明了标号噪声后果的重要性和多样性：分类性能的下降，学习要求的改变，增加了复杂性学习模型，观察频率失真，识别相关特征的难度增加等等。标号噪声带来的后果往往取决于噪声的类型和等级。因此，对于机器学习研究员来说，处理标号噪声并且考虑到这些因素对于处理污染数据是很重要的。

1.3.2.6 性能评价与测试

在处理标号噪声问题中，一个重要的问题是如何证明所提出算法的有效性。对标号噪声不同的处理方式，有不同的评价准则。一般地，一个好的方法满足以下两个条件中的一条：（1）如果能在标号噪声的背景下，保持了原有的评价性能；（2）相比其他标号噪声算法提高了相应

的评价性能。在已有的研究中，大部分的实验通过验证分类准确率来证明所提出方法的有效性[15][25]。这是因为标号噪声的引入最直接的影响就是分类精度的下降。

其他比较常用的准则是模型的复杂度[31][32],比如决策树的节点，归纳逻辑的规则数等。事实上，一些针对标号噪声的推理算法趋向于过拟合，这导致了过度复杂的模型。在一些情况下，模型本身的估计参数也很重要，这些方法往往专注于从观察频率去估计真实频率[33]。

还有一大类评价方法是针对数据清洗的效果[10]。不同的度量方法可以用图 1.3 来统一解释，其中主要分为两种错误，错误类型 1 反应正确标号的实例被错误地选中情况，相关度量见公式（1.1）。错误类型 2 反应错误标号的实例没有被选中的情况，度量公式如公式（1.2）。还有一类评价准则可以用上述两条来计算出来，也就是在选中的实例中真正包含错误标号的实例的比例，度量公式如公式（1.3）。

$$\mathrm{ER}_1=\frac{\text{正确标记的实例被选中的个数}}{\text{所有正确标记的实例的个数}} \tag{1.1}$$

$$\mathrm{ER}_2=\frac{\text{错误标记的实例没有被选中的个数}}{\text{所有错误标记的实例的个数}} \tag{1.2}$$

$$\mathrm{NEP}=\frac{\text{错误标记的实例被选中的个数}}{\text{所有选中的实例的个数}} \tag{1.3}$$

一个好的数据清洗方法会在 ER_1、ER_2 和 NEP 中找到一个折中。在一方面，保守的过滤器只会将其认为可能性极大的少数实例视为标号噪声，因此 ER_1 错误率较小，NEP 的精度较大，但是它们倾向于保留大多数错误的实例，所以 ER_2 错误率较大。因此，用这类过滤器清洗后的数据集可能不会有一个很高的分类精度。另一方面，有些过滤器为了提高分类的精度会选中大量的实例视为标号噪声，此时对应的 ER_2 错误率较小，但是这些方法往往也会选中过多的正确标记的实例，也就是 ER_1 的错误率较大，而 NEP 精度较小。

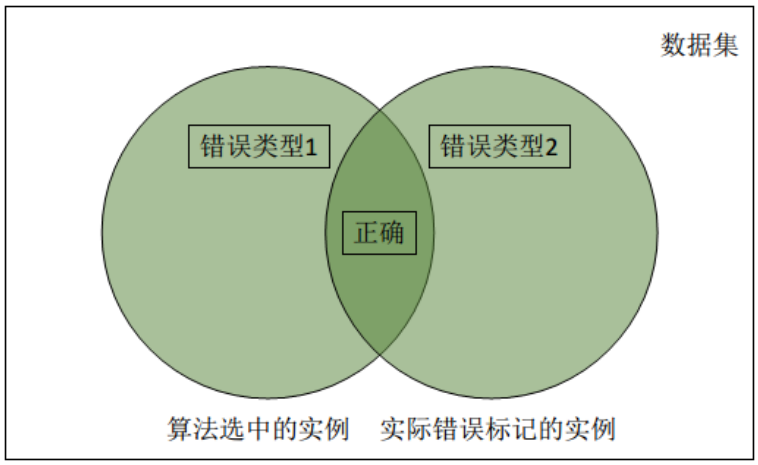


图 1.3 数据清洗的几种错误类型

1.4 本文的主要研究工作

本文以标号噪声下的图像分类问题作为研究对象，重点研究标号噪声的检测、纠正问题以及在特征学习过程中的鲁棒性问题。具体地，结合异常检测技术和最小判别误差思想，发现并纠正错误的标号，建立高效的数据清洗模型，并利用数据增广技术和改进的自编码网络学习可靠的类内特征表示，建立标号噪声下的鲁棒的特征学习框架。

针对标号噪声检测、纠正问题，我们将该问题转化为每一类数据的异常检测问题。通过这种分而治之的思想，可以降低标号噪声检测的难度。针对异常检测算法中假阳性样本过多的问题，提出组间最小判别误差算法。根据异常检测算法选出候选标号噪声数据，利用组间最小判别误差算法进一步找出真正的标号噪声数据并重新标记其标号，达到自动的数据清洗的目的。在 MNIST 手写数字图像数据集进行的实验表明，该方法较其他方法有更好的清洗准确率和效率。

针对标号噪声的环境下，自编码网络无法获得准确的特征表示，从而影响分类精度的问题，提出了基于数据增广和鲁棒的自编码网络的特征学习模型。首先提出了基于信息最大化生成对抗网络的有效样本生成算法，扩大数据集；接着提出基于重要度加权的特征学习算法，对每类数据学习一个良好的、可靠的自编码特征表示。由于图像数据集中包含一些标号噪声的数据，在对自编码网络进行梯度下降的过程中，根据重建误差大小计算每个数据的重要度，提高正确标记的数据的重要程度，降低潜在标号噪声数据的重要程度，从而降低标号噪声对特征学习的影响，学习更加可靠的特征表示。最后，整合两个模块提出类专属自编码网络的特征学习框架，在 MNIST 手写数字图像数据集和 Caltech-10 物体图像数据集的实验表明，该方法获得的特征表示可以更好的提高分类精度。

1.5 本文的内容安排

本文共由五章组成，具体的结构安排如下：

第一章：绪论。主要介绍了课题的研究背景、研究目的意义以及标号噪声问题的提出与相关概念的介绍，简要描述了本文的主要研究工作，并给出了本文的章节结构安排。

第二章：相关技术的介绍。首先对常见经典的标号噪声处理技术进行了介绍，重点包括：鲁棒的学习算法、数据过滤、纠正算法以及标号噪声容忍算法。之后概述了课题的主要研究对象自编码网络及其深度模型的介绍。最后介绍了生成对抗网络及其扩展模型的相关原理。

第三章：基于异常检测技术和组间最小重构误差算法的数据清洗。首先介绍了整个数据清洗框架的思路。之后分别介绍所使用的异常检测技术和组间最小重建误差算法，最后介绍实验中选用的数据集和评价方法，对提出方法的有效性进行实验验证分析，并与其他数据清洗算法进行对比。

第四章：基于数据增广和鲁棒的自编码网络的特征学习模型。首先详细介绍了基于生成对抗网络的数据增广技术，随后提出了基于重要度加权算法的自编码网络。最后，通过验证性实验验证所提出算法的有效性，并对数据集清洗、分类任务的性能进行了对比实验。

第五章：总结与展望。主要对本文的相关研究工作进行了详细地总结，同时指出未来研究中需要关注的方面。

第二章 相关技术介绍

2.1 标号噪声处理技术

数据集中的标号噪声问题一直是机器学习和模式识别领域中所关注的重点问题之一。在处理标号噪声的相关研究文献中，大致可以分为三类方法：（1）依靠算法本身的鲁棒性来缓解标号噪声，（2）过滤、纠正含有标号噪声的数据，以及（3）将标号噪声考虑在内的噪声容忍性算法。

2.1.1 鲁棒的学习算法

这类方法通常是依靠算法对标号噪声天然的鲁棒性来缓解标号噪声的影响。换句话说，这些算法不是为标号噪声而特别设计的算法，但是，通过鲁棒的学习算法所得到的分类器是对标号噪声不太敏感的。事实上，一些研究已经表明了在标号噪声较小的情况下，这些学习算法是有效果的。

在鲁棒的损失函数方面，Manwani 等人[34]提出不管标签噪声存在与否，如果推断模型的错误分类的概率是相同的，在给定的损失函数下的风险最小化被定义为标签噪声鲁棒性。同时，Thathachar 等人[35]证明了 0-1 损失对于均匀分布的标号噪声是鲁棒的，Sastry 等人[36]证明了当零错误率是可能的时候，0-1 损失也是对标号噪声鲁棒的。Beigman 等人[37]讨论了在 NNAR 标号噪声类型下的情况。对于其他常见的损失函数来说，如指数损失、log 损失、hinge 损失等，均被证明对标号噪声不鲁棒，这也表明常见的机器学习算法都不完全对标号噪声鲁棒。

还有部分学习算法通过避免过拟合来缓解标号噪声带来的影响。如在决策树（Decision Tree, DT）算法中，Masegosa 等人[38]利用不精确的信息增益来减少决策树的大小并观察到决策树的修剪可以用来避免过拟合现象，从而降低了标号噪声对模型带来的影响。Abellán 等人[39]将这个剪枝方法扩展到了连续特征和缺失数据中。还有如集成学习方法（Ensemble Methods）中，Dietterich 等人[40]发现在标号噪声情况下 bagging 算法的分类精度高于 boosting 算法。这是由于 bagging 算法重复选择数据集的不同子集，每次重采样导致不同的模型，因此在 bagging 算法中基分类器的多样性大大提高。

综上所述，机器学习中常用的损失函数对标号噪声是不完全鲁棒的。然而，过拟合避免技术，如鲁棒的损失函数、正则化技术、决策树剪枝、集成学习等，可以在一定程度上缓解标号噪声的影响。文献中大量实验表明标号噪声对分类器的性能影响仍然很大。鲁棒的学习算法似乎只适用于简单的情况，这些情况可以通过防止过拟合而安全地管理标号噪声。因此，为标号噪声特别设计的算法是很有必要的。

2.1.2 数据过滤、纠正算法

当数据集被标号噪声污染时，对其进行过滤、纠正是最直接的方法。数据清洗的一般过程如图 2.1 所示。利用清洗后的训练数据进行建模往往比第一类方法更有优势，这是因为清洗后的训练数据对模型推理的影响大大减小。

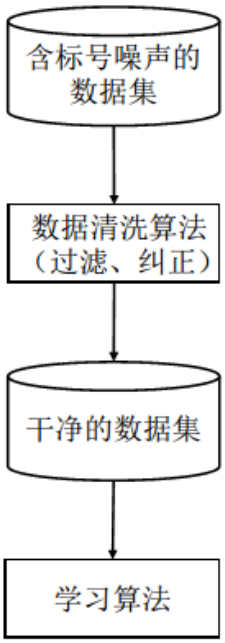


图 2.1 数据清洗的一般过程

一些方法基于分类器的错分思想来进行过滤数据，Jeatrakul 等人[41]利用含噪声的训练数据学习神经网络模型，将所有的错误分类的数据视为标号噪声数据并删除。Pruengkarn 等人[42]利用支持向量机分类器进行了相似的工作。然而由于分类器的错分问题，基于分类模型的过滤数据方法可能会删除一些正确标号的数据或删除较少部分的错误标号数据。事实上，这些方法陷入了“鸡生蛋，蛋生鸡”的矛盾现象，也就是说，一个基于分类模型的过滤数据方法依赖于一个好的分类器，而一个受污染的训练数据集往往得到一个较差的分类器。为了进一步提高分类器中错误分类的数据和标号噪声数据的相关性，Brodley 等人[25]使用多个学习算法构造多个分类器作为过滤器，并将在大多数分类器都错分的数据视为标号噪声数据并过滤，然后用过滤后的新数据构造分类器作为学习模型验证准确度。在分类准确度上取得了不错的提高，最高能应对 30%的噪声水平。

还有一些方法通过引入领域专家来进行过滤数据，Fefilatyev 等人[43]基于 SVM 分类器中的支持向量几乎包含所有的错误标号数据这一事实，引入一名人类专家从作为支持向量的数据中来检查可疑的错误标记数据，并将错误标记的数据重新标记。进一步地，Ekambaram 等人[44]通过迭代使用非支持向量数据学习得到的 SVM 分类器，降低了人类专家需要审查候选数据的

数量。尽管这一类方法可以得到目前最好的清洗结果，但是人类专家在大多数应用场景中是很难获得的。另一方面，我们无法保证新标号是正确的。

总而言之，尽管数据过滤、纠正的方法是最为直接的处理标号噪声的方法，但是一些方法可能会过滤掉过多的数据[45][46],与此同时，过多地过滤正确标记的数据可能会因为数据规模变小而对分类性能造成比标号噪声更大的影响[25]。因此一个好的数据过滤、纠正算法需要在这其中找到折中的解决办法。

2.1.3 标号噪声容忍算法

当我们事先知道与标号噪声有关的信息或者噪声对学习模型的具体影响时，可以在设计模型、算法的时候把标号噪声考虑在内。具体情况下,可以在学习一个分类器的同时学习一个噪声模型，将数据生成的过程和提高分类器性能的过程分开，也就是根据数据真实、未知的标号来学习分类器。还有方法通过改进传统的学习算法来达到降低标号噪声影响的目的。同时，数据清洗方法也可以直接嵌入到学习模型的过程。由于上述方法可以在建模过程中容忍标号噪声，因此被称为标号噪声容忍算法。

具体地，Wang 等人[47]将近邻成分分析度量学习方法扩展到一个标号噪声容忍的版本。通过分析标号噪声对转移矩阵的影响，提出利用给定观察标号预测真实标号的条件概率这一方法来减少标号噪声的影响。在这种方法中，每个数据点对假定模型的影响都是经过精心计算的，然而会存在降低学习效率、模型过拟合等风险。Ganapathiraju 等人[48]将数据清洗嵌入到 SVM 的学习过程中，把对应较大对偶权重的数据视为候选的标号噪声数据并过滤。还有一些方法从经典分类器的学习过程进行分析，提出了一系列改进方法来提高分类器对标号噪声环境下的预测性能。例如$\alpha-bound$ [49]、$\lambda-trick$ [50]等方法从学习策略出发，用于缓解感知器算法对标号噪声的偏置；MadaBoost[51]、AveBoost[52]等算法则通过改进学习参数的更新过程来修正传统的 boosting 方法。

总体而言，尽管通过考虑和分析标号噪声的影响来直接建模学习过程往往能获得较好的实验效果，但是这类方法最主要的问题是，对标号噪声的额外考虑增加了学习算法的复杂度，同时可能会由于学习过程中额外的学习参数而导致过拟合现象。

2.2 深度自编码网络

本节重点介绍经典神经网络中的自编码网络的基本原理，并进一步介绍当前研究热点的深度自编码网络的发展与应用。

2.2.1 自编码网络原理

自编码网络（Autoencoder）是一种典型的人工神经网络，主要以无人监督的方式并通过重建其输入数据来学习有效的数据编码[53]。一个常见的自编码网络如图 2.2 所示。具体地，假设我们有包含 n 个数据的数据集 $X=\{x_1,x_2,...,x_n\}$，其中 $x_i \in R^d$。自编码网络首先将每个输入数据 $x=\{x^1,x^2,...,x^d\}$ 传输到输入层，并在隐层获得特征编码 $h=\{h^1,h^2,...,h^k\}$，其中 $h \in R^k$。通常情况下，自编码网络用于提取数据隐含特征，因此隐层节点的个数通常比输入层节点数少，即 $k<d$。最后在输出层得到解码，得到重建结果 $\hat{x}=\{\hat{x}^1,\hat{x}^2,...,\hat{x}^d\}$。形式化地，重建结果 $\hat{x}$ 可以表示为：

$$\hat{x}=g(W_2 f(W_1 x+b_1)+b_2) \tag{2.1}$$

其中，W_1 和 b_1 分别是编码器的权重矩阵和偏置项，W_2 和 b_2 是解码器的权重矩阵和偏置项。$f(\cdot)$ 和 $g(\cdot)$ 是激活函数，通常选用 sigmoid 函数。为了使自编码器能较好的重构数据样本，损失函数可表示为：

$$L(\theta;\mathrm{X})=\frac{1}{n}\sum_{i=1}^{n}\| \hat{x}_i-x_i \|^2 \tag{2.2}$$

其中，$\theta=\{W_1,W_2,b_1,b_2\}$。由于 sigmoid 函数是光滑、连续可微的，该损失函数可以通过小批量梯度下降法进行求解。利用充分学习的编码器参数之后，可以在隐层得到紧致的隐层特征表示 $z=f(W_1 x+b_1)$，并作为特征提取器[54]。

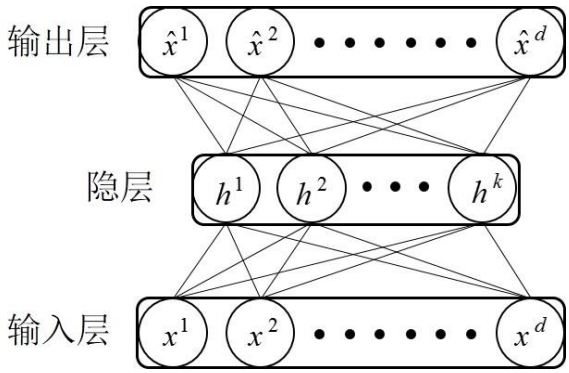


图 2.2 典型的自编码网络

基于原始的自编码模型，不同研究员提出了不同的扩展模型。为了提高特征的鲁棒性，Vincent 等人[55]提出了降噪自编码模型（denosing autoencoder，dAE），该模型通过将部分神经元随机地设置为零，重新构造一个被破坏的数据作为模型输入。Makhzani 等人[56]提出稀疏自编码网络（sparse autoencoder，SAE），通过手动地将所有响应最大的几个隐藏单元激活的方式来对隐藏单元施加稀疏以获得输入数据最有用的特性。Rifai 等人[57]通过给目标函数添加显式的正则化项来避免过拟合。Kingma 等人[58]则将原始自编码模型扩展为变分自编码模型（variational autoencoder，VAE），该生成式模型可以通过调节隐变量来生成变化的图像。

2.2.2 深度自编码网络

近几年，由于深度学习的发展，自编码模型通过堆叠的方式形成了深度自编码模型（deep autoencoder，DAE）[59]。深度自编码模型可以捕获更加复杂、抽象的特征，这类深度网络在很多应用都取得了巨大的成功，如特征学习[60]，物体检测[61]等。一个典型的深度自编码如图 2.3 所示。其损失函数可以表达为如下公式：

$$\arg\min_{W_1 \ldots W_L, W_1^{'} \ldots W_L^{'}} \| X - g \circ f(X) \|^2 \tag{2.3}$$

其中 $f = \varphi(W_L \varphi(W_{L-1} \ldots \varphi(W_1 X)))$ 为编码函数，$g = W_1^{'} \varphi(W_2^{'} \ldots \varphi W_L^{'}(f(X)))$ 为解码函数。解决以公式（2.3）为损失函数的问题会遇到计算能力不足的挑战，通常我们以一种贪婪的方式（即一次学习一层的方式）学习模型参数。具体流程分为以下两个步骤：

（1）无监督的预训练（unsupervised pre-training）：我们将每个隐层当做一个单独的自编码网络并以最小化重建误差作为目标进行训练。一旦训练好第 k 层，第 $k+1$ 层的输入即为第 k 层的输出。以此类推，依次训练所有的隐层。

（2）微调（fine-tuning）：一旦所有的隐层均被预训练完毕，借助监督信息并在监督任务中根据最小预测误差目标来对整体参数进行进一步的调整。

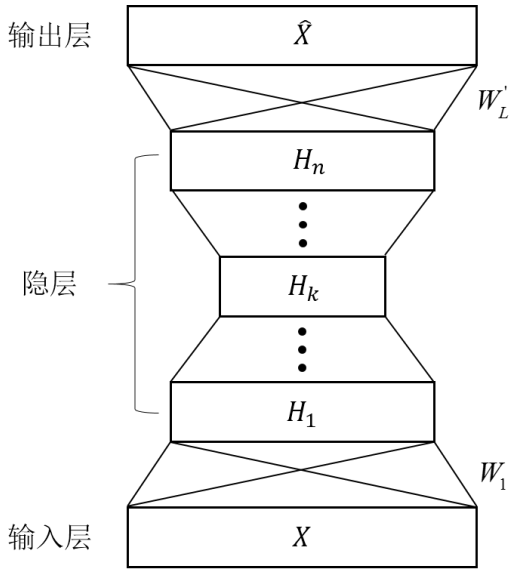


图 2.3 典型的深度自编码网络

2.3 生成对抗网络

本节重点介绍最近流行起来的生成对抗网络的基本原理，并进一步介绍有监督的生成对抗网络和信息最大化的生成对抗网络两种经典模型。

2.3.1 生成对抗网络

生成对抗网络（Generative Adversarial Networks，GAN）[62]是 2014 年由 Ian Goodfellow 首先提出的并广泛用于无监督的机器学习应用中。它描述了两个神经网络系统在博弈论的零和游戏框架中相互竞争的学习过程。该技术可以生成与训练数据非常相似的图像，甚至通过人眼无法区分开来。近两年内，该技术已经成为了机器学习领域最热门的方向之一，并有各种各样的扩展方法被提出[63][64]。

一个典型的 GAN 框架如图 2.4 所示。它包括由多层感知机构成的生成式网络和判别式网络。对于生成式网络，它首先接收一个随机噪声 $Z \sim \mathrm{N}(0,1)$，并通过 $G(x)$ 投影到生成的图像 X_{fake}；其目标在于生成尽可能逼真的图像数据。对于判别式网络，它首先接收一个图像数据，该图像数据可能来自于真实的数据集 X_{real} 也可能来自于生成式网络所生成的图像 X_{fake}，并通过 $D(x)$ 输出该图像是真实的图像还是虚假图像。

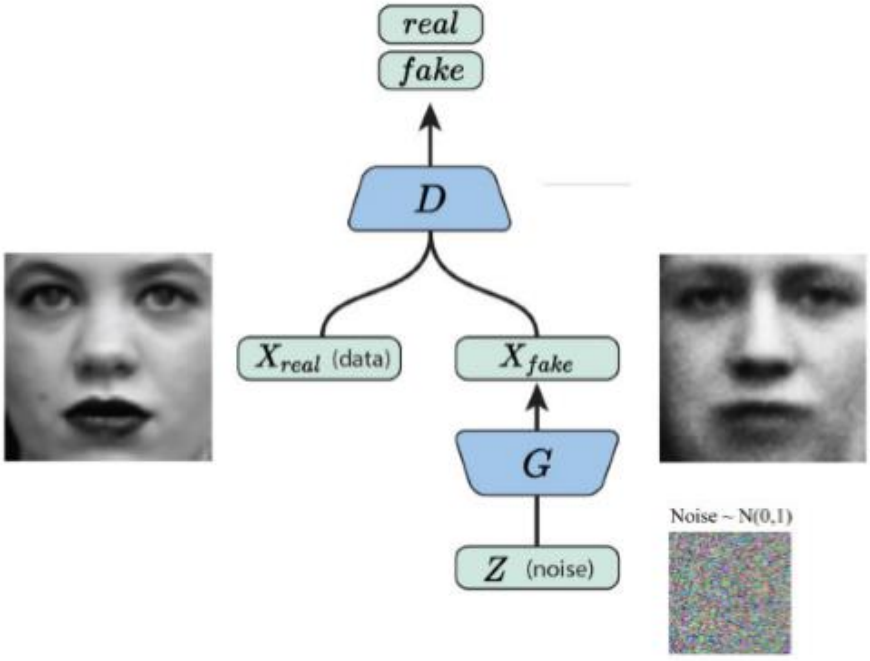


图 2.4 GAN 的基本框架

具体地，对于判别式网络，其目标是最大化正确标签的 log 似然函数，即：

$$\max_{D} E_{x\sim p_{data}(x)} \log D(x) + E_{z\sim p_z(z)} \log(1-D(G(z))) \qquad (2.4)$$

其中 $p_{data}(x)$ 和 $p_z(z)$ 分别表示真实图像数据的分布和随机噪声的分布。生成式网络 $D(x)$ 的输出结果为 0-1 的概率值，该值越大表明预测的图像是真实图像的可能性越大。对于生成式网络，由于它要生成逼真的图像数据来欺骗判别式网络，因此其目标是最大化将该生成图像预测为真实图像的概率，即最小化其预测为虚假图像的概率，即：

$$\min_{G} E_{z\sim p_z(z)} \log(1-D(G(z))) \tag{2.5}$$

将公式（2.4）和公式（2.5）结合起来，得到一个统一的优化目标函数如下所示：

$$\min_{G}\max_{D} E_{x\sim p_{data}(x)} \log D(x) + E_{z\sim p_z(z)} \log(1-D(G(z))) \tag{2.6}$$

可以看出，生成器和判别器在共同进行一个最小最大的博弈问题。由于两个网络都是多层感知器，因此可以通过基于 mini-batch 随机梯度下降法的迭代优化来进行学习。同时，该模型有全局最优解和良好的收敛性质，即当两个网络有足够的学习能力时，模型可以通过上述优化方式来达到其各自的最优解，最终使得生成的图像分布于真实图像分布一致。

2.3.2 有监督的生成对抗网络

由于 GAN 是无条件的生成模型，因此所生成的图像是随机生成的。然而，通过对生成模型提供额外的信息，可以更好的指导图像的生成过程，这样的生成对抗网络被称为条件生成对抗网络（Conditional Generative Adversarial Networks，CGAN）[65]。

一个基本的 CGAN 在生成器和判别器中引入辅助信息 y 作为额外的输入层。最常见且应用最广泛的辅助信息是图像数据对应的类别标号。其具体框架如图 2.5 所示，在生成器中，将类别标号 C 与先验噪声 $Z \sim \mathrm{N}(0,1)$ 结合输入到生成网络中；在判别器中，将其与真实图像数据 X_{real} 或生成图像 X_{fake} 结合输入到判别网络中。通常将类别标签展开成 one-hot 向量。CGAN 的目标函数如下所示：

$$\min_{G}\max_{D} E_{x\sim p_{data}(x)} \log D(x\,|\,y) + E_{z\sim p_z(z)} \log(1-D(G(z\,|\,y)\,|\,y)) \tag{2.7}$$

通过引入类别标签作为辅助信息，CGAN 可以通过利用生成器根据指定的标号来生成相应类别的图像数据。Gauthier 等人[66]提出利用卷积神经网络代替多层感知器模型并将人脸属性作为辅助信息输入到 GAN 模型中得到 CGAN 模型。该模型不仅可以生成拥有特定属性的人脸图像，而且由于使用卷积神经网络的原因，使之推广到 RGB 图像数据且生成的图像更加逼真与稳定。

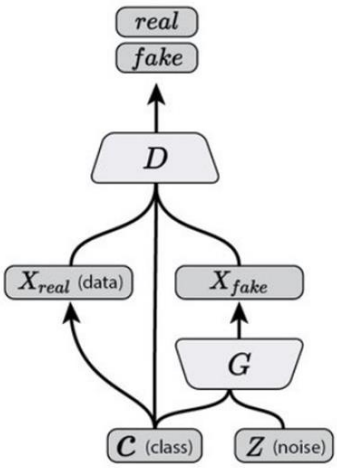


图 2.5 CGAN 的基本框架

上述方法在低精度的图像数据集中获得了良好的效果，然而面对更高精度、更复杂的图像数据其生成效果仍然不够逼真。因此，Odena 等人[67]提出了基于辅助分类的生成对抗式网络（auxiliary classifier GAN，AC-GAN），其基本框架如图 2.6 所示。可以看出，相比较于 CGAN 模型，AC-GAN 模型改进了判别器的目标，即除了预测样本是真实样本还是生成的样本，还要预测样本的类别。具体地，其目标函数包含两部分：1）样本正确来源的 log 似然 L_S，以及 2）样本正确类别的 log 似然 L_C，如公式（2.8）和公式（2.9）所示。

$$L_S = E[\log P(S = real \mid X_{real})] + E[\log P(S = fake \mid X_{fake})] \quad (2.8)$$

$$L_C = E[\log P(C = c \mid X_{real})] + E[\log P(C = c \mid X_{fake})] \quad (2.9)$$

其中 $X_{fake} = G(c, z)$。在训练过程中判别器目标函数为最大化损失 $L_S + L_C$，而生成器的目标函数为 $L_C - L_S$。

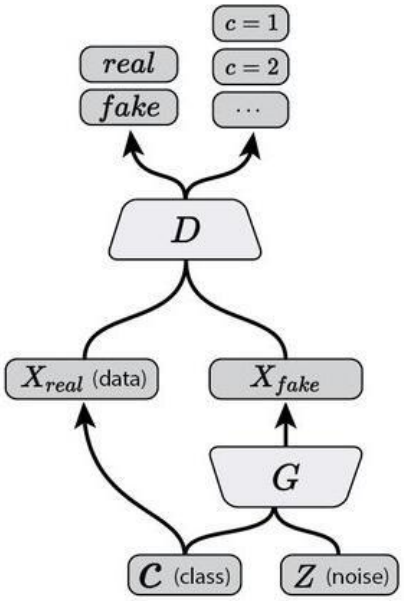


图 2.6 AC-GAN 的基本框架

通过对判别器增加分类损失最小化这一目标，该模型在 ImageNet 大规模复杂图像数据集进行学习。实验结果表明，相比传统的 CGAN 模型，该 AC-GAN 模型生成了具有更高逼真程度和多样性的图像数据。

2.3.3 信息最大化的生成对抗网络

CGAN 模型与 AC-GAN 模型在训练过程中均引入了类别标号这一监督信息来提高模型的稳定性并可较好生成特定类别的图像数据。然而，在监督信息不准确的情况下（即标号噪声现象），这类方法均会在一定程度上失效。Chen 等人[68]提出的信息最大化生成对抗式网络（Information Maximizing GAN，InfoGAN）可以很好地规避标号噪声问题。InfoGAN 将信息最大化准则引入 GAN 模型，从而可以以无监督学习的方式实现可解释的表示学习过程。

具体地，相比 GAN 模型直接根据随机噪声来生成图像，InfoGAN 模型则引入了可解释的隐变量c，通过约束隐变量与生成图像数据的关系来学习数据可解释的信息。该模型的损失函数表示如下：

$$\min_{G}\max_{D}V(D,G)-\lambda I(c;G(z,c)) \tag{2.10}$$

其中$V(D,G)$为 GAN 模型的损失函数，$G(z,c)$为根据随机噪声 z 和隐变量c生成的图像数据，$I(c;G(z,c))$则表示隐变量和生成图像的互信息。互信息越大，隐变量与生成样本之间的相关性越高。然而由于在$I(c;G(z,c))$的计算中，真实的后验概率$P(c\mid x)$无法获得，因此在优化过程中，通过变分推断的思想，引入$Q(c\mid x)$来近似$P(c\mid x)$。最终，InfoGAN 模型的目标函数如下：

$$\min_{G,\mathrm{Q}}\max_{D}V(D,G)-\lambda L_I(G,Q) \tag{2.11}$$

由于Q和D可共用同一个深度卷积网络且在最后一层增加一个全连接层来得到$Q(c\mid x)$，因此在模型上的改动较小。当模型学习完毕时，一方面，通过设定隐变量c为指定的标号来生成对应标号的图像数据。另一方面，通过改变随机噪声 z 的取值来获得同一类别下更加丰富和多样化的图像数据。该方法以无监督的方式生成指定标号的样本在数据增广任务中潜力巨大，尤其是在监督信息不可用的场景中。

2.4 本章小结

在本章内容中，首先对文献中针对标号噪声问题的处理技术进行了简要介绍，共包括鲁棒的学习算法、数据过滤纠正算法以及标号噪声容忍算法；其次，介绍了经典的特征学习方法自编码网络的原理以及深度自编码模型的构造和训练过程。最后详细介绍了基本的生成对抗式网络、可生成指定类型的图像数据的有监督生成对抗式网络以及信息最大化的生成对抗式网络的基本原理。为后面的章节做好了充足的技术背景介绍以及模型理论支撑。

第三章 基于异常检测技术和重建误差最小化的数据清洗

3.1 引言

通常，为了减少标号噪声的影响，清洗数据集中的数据是一种最直接的方式。一些方法通过利用异常检测技术来清洗错误标签的数据，如 Local Outlier Factor（LOF）[69]、One Class Support Vector Machines（OC-SVM）[70]等，这些方法直接将有标号噪声的数据视为其对应类别数据中的异常值。然而，并不是所有的异常值都是错误标签的数据。正如对异常值的定义一样，“*异常值是一种观察结果，它偏离了其他大多数的观察结果，并让人觉得它是由其他不同机制产生的样本*”[71]，在一个类别边界区域中的样本同样也满足上述定义，但并不一定包含标签噪声。因此，我们尝试纠正潜在、不准确的假设（即将标号噪声的数据当做异常值）。相反地，将异常值视为候选、高概率的错误标记的数据，这是一种更直观、更恰当的假设。

基于上述合理的假设，我们不把异常值直接视为带有标号噪声的数据点，我们只把它们当作**候选**的错误标记的数据，然后用另一个步骤来验证他们是否真的有不正确的标签。为此，通过将异常检测技术与重建误差最小化准则相结合，我们提出了一种新颖有效的方法来减少标号噪声的影响。具体地，首先将含标号噪声的数据集根据标号分为多个子集。然后利用基于鲁棒的深度自编码网络（robust deep autoencoder，RDA）的异常检测方法在每个子集中找到异常值，并获得子集的类内特征表示。这些异常值被视为候选、高概率的标号噪声数据。由于并不是所有的异常值都是标号噪声的数据，我们进而利用重建误差最小化（reconstruction error minimization，REM）准则，通过检查它们是否可以被相应的类重新构造，从而进一步验证这些候选者。最终，找到真正的标号噪声数据并修正错误标号完成数据清洗任务。

3.2 数据清洗框架

本文所提出的基于异常检测技术和重建误差最小化的数据清洗框架如图 3.1 所示，共包括两个步骤：

步骤 1：基于鲁棒的深度自编码网络的异常检测

首先，针对含标号噪声的训练数据集，根据数据各自相应的标号来划分为多个子集。然后，在每个子集中利用鲁棒的深度自编码网络异常检测技术检测异常数据。最后，我们获得每个子集中候选、高概率的标号噪声数据（即异常数据）和学习良好的类自编码网络。

步骤 2：基于重建误差最小化准则的标号噪声清洗

首先，将候选标号噪声数据输入到各个学习良好的类自编码网络。然后计算各自的重建误差来进一步判断是否为标号噪声数据并基于重建误差最小化准则来纠正标号，达到清洗数据的目的。

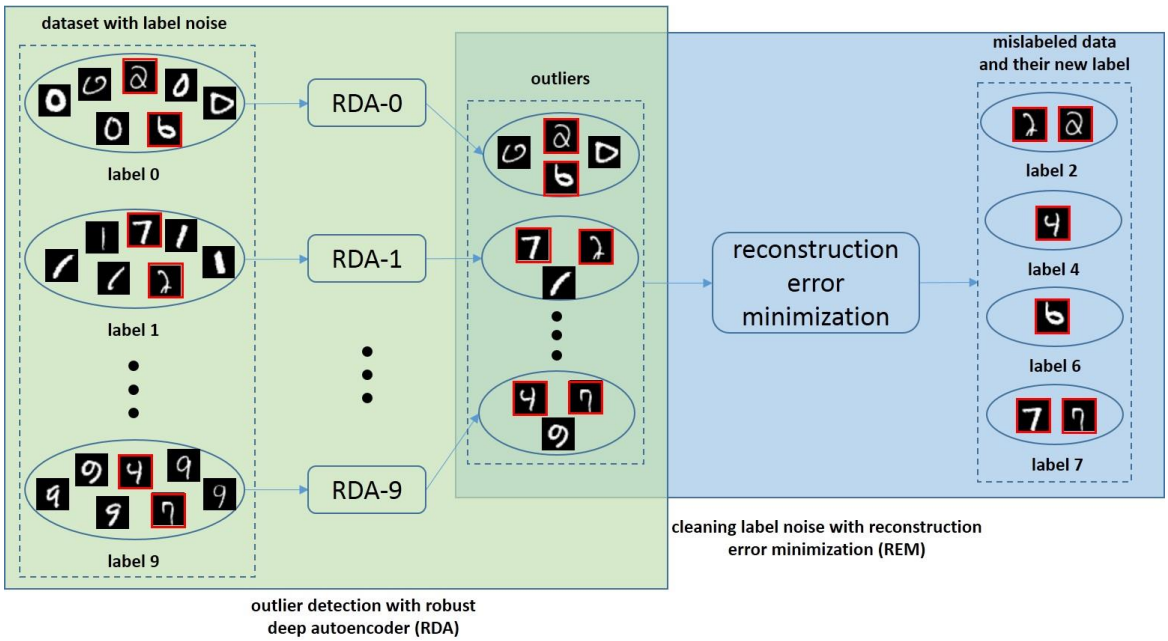


图 3.1 基于异常检测和重建误差最小化的数据清洗框架

3.3 基于鲁棒的深度自编码网络的异常检测

本节重点介绍了基于鲁棒深度自编码网络的异常检测技术，分别详细介绍了该模型的基本原理以及模型的训练方式。

3.3.1 鲁棒的深度自编码网络

由于提出的数据清洗框架建立在异常检测技术之上，而且大多数异常检测方法（例如：one-class SVM[72]和 one-class neural network[73]等）都可以被使用。在此，本文利用一种先进的异常检测方法即鲁棒的深度自编码网络（robust deep autoencoder, RDA）[74]来寻找异常值。具体地，假设有数据集矩阵 X，其中每一行代表一个数据项且部分数据为异常值。RDA 认为数据矩阵 X 是由两部分构成，即 $X = L_D + S$。其中 L_D 是能够通过一定方式下解释的部分，即我们认为的正常值，可以通过深度自编码网络进行很好的重构、解释，而 S 则表示为难以重构的部分，即我们认为的异常值。该异常检测即时的目标函数如下所示：

$$\begin{aligned} &\min_{\theta,S} \| L_D - D_\theta(\mathrm{E}_\theta(L_D)) \|_2 + \lambda \| S^T \|_{2,1} \\ &s.t.\ X - L_D - S = 0 \end{aligned} \tag{3.1}$$

共分为两个部分，第一部分是 L_D 的重构误差损失，$E_\theta(\cdot)$ 和 $D_\theta(\cdot)$ 分别表示为编码函数和解码函数；第二部分是 S^T 的 $l_{2,1}$ 范数损失，具体可以被定义为：

$$\| S^T \|_{2,1} = \sum_{i=1}^{n} \| s_i \|_2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} | x_{ij} |^2 \right)^{1/2} \tag{3.2}$$

该式可以被看成是对每一行求 l_2 范数，再对整体求 l_1 范数。通过 $l_{2,1}$ 范数的约束，可以从每一个

数据中找到稀疏的异常值。λ 则为两个部分的平衡因子。

3.3.2 模型训练

模型的训练主要是基于交替方向乘子法(Alternating Direction Method of Multipliers，ADMM)的相关思想，通过 ADMM[75]算法可以将最小化目标函数拆分为几部分，在保持一部分不变的同时，优化另一部分。具体地，针对公式（3.1）的目标函数，将其分解为两部分，分别是依赖 L_D 的 $\| L_D - D_\theta(\mathrm{E}_\theta(L_D)) \|_2$ 和依赖 S 的 $\| S^T \|_{2,1}$。当 S 保持不变时，利用经典的深度学习优化算法即反向传播算法（back-propagation，BP）[76]来优化第一部分的深度自编码模型。对于 $\| S^T \|_{2,1}$ 的优化，由于其不可微，借助近端梯度算法（proximal gradient method，PG）[75]来求解。最终，利用 ADMM 和迪杰斯特拉交替投影方法（Dykstra's alternating projection method）[77]迭代优化整合上述两部分。整个优化算法如下所示。

算法 3.1 鲁棒的深度自编码优化算法

输入：训练数据 $X \in R^{N\times n}$，收敛阈值 ε

输出：深度自编码模型参数 θ 和稀疏矩阵 S

初始化 $L_D \in R^{N\times n}$ 和 $S \in R^{N\times n}$ 为零矩阵，令 $LS = X$，并初始化深度自编码模型参数 θ。

While (Ture)：

(1)令 $L_D = X - S$

(2)利用反向传播算法，最小化第一部分 $\| L_D - D_\theta(\mathrm{E}_\theta(L_D)) \|_2$

(3)根据（2）中训练好的模型，令 $L_D = D(E(L_D))$

(4)令 $S = X - L_D$

(5)利用近似算子优化 S，对 S 中每行数据 s 计算 $prox_\beta(s)$

(6)计算 $c_1 = \| X - L_D - S \|_2 / \| X \|_2$ 和 $c_2 = \| LS - L_D - S \|_2 / \| X \|_2$

(7)检查收敛条件。如果 $c_1 \le \varepsilon$ 或者 $c_2 \le \varepsilon$ 则跳出循环，否则进入(8)

(8)更新 $LS = L_D + S$,并进入下一个循环。

End while

其中 $l_{2,1}$ 范数的近端算子 $prox_\beta(x)$ 表示为如下：

$$prox_\beta(x) = \begin{cases} x_j - \beta \dfrac{x_j}{\| x \|_2}, & \|x\|_2 > \beta \\ 0, & \|x\|_2 \le \beta \end{cases} \qquad (3.3)$$

在公式（3.3）中，x_j 表示数据 x 的第 j 维。由于该算子将数据的每一维度结合在对应行数据中，因此是行稀疏。

当模型训练完成后，我们认为稀疏矩阵 S 中的所有非零行为异常样本，并将他们作为候选、高概率的标号噪声数据作为下一部分数据清洗的输入数据。

3.4 基于重建误差最小化的数据清洗

通过利用 RDA 异常检测方法找出异常值之后，不仅获得了每个子集中的异常值，同时也得到了 N 个学习良好的深度自编码网络 $<E_{\theta_i}(\cdot), D_{\theta_i}(\cdot)>, i=1,2,...,N$ 。由于这些自编码模型可以视为相应类别的模板，可以在一定程度上反应每个类的特性。因此，我们利用这类信息从异常数据中进一步找到真正的标号噪声数据。整个流程如图 3.2 所示。

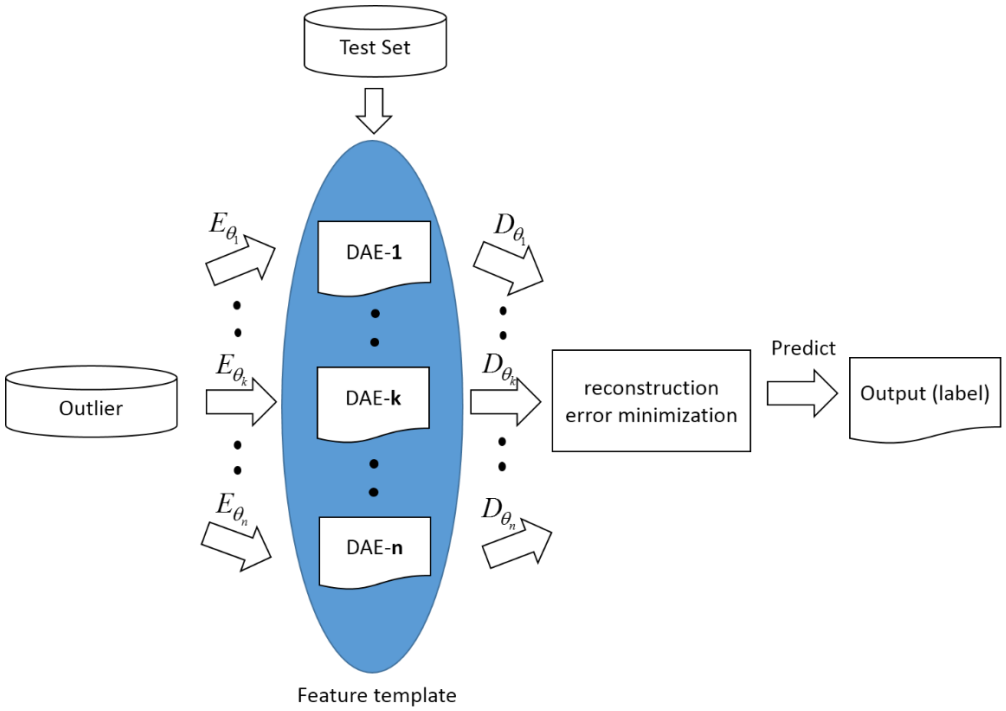


图 3.2 重建误差最小化流程图

具体地，对于异常数据 x，将其输入到已获得的 N 个深度自编码网络中，通过编码与解码过程得到重构结果，计算重构误差并预测其真实的标签为对应重建误差最小的类别。基于重建误差最小化的分类函数如下所示：

$$y_{true} = \arg\min_{j=1,2,...,N} \| D_{\theta_j}(E_{\theta_j}(x)) - x \|^2 \quad (3.4)$$

重建误差是一个很好的指标，可以将真实的错误标签数据与其他异常值（例如，在其类的边界区域中的数据）区分开来，因为来自某一个类的数据往往会从它自己的类中获得比其他类更低的重建误差。

进一步地，通过一个指示函数来决定异常值是否是真正的标号噪声数据，该指示函数如下所示：

$$I(x) = \begin{cases} 1, & y \neq y_{true} \\ 0, & otherwise \end{cases} \quad (3.5)$$

其中 y 是原始的标号，y_{true} 是预测的通过公式（3.4）预测的标号。当两个标号相同时，我们认为该异常数据为标号噪声数据。注意到，公式（3.5）也可以被当做一个分类器用来修正错误标记的数据并在测试集中用于分类。

3.5 实验及结果分析

在本节中，共在两个任务中验证所提出算法的有效性，分别是 1）在训练数据集中找出被错误标记的数据即数据清洗任务；2）利用训练集学习的分类模型，在测试数据集中进行测试即标号噪声下的分类任务。

3.5.1 实验数据及平台

为了验证模型的有效性，利用 MNIST 手写数字集来进行相应的对比实验。MNIST 数据集是计算机视觉领域中知名的手写数字数据集，共包含 60000 个训练数据和 10000 个测试数据。每个图片的像素大小是 28*28。由于整个实验环境下需要有标号噪声的训练集，因此我们遵从标准规则[10]引入标号噪声数据，1）在每一个类别中随机选择样本；2）翻转其真实标号为其他类别标号。注意到，只有训练数据集包含标号噪声数据，而测试数据的标号保持不变。图 3.3 显示了部分标号噪声数据。其中每一行的图像数据属于同一个类别，红色框中的图像为该类别的标号噪声图像。

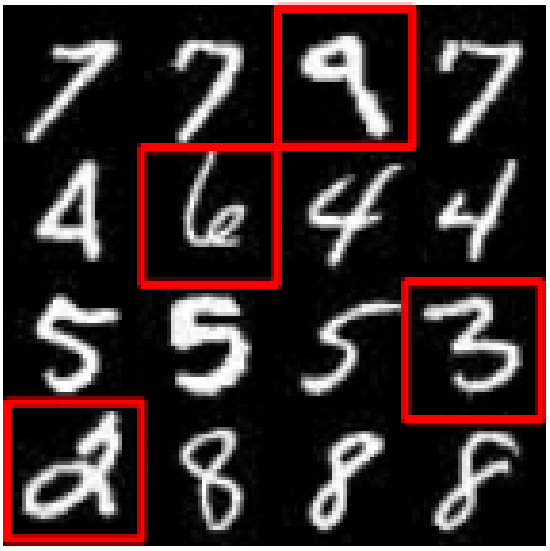


图 3.3 标号噪声数据示例

整个实验平台是在 Windows7 操作系统、1 个 CPU 以及 8GB 内存下进行的，所有方法均在 Pyhton 3.5 的环境下实现。具体地，所提出方法是在 Google 的机器学习库 Tensorflow 0.12.0rc0[78] 实现，对照方法则是利用 scikit-learn 0.19.0 机器学习库实现的。

3.5.2 评价准则

对于评价准则，选取四种度量方式作为我们的评价方法，分别是准确率（accuracy）、查准率（precision）、查全率（recall）和 F1-score，其各自的定义如下所示：

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.6}$$

$$precision = \frac{TP}{TP + FP} \tag{3.7}$$

$$recall = \frac{TP}{TP + FN} \tag{3.8}$$

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.9}$$

其中 TP、FP、TN 和 FN 分别表示为真正类、假正类、真负类和假负类。具体地，在数据清洗任务中，我们利用了上述四种评价方法。对于在测试集的分类任务，我们仅考虑准确率的度量指标。

3.5.3 异常检测性能分析

本节利用数字“1”作为实例，来分析基于鲁棒的深度自编码网络的异常检测性能。具体地，将含 5%标号噪声且标号为“1”的图像数据作为模型的训练集，并随机选取了 100 个标号为数字“1”的图像作为测试矩阵 X 用来验证其性能，可视化如图 3.4 所示，其中部分为标号噪声数据。由于该异常检测模型受平衡因子 λ 的影响，设置了以 0.00005 为间隔从 0.00035 到 0.00105 的阈值。对于不同的阈值，训练了对应的异常检测模型，并可视化了两种矩阵，分别为可重构矩阵 L_D 和稀疏矩阵 S 。两种矩阵的可视化结果如图 3.5 所示。

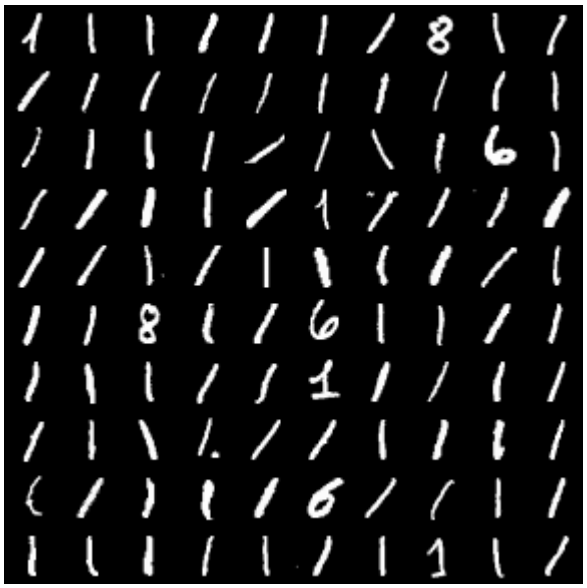
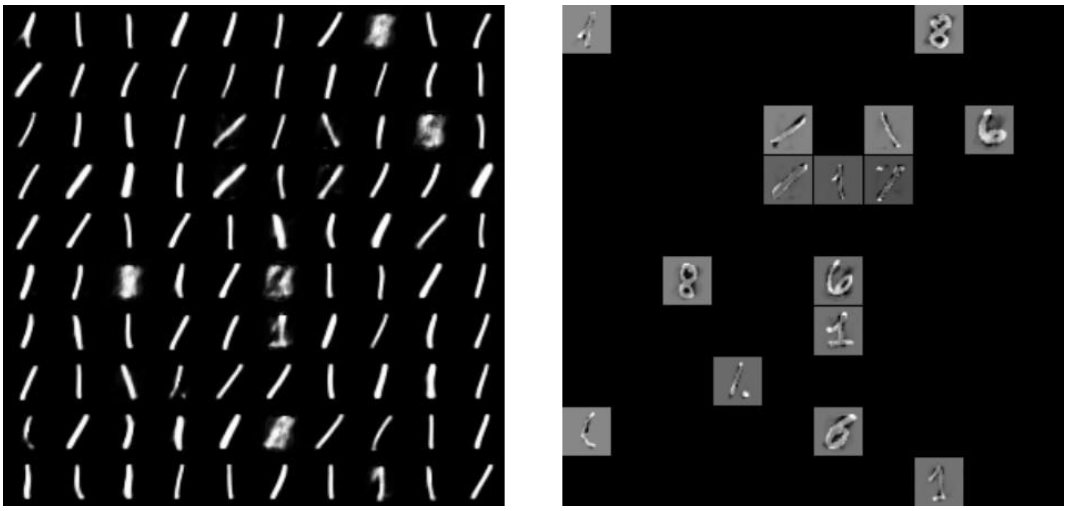
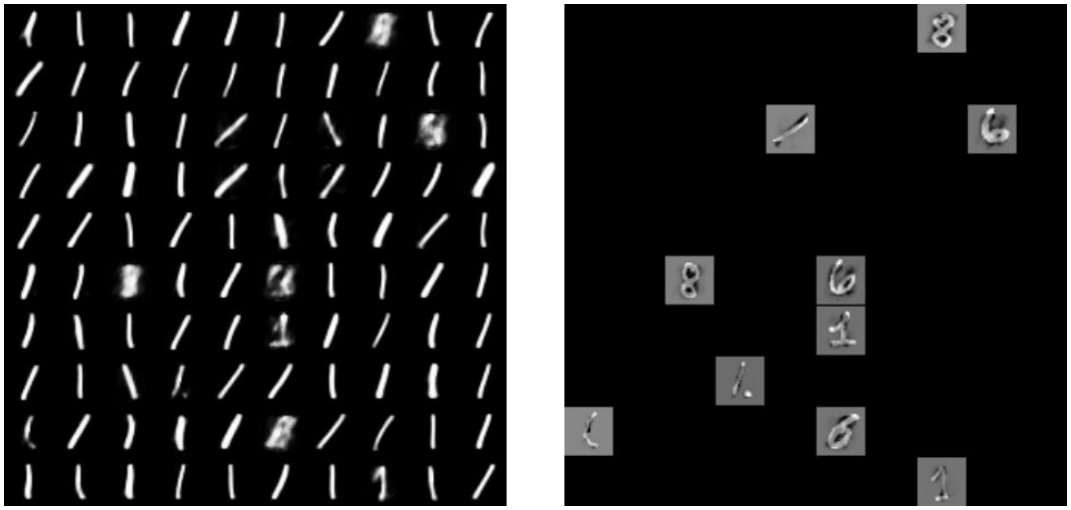


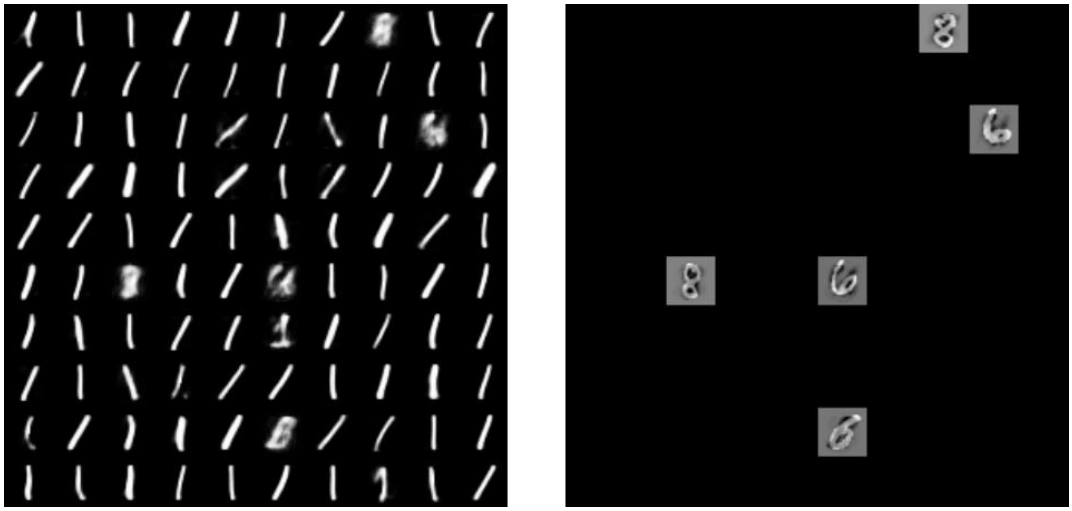
图 3.4 含标号噪声的数字图像“1”



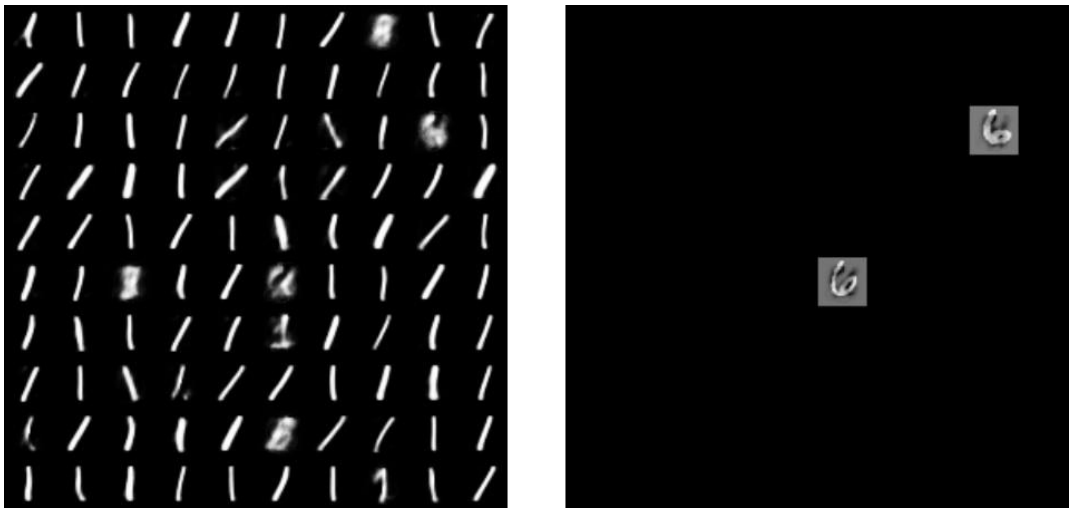
(a) $\lambda = 0.00035$



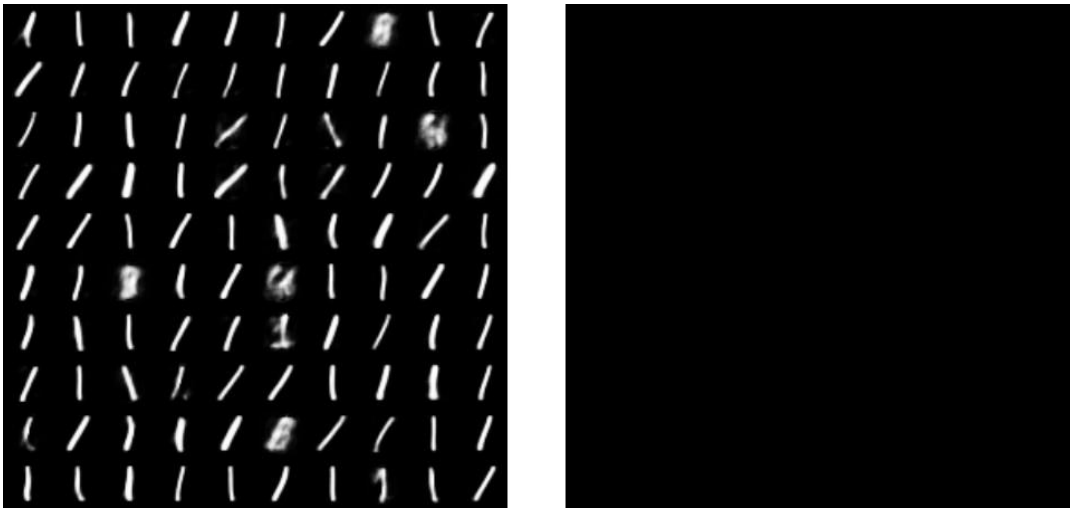
(b) $\lambda = 0.00055$



(c) $\lambda = 0.00075$



(d) $\lambda = 0.00095$



(e) $\lambda = 0.00105$

图 3.5 不同 λ 取值对应的重构矩阵 L_D 和稀疏矩阵 S

从图 3.5 中可以看出不同的 λ 对应的重构矩阵 L_D 仅有轻微的区别，而稀疏矩阵 S 的变化情况较大。具体地，当 λ 小于 0.00075 时，由于对 S^T 的 $l_{2,1}$ 范数损失惩罚较小，则会检测出较多的异常值，存在正确标记的样本被错检现象（即假正类样本）；当 λ 等于 0.00075 时，异常检测效果最佳，所有异常值均检测出来且没有多余的假正类样本；当 λ 大于 0.00075 时，由于对 S^T 的 $l_{2,1}$ 范数损失惩罚较大，则会检测出较少的异常值。当 λ 等于 0.00105 时，所有数据均被当作正确标记的数据，没有异常值被检测出来。从上述现象，可以看出 λ 的取值对异常检测性能影响较大。另一方面，针对不同标号的数字子集，统计了在训练过程中获得最优异常检测效果对应的 λ 取值，见表 3.1。可以看出不同数字子集对应的最优 λ 较为接近，一般情况下，选取 0.00070 或 0.00075 即可获得最优检测结果。

表 3.1 不同数字子集训练模型对应的最优的λ

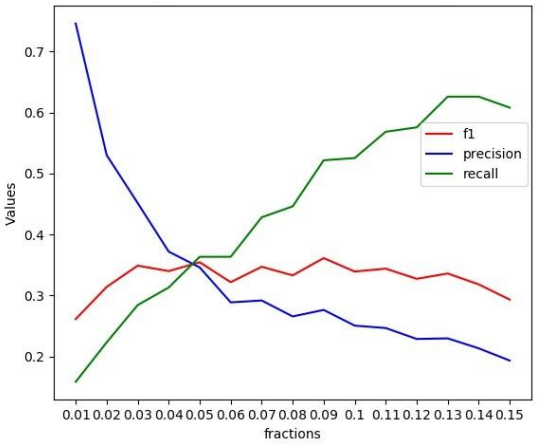
Number	0	1	2	3	4	5	6	7	8	9
λ（10^{-3}）	0.70	0.75	0.75	0.70	0.75	0.70	0.75	0.70	0.80	0.75

3.5.4 数据清洗任务

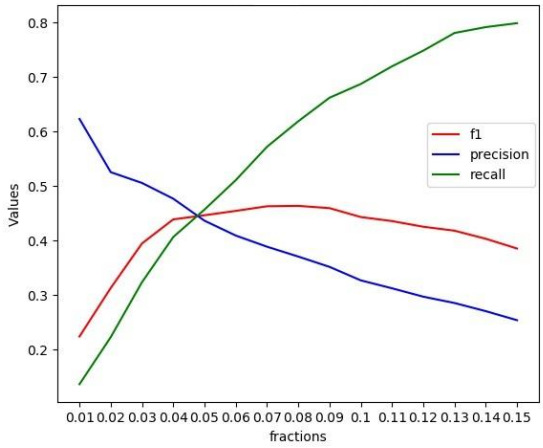
为了证明所提出方法在数据清洗任务中的有效性，将其与经典的错误标号噪声检测方法进行了对比实验。具体地，分别对比了 Isolation forest（iForest）[79]、one class SVM（OC-SVM）[72]和 robust deep autoencoder（RDA）[74]三种方法，并将所提出的方法命名为 LN-RDA。

各模型的参数设置如下：对于 iForest 方法，森林中树的棵数设置为 100，样本中异常值的比例设置为以 0.01 为间隔从 0.01 到 0.15；对于 OC-SVM 方法，针对 SVM 选择了 RBF 核函数并设置了与 iForest 相同的异常值比例；由于所提出的方法是基于 RDA 方法，所以两种方法中深度自编码网络的模型参数设置是一样的，共包含 5 层结构，每层神经元的个数分别是 784、400、200、400 和 784，其中激活函数默认为 Sigmoid 函数。对于超参数λ，均设置为以 0.0001 为间隔从 0.00035 到 0.00125。

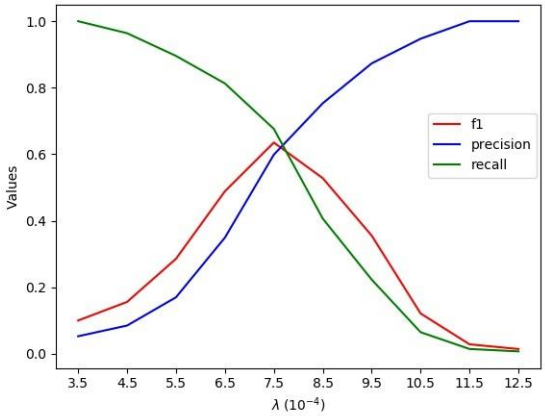
在添加 5%的标号噪声数据后，异常检测的对比实验结果如图 3.6 所示。四种方法最好的 F1-score 值分别为 0.36、0.46、0.64 和 0.82，所提出的方法获得了最好的检测性能。与 RDA 模型相比，LN-RDA 有更高的查准率，这表明了重构误差最小化的方式可以很大程度上减少假正类的数量。对于召回率而言，LN-RDA 和 RDA 性能相近，这是由于所提出方法受 RDA 异常检测模型性能的限制所导致。另一方面，对于 iForest 和 OC-SVM 方法，它们均在异常值比例约为 5%的情况下获得其最高的 F1-score 值，而在其他设置情况下较小。这一现象反映了该两者方法对异常值比例敏感，较依赖于异常值比例这一先验知识。



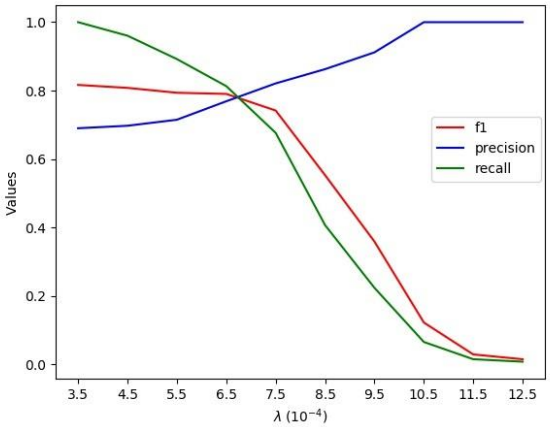
(a) iForest 模型的 recall、precision 和 f1



(b) OC-SVM 模型的 recall、precision 和 f1



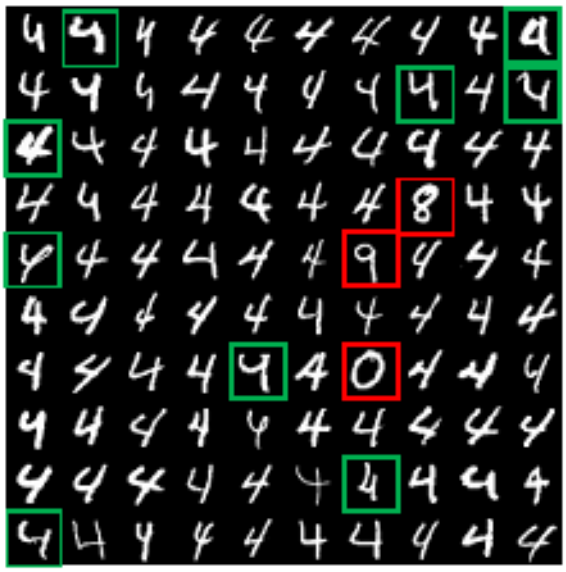
(c) RDA 模型的 recall、precision 和 f1



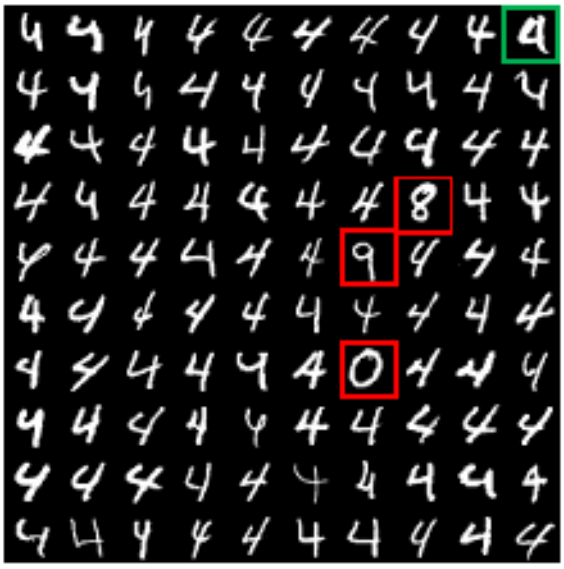
(d) LN-RDA 模型的 recall、precision 和 f1

图 3.6 四种模型的异常检测结果

另一方面，我们可视化了部分检测结果用来从直观的感受上反应所提出方法的有效性。如图 3.7 所示，展示了对训练集中标号为‘4’的训练数据的检测结果。在图 3.7(a)中，RDA 模型检测到的所有异常值都将被视为带有标签噪声的数据点并由一个正方形表示，其中一个红色的正方形表示一个真正类（也就是带有标签噪音的数据点），而一个绿色的正方形表示一个假阳性（也就是一个带有正确标号的数据被错误地认为是标号噪声）。可以看到，RDA 异常检测方法会导致大量的误报，会出现很多假正类的现象。但是本文所提出的 LN-RDA 模型却会显著减少它们的数量。



(a) RDA 模型的检测结果



(b) LN-RDA 模型的检测结果

图 3.7 RDA 和 LN-RDA 模型的检测结果

然后，根据重构误差最小化分类器，对找出的标号噪声数据重新贴标号，并将其结果与其他经典模型来比较。具体地，将其与 ICCN-SMO[80]、TC-SVM[43]和 ALNR[44]模型进行比较。这些被比较模型的超参数设置是基于相应的文献。注意到，这些方法都是通过引入领域专家来重新标记那些候选的标号噪声数据。从这方面来说，因为我们不假设存在这样的监督信息，所提出的方法是完全自动的纠正标号。

表 3.2 给出了各个方法在时间代价和准确度上的性能结果。可以看出，虽然没有人类专家的引入，与其他基于人类专家的方法相比，所提出的方法在时间代价和预测精度上均达到了最佳的性能。此外，由于这些被比较方法是通过人类专家对候选标号噪声数据进行检查并通过视觉修正，因此所提出的模型在时间代价上具有绝对优势。

表 3.2 重新标记错误标号数据的精度和时间代价

Methods	ICCN-SMO	TC-SVM	ALNR	LN-RDA
Time cost (s)	938.6	352.1	70.8	**6.5**
Accuracy(%)	65.47	96.45	94.10	**98.06**

3.5.5 测试集的分类任务

在训练数据含标号噪声的情况下，测试数据的分类精度也是度量方法好坏的一个重要标准。因此，本节分析了在测试集中的分类性能。在所提出的方法中，使用重构误差最小化准则来作为分类器。注意到，对于每个类的 RDA 模型，选择与最好的 F1-score 对应的 λ 作为相应最优超参数。此外，在每个类中训练一个普通的深度自动编码器，并通过所提出的分类器对测试数据进行分类。该方法作为基线方法，被表示为 LN-DA。另一方面，还将所提出的方法与两类最先进的方法进行比较，包括：1）基于数据清洗的方法 TC-SVM[43]和 ALNR[44]以及 2）标号噪声鲁棒的方法 L1-norm[81]、BML[82]和 RNCA[47]。所有被比较方法都是基于相应的论文的原始实现，并通过交叉验证选择相关的超参数。我们设置了 4 种不同程度的标号噪声（5%、10%、20%和 30%），对每个实验重复 30 次并展示了平均分类精度以及相应的标准差。

如表 3.3 所示，与其他方法相比，我们所提出的方法在较低（5%和 10%）和较高（20%和 30%）水平噪声下均能达到最佳性能，表明了所提出框架在处理标签噪声问题的有效性。当训练集无标号噪声时，LN-DA 获得最佳的分类精度，表明了重建误差最小化的分类器是一个合适的选择。然而，由于标签噪声的影响，LN-DA 的性能变得更差，尤其是在高噪音环境下。

针对基于数据清洗的方法，与基线方法相比，他们获得了更高的精度，因为他们使用更干净的训练集来学习分类器。另一方面，标号噪声鲁棒的方法相比基于数据清洗的方法在低噪音

水平上取得了更好的效果。然而，这些方法在更高的噪声水平上表现相对较差，这反应了在较高水平的标号噪声下获得可靠的点估计的难度。

表 3.3 在不同标号噪声下的测试集的分类性能

（星号表示在 0.05 的显著性水平上次优的方法与所提出方法有统计上显著性差异）

Methods	Noise level (%)				
	0	5	10	20	30
TC-SVM	98.27±0.06	95.83±0.06	94.95±0.16	91.45±0.20*	84.43±0.27*
ALNR	98.27±0.10	95.19±0.10	94.47±0.21	90.21±0.27	83.39±0.33
L1-norm	98.00±0.10	96.67±0.15	95.12±0.22*	89.78±0.31	81.39±0.49
BML	97.95±0.10	96.58±0.16	95.08±0.23	89.68±0.31	81.63±0.49
RNCA	98.15±0.11	96.73±0.17*	95.05±0.22	89.77±0.31	82.31±0.49
LN-DA	**99.12±0.15**	95.37±0.20	94.28±0.28	86.33±0.40	79.26±0.55
LN-RDA	99.04±0.11	**98.06±0.18**	**96.24±0.25**	**92.50±0.37**	**86.39±0.50**

3.6 本章小结

本章基于一个简单但新奇的想法，也就是把异常值检测过程视为初步检测候选标号噪声数据的过程。提出使用重建误差最小化准则来进一步验证被异常检测方法检测到的数据是否为标号噪声数据。我们认为对于正确标号的数据点来说，它们的重构误差在由仅有少量非同类别的数据训练的鲁棒模型中要远小于由大量非本类别数据训练的模型。在 MNIST 数据集上的各种实验表明，所提出的方法大大降低了直接使用异常检测算法来识别带有标号噪声的数据的假正类的数量。另一方面，还展示了所提出的方法相比当前最先进的方法在数据清洗任务和测试集的分类任务中均获得了更好的性能。

第四章 基于数据增广和鲁棒的自编码网络的特征学习模型

4.1 引言

在第三章中，传统的深度自编码网络因训练集标号噪声的存在而影响了特征学习的准确性，从而间接影响了测试集的分类性能。因此，一个对标号噪声鲁棒的特征学习方法是处理问题的关键。第三章中用到的基于鲁棒的深度自编码网络的异常检测方法就可以看做一种鲁棒的特征学习方法。另一方面，一个大规模的数据集是学习一个良好的特征空间的基础，而标号噪声问题直接使可利用的有效数据量变少。在样本数量受限的情况下，特征学习过程也会因为训练过程不够充分而导致分类等任务的精度下降。

因此，本章仍以自编码网络模型为基础，在因标号噪声存在而导致有效样本数较少的应用场景下，提出了一种基于数据增广和鲁棒的自编码网络的特征学习模型。具体地，针对有效样本数量较少的情况，利用信息最大化生成对抗式网络（InfoGAN）来对样本集进行数据扩增，该方法可在无监督的条件下生成指定标号的样本。接着将增广后的数据集根据相应的标号划分成各个子集并利用自编码网络为每个类别学习一个单独、可靠的特征空间。由于原始数据集和生成的数据集中的标号噪声对自编码网络特征学习能力有一定的影响，提出了一种鲁棒的自编码网络。特别地，针对自编码网络训练的优化方法 mini-batch 梯度下降法，提出了一种重要性加权的 mini-batch 梯度下降的优化过程。该方法通过将每个数据的重构误差大小转化为优化过程中数据的重要性权重，集成在整个模型学习过程中。通过考虑每个数据点对模型学习的影响，自编码网络的特征学习可以在一定程度上缓解标号噪声带来的影响。验证性实验表明，所提出的加权优化方法可以得到更可靠、准确的特征表示。为了充分利用学到的类专属特征空间，使用重建误差最小化的分类器找出标签噪声并重新标记，同时将整个过程扩展为迭代的方式进行更加充分的特征学习。在 MNIST 和 Caltech-10 数据集上的实验表明，数据扩增对特征学习能力有明显的提高。同时，相比其他最先进的方法，所提出的鲁棒的自编码网络在数据清洗和测试集分类任务均达到了最佳的性能。

4.2 基于生成对抗网络的数据增广

本节中，首先通过分析在标号噪声存在的情况下，有效样本数量对经典分类模型精度上的影响。然后提出利用信息最大化生成对抗网络来生成特定标号的样本。最后通过验证性实验对生成样本的可用性进行了详细地分析。

4.2.1 有效样本数量对分类精度的影响

为了验证有效样本数量对模型特征学习以及分类任务的影响，我们以计算机视觉中经典的卷积神经网络 LeNet-5 作为基本的特征学习和分类任务的模型。具体地，采用 MNIST 手写数字

数据集进行实验，噪声数据与有效样本数量的比率设置为 1:1、2:1 和 4:1。其中标号噪声的引入通过随机更改正确标号为其他类别标号，属于完全随机噪声。针对不同比率，有效样本数量设置为 100、500、1000、5000。模型采用 AdaDelta 作为优化器且批量大小设置为 100。对不同噪声比率的模型，学习率的范围设置为 0.01 到 1，并且选取得到最好性能的学习率。图 4.1 显示了分类精度。

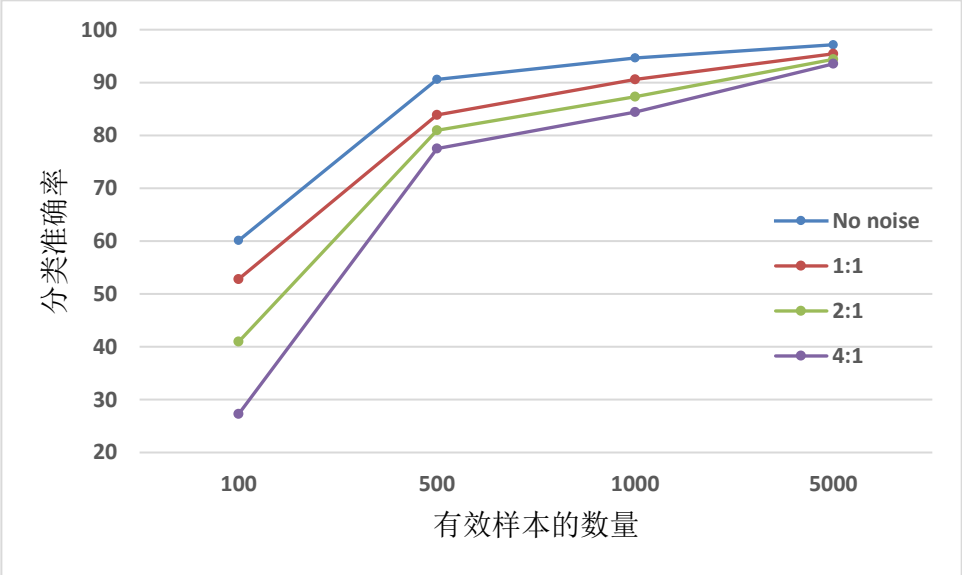


图 4.1 不同噪声水平下的分类性能

从图 4.1 可以看出，在不同的数据规模中，标号噪声的存在都会在一定程度影响分类器的分类性能。具体地，在标号噪声样本与有效样本的比率一定的情况下，当有效样本数量较少时，噪声数据影响较大；而在有效样本数量较多的时候，则能在一定程度上降低噪声数据带来的影响。具体地，可以看出，当没有噪声数据存在的时候，仅需要 500 个有效样本数据就可达到 90%的分类精度。而在其他比率设置为 1:1、2:1 和 4:1 的情况下，则分别需要 1000、2000 和 3000 左右数量的有效样本。因此，通过扩充有效样本的数量是提高模型特征学习能力以及分类精度的一个重要途径。

4.2.2 数据增广策略

未来大量训练数据的获取费时费力，同时当前的生成对抗网络及其变形模型已经可以生成相当逼真的图片。因此，本节尝试使用生成对抗网络生成特定标号的有效样本来提高特征学习及分类器的分类能力。

具体地，由于实验环境是在标号噪声的背景下，数据的标号信息是无法直接使用的。因此，选取信息最大化的生成对抗网络（InfoGAN）作为基本生成模型。通过引入信息最大化概念使

之可以学习到潜在的可解释性特征。该模型最大的优势在于训练过程是无监督的，不需要额外的标号信息，可以直接适用于标号噪声问题中。基于 InfoGAN 模型的数据增广算法如下所示。

算法 4.1 基于 InfoGAN 模型的数据增广算法

输入：训练样本集 X ，样本类别集合 N ，各类样本待生成数量 n

输出：含样本标签的生成样本集 F

根据训练样本集 X 训练 InfoGAN 网络模型的生成器 G 、判别器 D 与 Q 。

初始化 $i=1$

For $i \le n$

(1)根据离散均匀分布生成类别隐变量 $c_1 \sim \mathrm{Cat}(K=10, p=0.1)$

(2)根据连续均匀分布生成连续隐变量 $c_2 \sim \mathrm{Unif}(-1,1)$

(3)根据标准正态分布生成 62 维随机噪声变量 $z \sim \mathrm{N}(0,1)$

(4)将各变量拼接并作为生成器的输入参数，生成样本 $I_{fake}=G(c_1,c_2,z)$

(5)将生成样本 I_{fake} 及对应标号 c_1 加生成样本集 $\{I_{fake},c_1\} \to F$

(6) $i=i+1$

End For

由于 InfoGAN 模型的训练过程及具体的模型设置不是本节重点内容，因此省略相关内容。详细内容请见本文 2.3.3 节以及相关参考文献[68]。

4.2.3 生成样本可用性验证

为了验证生成样本的效果，本节首先从生成样本可视化的角度进行简要分析。具体地，仍以 MNIST 手写数字集作为实验数据来源，展示了该数据集中的真实样本与通过 InfoGAN 模型生成样本的部分数据。如图 4.2 所示，可以看到，无论从图像数据多样性还是从分辨率高低都很难区分两者的不同（左：真实数据；右：生成数据）。

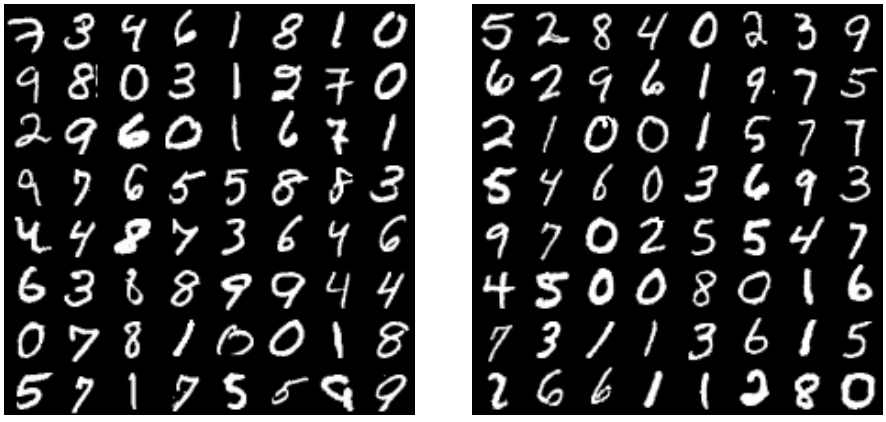


图 4.2 真实数据与生成数据对比

进一步地，分析了类别隐变量c_1、连续隐变量c_2以及随机噪声变量z对生成图像的影响。如图 4.3 所示，每一行数据表示固定类别隐变量c_1和随机噪声变量z不变，从-2 到 2 均匀调整连续隐变量c_2后的生成结果；每一列数据表示固定连续隐变量c_2和随机噪声变量z不变，从 0 到 9 调整类别隐变量c_1后的生成结果。可以看出，InfoGAN 模型在无监督的条件下，准确地学习到了数字类型（类别隐变量c_1）和角度旋转（连续隐变量c_2）等可解释信息。其中，将类别隐变量c_1作为生成样本的标号，给扩大有监督数据集的规模提供了可能。

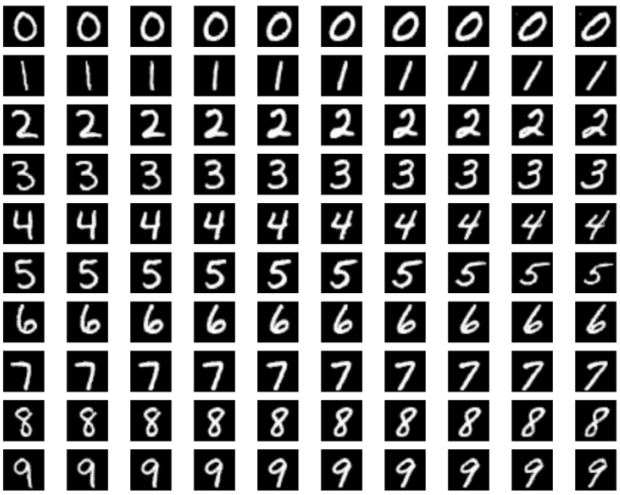


图 4.3 隐变量c_1, c_2影响下的生成样本

同时，我们还分析了随机噪声变量z对生成图像的影响。如图 4.4 所示，每一行数据表示在固定类别隐变量c_1和连续隐变量c_2条件下，随机采样多组噪声变量z后的生成结果。可以看出，噪声变量z的不同取值会生成不同风格的图像，这反应了噪声变量z能够有效提高样本的多样性。该现象也从侧面反映了利用生成数据在扩增样本数量这一问题上的可行性。

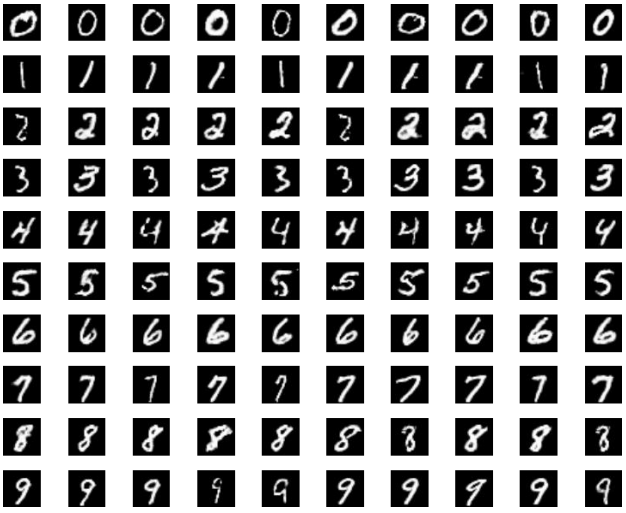


图 4.4 随机噪声变量z影响下的生成样本

另一方面，为了更充分探究生成样本对特征学习以及分类任务的影响，在 MNIST 数据集中进行了验证性的对比实验。具体地，设置了四种训练数据集，分别由 N 个生成样本、N 个真实样本、N 个生成样本与 N 个真实样本混合以及 $2*N$ 个真实样本组成。其中 N 的取值设置为 $\{250,500,1000\}$。将上述训练集应用于两种经典分类模型，分别为 CNN 模型和核 SVM 模型。对于 CNN 模型，选用卷积神经网络 LeNet-5；对于核 SVM 模型，其核函数选用径向基函数（Radial Basis Function，RBF）。两个模型相关的超参数均选取为在测试集下获得最好性能所对应的超参数。

在测试集中的分类精度如图 4.5 和图 4.6 所示。可以看到，两个分类模型得到了较为一致的规律。在样本数量相同的情况下，生成样本的分类精度低于真实样本，这表明了生成样本较真实样本在样本多样性上仍有一定的不足。然而，相较于仅用真实样本，真实样本与生成样本混合使用的训练集其分类精度均有一定的提升，且当样本数量较少的情况下效果更加明显。该现象在一定程度上反映了生成样本能够上丰富原有数据集的多样性，从而进一步提高模型特征的表达能力。

综合上述分析验证，使用 InfoGAN 模型生成特定标号的样本能够在一定程度上起到数据增广的作用，尤其是在样本数据受限的场景中。因此，在后续的实验部分均使用该数据增广策略作为必备的数据预处理步骤，默认增广的样本数量等同于已有的样本数量。

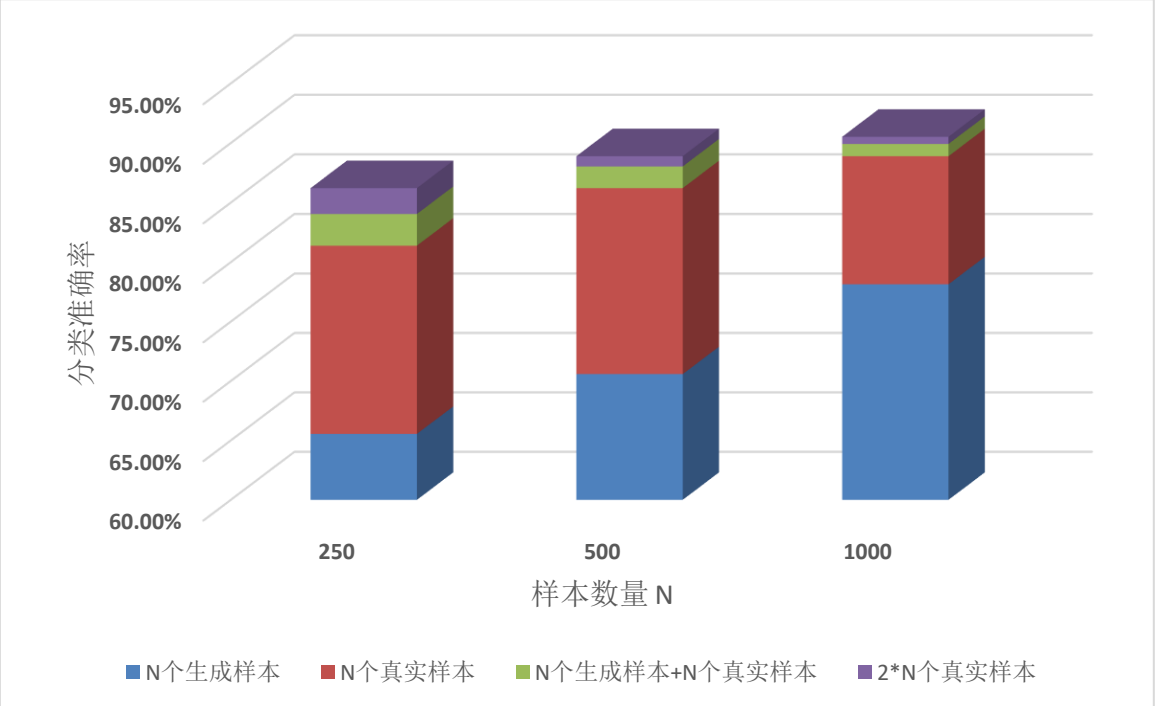


图 4.5 SVM 分类器在测试集上的分类准确率

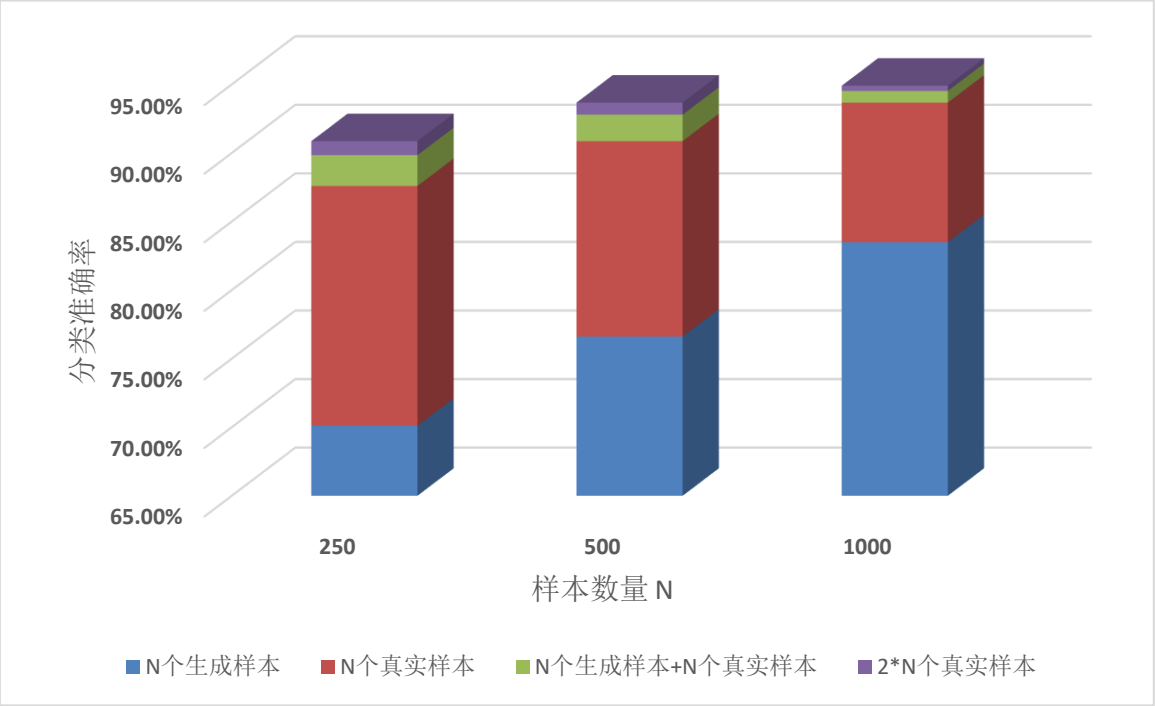


图 4.6 CNN 分类器在测试集上的分类准确率

4.3 基于重要性加权的自编码网络

本节中，首先通过验证性实验来分析标号噪声对自编码网络特征学习过程的影响。然后将重要性加权的优化策略引入传统的自编码网络优化过程。最后，通过验证性实验说明了该加权优化方法的有效性。

4.3.1 标号噪声对自编码网络的影响

为了探究标号噪声对特征学习过程的影响，选取自编码网络作为经典的特征学习模型进行分析。一般地，对于包含 N 个数据共 K 个类别的训练集 $\{x_i, y_i\}, i=1,2...,N$，首先根据它们对应的标签将它们分为不同类别子集（注意到，由于标号噪声的存在，有些样本实际上并不属于相应子集），然后给每个类别子集分别学习一个类内自编码网络，其损失函数如公式（4.1）所示。

$$L(\theta;X)=\frac{1}{n}\sum_{i=1}^{n}\| \hat{x}_i - x_i \|^2 + \lambda \| \theta \|^2 \tag{4.1}$$

整个损失函数分为两部分，其中第一项是子集的平均重建误差，第二项是模型参数正则化项。正则化项可以通过避免模型过拟合的方式在一定程度上减轻标签噪声带来的影响。λ 则用于平衡模型的复杂性和重建能力。这种损失函数可以通过小批量梯度下降优化方法来达到一个局部最优解，如公式（4.2）所示。

$$\theta_{j+1}=\theta_j-\alpha\sum_{k=1}^{m}\frac{1}{m}\nabla_\theta f(x_k) \tag{4.2}$$

其中θ是模型待学习参数，α是学习率，m是每次学习的样本数，$f(\cdot)$是模型的损失函数。可以看出，每次参数θ的更新方向是基于m个样本的平均梯度，这表明每个样本的重要程度在优化过程中是相同的。因此，当数据集中存在标号噪声现象时，错误标记的数据通过影响参数更新的方向，影响了自编码网络的特征学习过程，最终导致分类精度下降的结果。

进一步地，在 MNIST 数据集中进行了验证性实验来分析标号噪声的影响。具体地，在数据集中引入 10%的完全随机噪声并在每个类别中随机选择 100 个数据组成新的数据集。在该数据集上，为每个类别子集分别学习一个类内自编码网络。其中，标号为“1”，“3”和“8”的部分图像数据的重建结果如图 4.7 所示。图中每一对数据表示原始数据和其重构结果，第一行的数据为正确标号的数据，而第二、三行则是错误标记的数据。可以看出，尽管存在 10%的标号噪声，正确标记的图像均被类内自编码网络较好的重构出来；对于错误标记的图像而言，其重构是模糊和不准确的。这也从侧面反映了在较少的标号噪声情况下，该自编码网络的特征学习过程是具有一定的鲁棒性的。

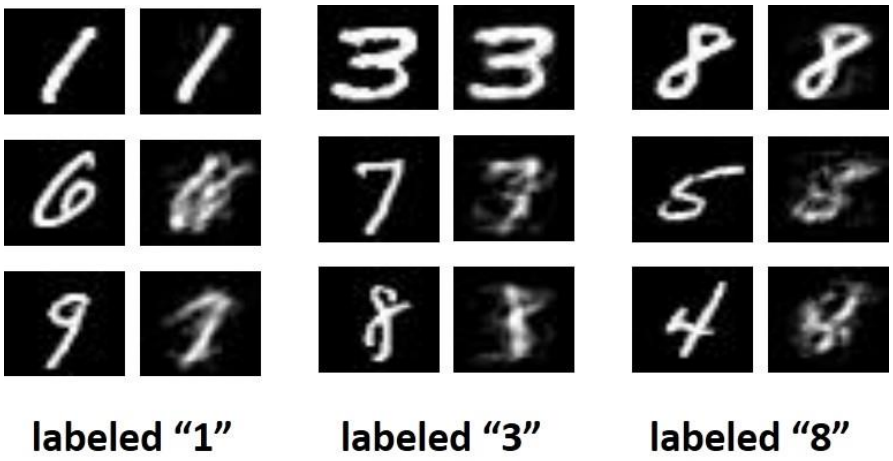


图 4.7 不同标号原始图像和重构图像对比

在此基础上，借助盒形图从统计的角度分析了正确标记的样本和错误标记的样本重建误差分布情况，结果如图 4.8 所示。可以看到，在 10%标号噪声的条件下，正确标记的数据的重构误差在均值和浮动范围方面均比错误标记的数据要低得多，几乎可以完全根据重建误差区分开来。这与图 4.7 的实验结果保持一致。另一方面，也从侧面反应了通过防止过拟合的方式能够在一定程度上缓解标号噪声带来的影响。

为了进一步探究在标号噪声水平较高的情况下自编码器学习的性能，将实验扩展到了含 30%标号噪声的数据集中，在相同的实验设置下统计了数据重建误差分布，如图 4.9 所示。可以看出，相比 10%标号噪声条件下，30%标号噪声下正确标号的数据其重建误差变大且错误标号的

数据其重建误差变小，两类数据的重建误差分布有了明显的重叠。这也表明当噪声水平相对较高时，仅仅借助防止过拟合的方式是完全不够的，仍需通过其他方式进一步提高自编码模型对标号噪声的鲁棒性。

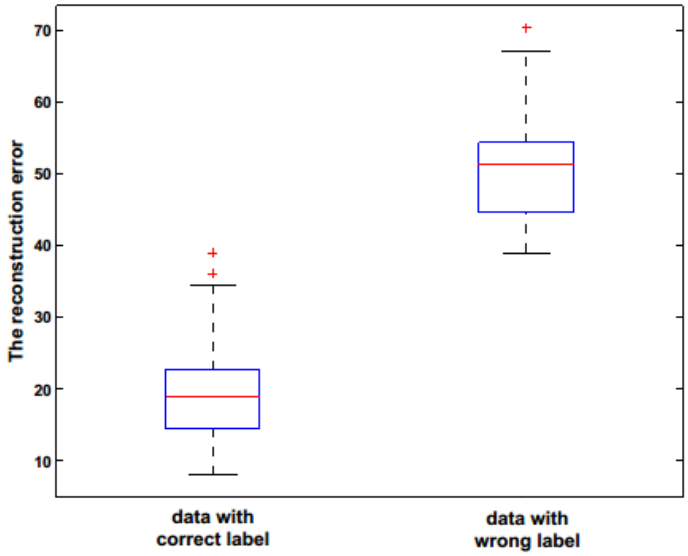


图 4.8 在 10%的标号噪声下数据重构误差的分布

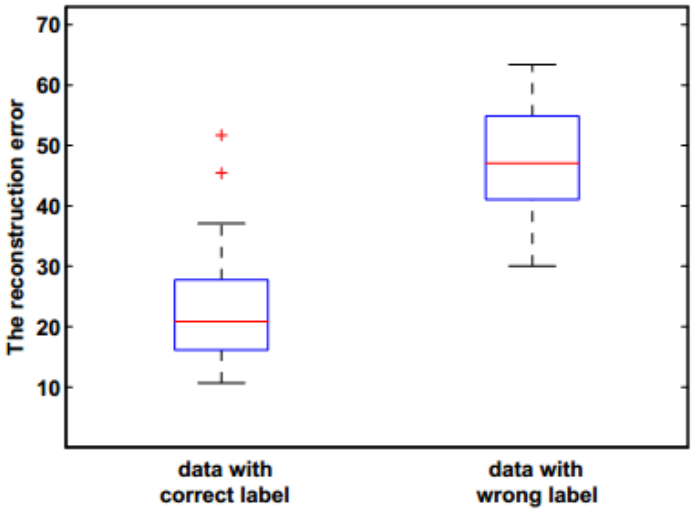


图 4.9 在 30%的标号噪声下数据重构误差的分布

4.3.2 重要性加权策略

根据上一节标号噪声对自编码网络学习影响的分析，需要进一步使用标号噪声鲁棒的方法来提高特性空间学习的质量。因此，考虑到每个数据对模型优化过程中的影响，提出了重要性加权的优化策略。具体地，当更新自编码网络模型参数时，利用加权的小批量梯度下降算法来进行优化，如公式（4.3）所示。

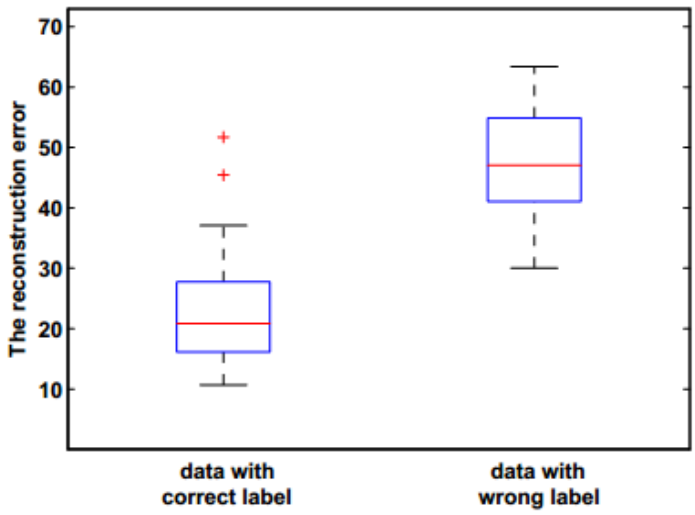
$$\theta_{j+1}=\theta_j-\alpha\sum_{k=1}^{m}w_k\nabla_\theta f(x_k) \tag{4.3}$$

与传统小批量梯度下降优化算法将所有数据的重要性视为一致不同的是，通过引入重要性权重 w_k 来考虑到每个数据对模型优化的影响。由于标号噪声的存在，每个数据对模型学习的影响程度是不同的。对于数据 x_k 来说，其重要性权重 w_k 由公式（4.4）计算得到：

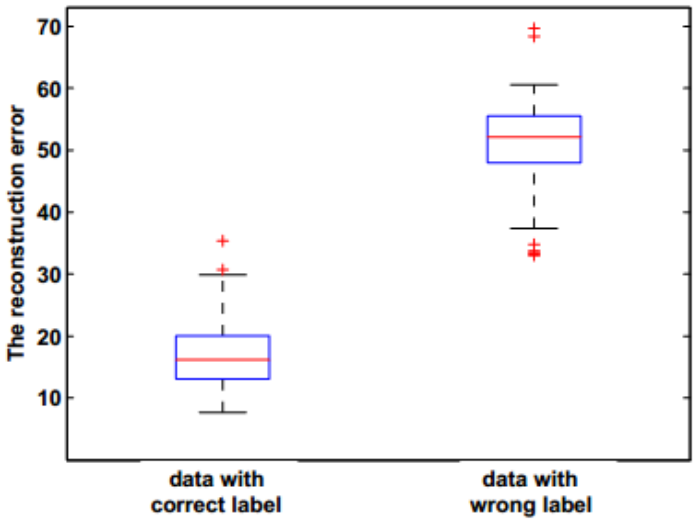
$$w_k=\frac{e^{-\|\hat{x}_k-x_k\|^2}}{\sum_{i=1}^{m}e^{-\|\hat{x}_i-x_i\|^2}} \tag{4.4}$$

可以看出，不同数据点的重要性权重是基于其重建误差的大小来定。数据的重建误差越大，则它们越有可能是带有标号噪声的数据，因此应该在更新梯度的过程中给一个较小的权重，从而减小其对模型的影响程度，反之则反。注意，所提出的方法类似于成本敏感的学习，因为它们都考虑不同样本的不同权重。然而，所提出的方法更强调为一种鲁棒的加权优化方法，在训练过程中，每个样本的重要性权重都是可变的。

通过进一步的实验分析，验证了所提出的优化方法的有效性。具体地，仍考虑标号噪声为30%的情况，分析了重要性加权策略下其数据重建误差的分布结果，并将其与标准优化方法进行比较。注意到，加权优化方法是在原模型的训练结果基础上进行进一步的训练，主要起到模型微调的作用。实验结果如图 4.3 所示。可以看到，相比原始优化结果，重要性加权优化算法的引入使得正确标记的数据其重建误差的均值更小、分布更集中，而错误标记的数据其重建误差的均值更大，分布也更为集中。同时，正确标记与错误标记数据的重建误差分布重叠程度也大大减少，这从一定程度上反应了加权优化策略可以很好地缓解标号噪声带来的问题，学习到更加鲁棒的特征。



(a) 基于原始优化方法的重构结果



(b) 基于重要性加权优化方法的重构结果

图 4.10 在 30%标号噪声下两种优化方法的重构结果

另一方面，在基于重要性加权优化方法的模型训练完成后，按照训练样本权重从大到小的顺序，对部分图像数据进行可视化。结果如图 4.11 所示，其中红色框中的样本是标号噪声数据，每一行是一个类别子集。从图中可以看出，错误标记的样本其权重普遍较低。同时，在正确标记的样本中，外观正常的样本其权重较高，而外观易混淆、有歧义的样本其权重也相对较低。上述实验现象的一致性反应了重要性权重在模型优化过程中能够较好地识别正确标记的样本和错误标记的样本，从而提高自编码网络模型的鲁棒性。

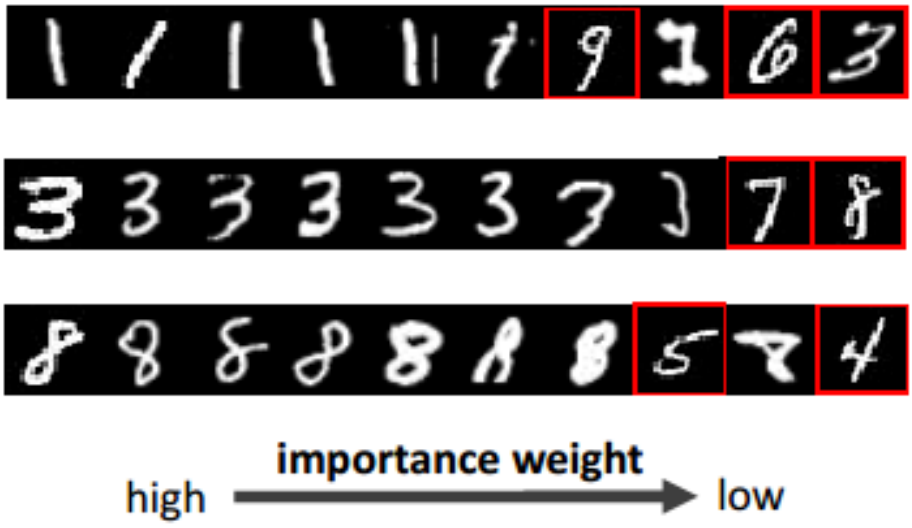


图 4.11 重要度由高到低样本可视化

4.4 类专属自编码网络的特征学习模型

根据前两节所介绍的数据增广策略和重要性加权优化策略，设计了基于自编码网络的类专属特征学习模型。此外，为了更好地利用学到的类专属特征空间，使用重建误差最小化的分类器找出标签噪声并重新标记，将整个过程扩展为迭代的方式进行更加充分的特征学习。其完整的流程如图 4.12 所示。

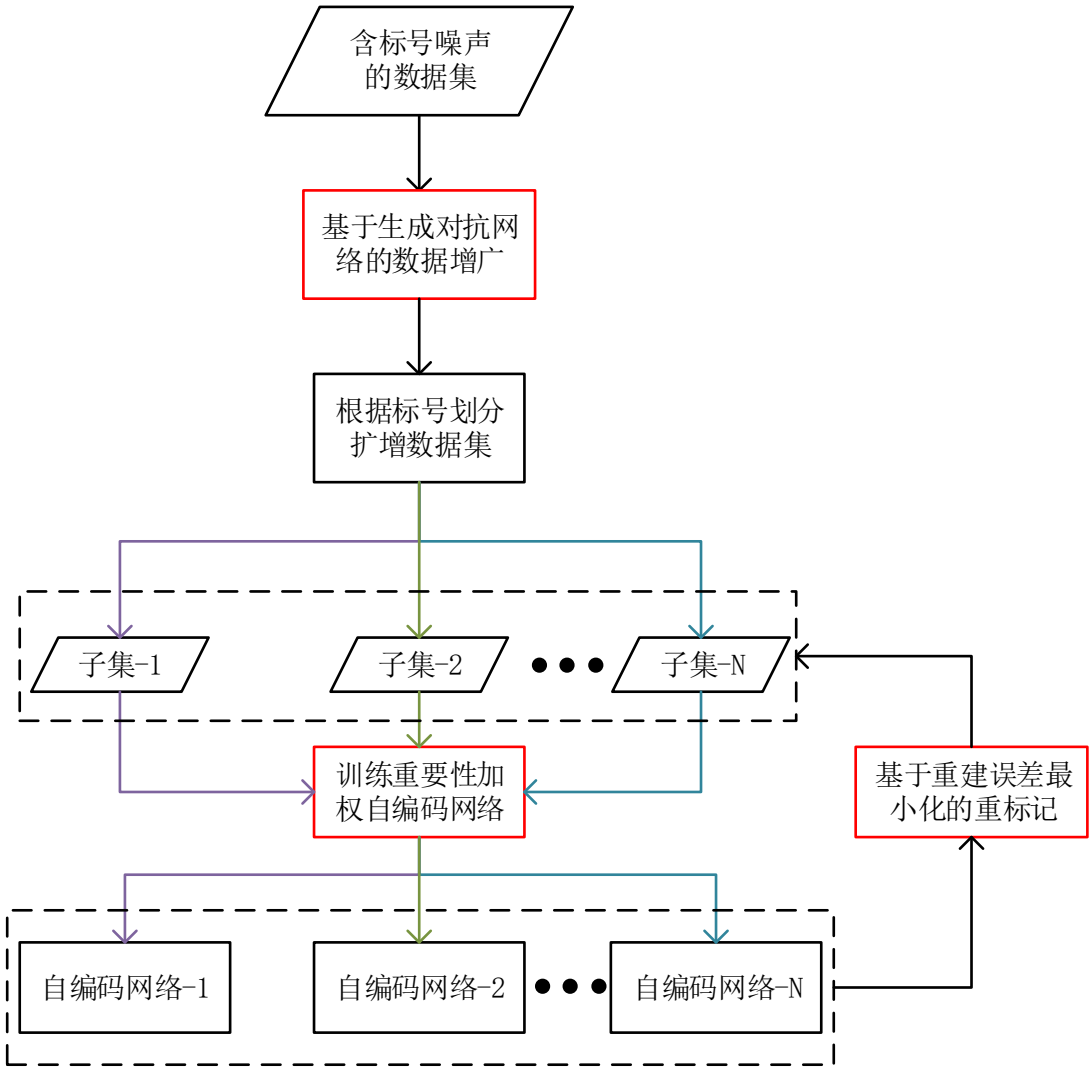


图 4.12 类专属自编码网络特征学习流程图

根据图 4.12 所示，所提出的特征学习模型共有三个核心组件，即数据增广、重要性加权优化以及重新标号并迭代训练。其中，通过生成对抗网络的数据增广组件也会引起一定程度的标号噪声现象，后续步骤的重要性加权与重标记过程同样适用于该生成的数据。对于迭代的重标记过程，根据实验结果分析，通常迭代次数仅需要 1 次可达到较好的效果，即学习两次重要性加权自编码网络，第一次是根据现有标号学习自编码网络，第二次是根据重新标记后的标号来再次学习。因此，如果没有特别说明，本文后续相关实验均默认迭代 1 次。

整个类专属特征学习模型被简称为 DA-WAE-Relabel，为了对比所提出模型的效果，设置了多种参照模型，分别是不进行重新标记迭代过程的特征学习模型 DA-WAE 以及将错误标记的数据剔除后再次进行训练的特征学习模型 DA-WAE-Remove。同时，设置只是用数据增广的模型 DA-AE 作为基线模型。各组件的详细使用情况对比见表 4.1。

表 4.1 各模型组件使用情况

模型	数据增广	重要性加权	迭代重标记	迭代剔除
DA-WAE-Relabel	√	√	√	
DA-WAE-Remove	√	√		√
DA-WAE	√	√		
DA-AE	√			

4.5 实验及结果分析

在本节中，针对所提出的类专属自编码网络特征学习模型，对两个经典任务的进行有效性评估，分别为：1）找出在训练集中的错误标记的数据，即数据清洗任务；2）根据带有标号噪声的训练集对测试集进行分类，即在标号噪声背景下测试集的分类任务。

4.5.1 实验设置

对于参数设置，主要涉及到自编码模型中本身的参数以及其训练阶段的相关参数。具体地，自编码模型的隐含层设置了 200 个神经元，训练过程中模型的正则化权重 λ 设置为 0.1，学习率 α 设置为 0.01，批大小设置为训练样本数量的大小，轮数设置为 100。数据增广的数量等同于原始样本的数量。

针对数据清洗任务，主要使用 DA-AE 和 DA-WAE 两个模型进行实验。而对于在测试集的分类任务，则使用 DA-WAE-Relabel、DA-WAE-Remove 以及 DA-WAE 模型。注意到，DA-AE 和 DA-WAE 两个模型仅进行一次类专属自编码特征学习过程，而 DA-WAE-Relabel 和 DA-WAE-Remove 模型则需要两次特征学习过程。

4.5.2 数据清洗任务

在本节中，将验证所提出方法在数据清洗任务中的性能，也就是识别那些最有可能是标号噪声的数据。特别地，遵循文献中的实验协议[43]，并在两个数据集上进行实验，它们分别是 MNIST 手写数字数据集和 Caltech-10 图像数据集。针对 MNIST 数据集，从两个视觉上容易混

淆的类别数字“7”和数字“9”分别随机选取 1000 张图像样本；而对于 Caltech-10 图像数据集，首先利用先进的无监督特征提取器 C-SVDDNet[84]提取图像数据特征，接着同样选取两个极易混淆的类别“yo-yo”和“roulette-wheel”，并在每个类别上随机选取了 60 个图像样本。进而，在其中分别引入 10%、20%和 30%的完全随机标号噪声。图 4.13 中可视化了两个数据集的部分样本，其中每一行样本属于同一个类别，而红色框中的样本为含标号噪声的图片。为了排除所选择的样本所导致偏差结果的可能性，重复了 30 次实验且每次都对样本进行随机选择。

(a) “7” v.s. ”9”

(b) “yo-yo” v.s. “roulette-wheel”

图 4.13 两个数据集的样本可视化

将所提出的 DA-AE、DA-WAE 两个模型与三种最先进的数据清洗方法进行了比较，分别是 ICCN-SMO[80]、TC-SVM[43]和 ALNR[44]模型。这些方法所涉及到的超参数是根据其对应的论文中选取。注意到，这些方法均是筛选出候选的标号噪声数据并引入领域专家进行进一步判断哪些样本是真正的错误标记的数据。从这个角度来看，所提出的方法由于不假设存在这样的监督信息而是完全无监督的，其通过判断标号与最小重构误差所对应的类别是否一致来确定是否为真正的错误标记的数据，这是基于类专属自编码网络的特征学习能力。

表 4.2 给出了标号噪声数据的平均检测精度。可以看出，虽然没有人类专家的修正，但所提出的方法与最先进的方法取得了相近的结果。这表明，尽管有一定的标号噪声作为干扰，所提出的方法仍能够较好地捕捉到各个类别的专有特征。此外，与 DA-AE 方法相比，DA-WAE 方法的精度更高，这也侧面反映了 DA-WAE 方法中使用的重要性加权优化方法可以减少标号噪声对特征空间学习的影响。

表 4.2 成对混淆类的标号噪声检测性能（%）

成对的类别	Noise level	ICCN-SMO	TC-SVM	ALNR	DA-AE	DA-WAE
7 vs 9	10%	71.38	**98.83**	95.84	97.62	98.77
	20%	78.60	97.21	96.01	95.73	**97.85**
	30%	82.49	95.03	95.69	95.01	**96.82**
yo-yo vs roulette-wheel	10%	54.94	**83.31**	79.67	82.80	83.21
	20%	57.81	80.69	80.29	79.70	**82.46**
	30%	59.33	73.45	70.10	70.41	**77.62**

此外，所提出的方法将会遇到一个额外的问题，即一些正确标记的样本被错误地预测为含标号噪声的样本。为了进一步分析所提出方法的有效性，在 MNIST 手写数字数据集上进行了一个更具有挑战性的验证实验。考虑数据集中所有的类别，并设置训练数据规模的范围为 $\{500,1000,2000,4000\}$。同时，使用三个在数据清洗任务中常见的性能指标进行评估[10]，即 ER1、ER2 以及 NEP。其中 ER1（类型 1 错误）表明了被错误检测的正确标记样本的百分比，而 ER2（类型 2 错误）表明了标号噪声样本没有被检测出的百分比。这两个度量指标越低，所学到的模型的效果越好。对于 NEP（噪声检测的综合精度），它表明了在所有被检测出来的样本中，标号噪声样本所占的百分比。一个 NEP 高的模型意味着一个更有效的模型。

表 4.3 AE 和 WAE 在 MNIST 数据集的 ER1（%）

数据规模	噪声级别		
	10%	20%	30%
N=500	2.67 **2.40**	3.25 **3.00**	4.00 **3.71**
N=1000	0.89 **0.78**	2.38 **2.13**	2.86 **2.43**
N=2000	0.39 **0.39**	1.44 **1.31**	1.86 **1.64**
N=4000	0.28 **0.28**	0.97 **0.84**	1.39 **1.29**

表 4.4 AE 和 WAE 在 MNIST 数据集的 ER2（%）

数据规模	噪声级别		
	10%	20%	30%
N=500	14.00 **14.00**	17.00 **16.00**	18.67 **18.00**
N=1000	11.00 **11.00**	13.00 **12.00**	14.00 **12.67**
N=2000	4.00 **3.50**	8.25 **7.50**	9.33 **8.33**
N=4000	3.50 **3.25**	4.63 **4.13**	6.75 **5.58**

表 4.5 AE 和 WAE 在 MNIST 数据集的 NEP（%）

数据规模	噪声级别		
	10%	20%	30%
N=500	78.18 **79.63**	86.46 **87.50**	89.71 **90.44**
N=1000	91.75 **92.71**	90.16 **91.19**	92.81 **93.91**
N=2000	96.48 **96.50**	94.10 **94.63**	95.44 **95.99**
N=4000	97.47 **97.48**	96.10 **96.60**	96.63 **96.92**

表 4.3-4.5 给出了各个性能指标的结果，这里不考虑使用数据增广的情况。可以看到，无论标号噪声级别和数据规模的大小，所提出的方法其 ER1 性能都小于 4%，这表明所提出的类专属特征表示是准确的，很少改变正确标号的数据。更重要地，可以观察到随着数据规模的增大，ER1 和 ER2 越来越低，NEP 越来越高。这表明当训练样本的规模更大时，所提出的方法更有效。同时，与 AE 相比，WAE 具有更好的性能，特别是在高标号噪声的环境下。这一观测结果充分显示了 WAE 中所提出的重要性加权优化方法的优势。

4.5.3 测试集的分类任务

针对测试集的分类任务，在 MNIST 的手写数字数据集和 Caltech-10 图像数据集进行分类实验，这是两个包含 10 类样本的经典分类基准数据集。在这两个数据集中随机选择每类 60 个图

像数据共 600 个数据，并将它们被划分为相等数量的训练集和测试集。同时，引入不同级别的标号噪声，分别是 10%，20%和 30%。注意，在我们的实验设置中，只有训练数据包含标号噪声，而测试集保持干净。

在所提出的方法中使用 DA-WAE-Relabel、DA-WAE-Remove 以及 DA-WAE 三个模型，其在测试集上的分类器选用重建误差最小分类器。同时，还使用整个训练集（不划分为各个类别子集）来学习全局PCA子空间，并使用K-NN对测试集进行分类。这种方法被命名为PCA-KNN，并被用作基线方法。除此之外，还将所提出的方法与两大类方法进行比较，这两类方法都是为处理标号噪声问题而设计的最先进的方法，具体方法如下所示：

1）标号噪声清洗算法：包括经典的基于 PCA 的异常检测方法（PCA-Outlier）[83]和基于互补的模糊支持向量机的噪声清洗方法（FSVM-Clean）[42]。

2）标号噪声鲁棒算法：包括 L1 正则化度量学习（L1-norm）[81]、贝叶斯度量学习（BML）[82]以及鲁棒的近邻成分分析（RNCA）[47]。

由于实验中的训练样本是随机选择的，因此重复 30 次来排除有偏差结果的可能性。在每个实验中，被比较方法的相关超参数都是通过 10 折交叉验证来选择的。图 4.14 和图 4.15 给出了测试集的平均精度和标准偏差。

从图中可以看出，在低水平和高水平的标号噪声下，DA-WAE-Relabel 方法相比其他被比较的方法均获得了最好的平均分类精度，这表明了整个特征学习框架的有效性。特别地，对比 DA-WAE-Remove 方法和 DA-WAE-Relabel 方法，可以发现相比去掉标号噪声数据，对含标号噪声数据进行重新标号促进了自编码网络特征学习能力。此外，当训练集不含标号噪声的时候，所提出的方法 DA-WAE 得到了最好的平均分类精度，这表明基于最小重建误差的分类器是一个合适的选择。另一方面，当噪声水平相对较低时，基线 PCA-KNN 方法可以在一定程度上容忍标号噪声，但如果当噪声水平超过某个阈值（例如，20%），它的预测性能就会显著下降。与基线 PCA-KNN 方法相比，第一类方法即 PCA-Outlier 和 FSVM-Clean 方法的预测精度更高，因为这些算法中包含处理错误标记数据的内在机制。对于第二种方法，也就是标号噪声鲁棒的算法，相比第一类方法，它们在 10%的标号噪声水平上取得了更好的效果。然而，这些方法在相对较高的噪声水平（20%和 30%）上表现较差，特别是在 30%的标号噪声中，这体现了在高标号噪声下这类方法获得可靠的点估计是相当困难的。

此外，由于图 4.14 和图 4.15 所提供的结果彼此接近，因此进一步统计了 DA-WAE-Relabel 方法和其他被比较方法之间的成对 t-test 检验。表 4.5 显示了 p-value 的实验结果。具体地，在 10%标号噪声水平下，所提出的方法与 RNCA 比较；而在 20%和 30%的标签噪声中，所提出的方法与 FSVM 比较。（因为 RNCA 和 FSVM 在其相应的标号噪声级别上分别获得了第二好的性能，除了所提出的方法外）。可以看出，在 10%标号噪声的情况下，p-value 的值小于 0.001，这

意味着所提出的方法与RNCA相比有绝对的改进。而对于标号噪声较高的情况（即20%和30%），p-value值也小于0.05，这表明所提出的方法和FSVM-Clean之间的差异具有统计学意义，即所提出方法相比其他方法确实有一定程度的提高。

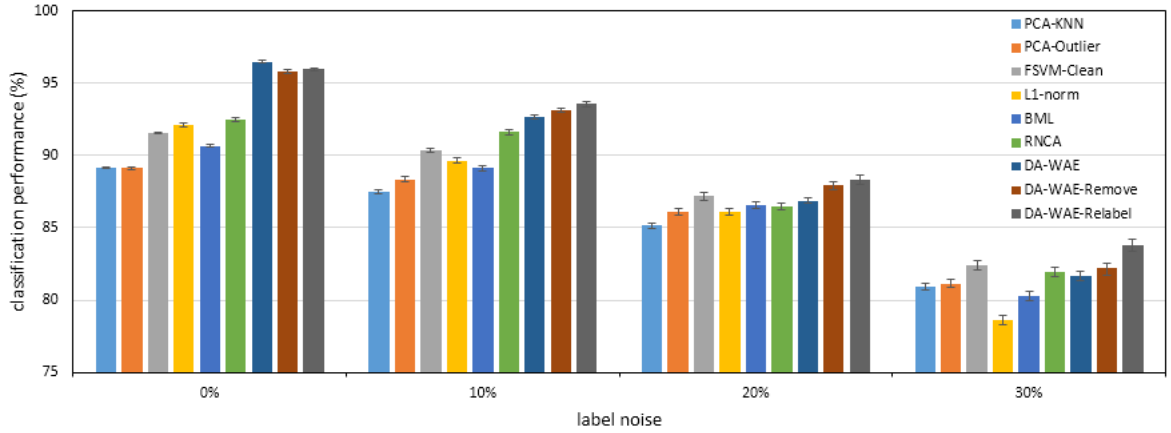


图 4.14 MNIST 手写数字数据集的分类精度（%）

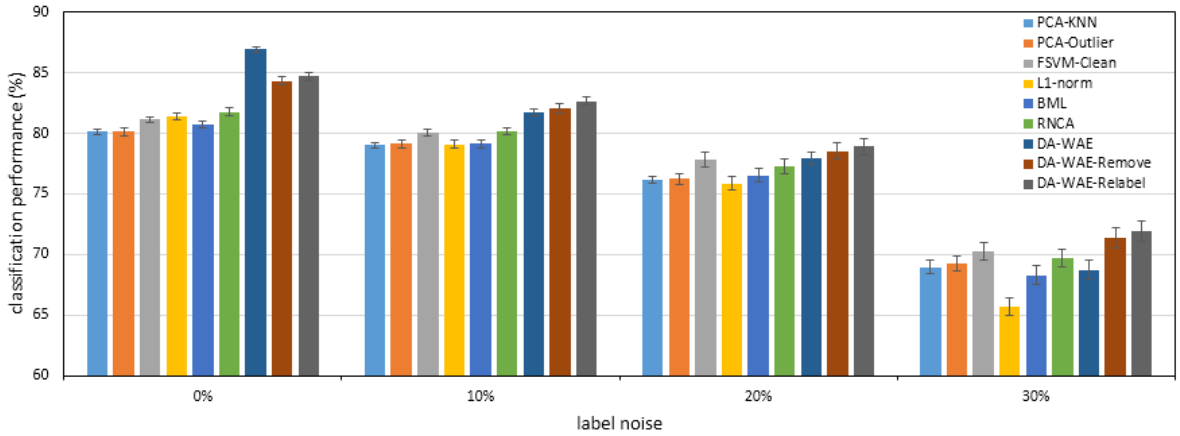


图 4.15 Caltech-10 图像数据集的分类精度（%）

表 4.5 不同模型的 t-test 结果

对比方法（噪声水平）	P-value	
	MNIST	Caltech-10
DA-WAE-Relabel v.s. RNCA（10%）	<0.001	<0.001
DA-WAE-Relabel v.s. FSVM-Clean（20%）	0.001	0.010
DA-WAE-Relabel v.s. FSVM-Clean（30%）	0.003	0.017

4.6 本章总结

本章提出了一个基于类专属自编码网络的特征学习框架。特别地，该框架适用于数据集中存在标号噪声或监督数据有效数量较少的场景中。整个框架主要由三个组件构成，第一个是基于信息最大化生成对抗网络的数据增广策略，其可在无监督的条件下生成带有特定标号的样本；第二个是基于重要性加权的优化策略，通过该优化策略可以降低标号噪声对自编码模型训练过程的影响；第三个是基于重建误差最小化的迭代过程，该迭代过程将其认为的标号噪声数据重新标号并进行迭代训练。验证性实验结果表明，所提出的三个组件均对在标号噪声背景下的特征学习有一定的帮助。在 MNIST 手写数字数据集和 Caltech-10 图像数据集中进行了广泛的对比实验，结果表明所提出的方法在数据清洗任务和测试集中的分类任务均优于当前文献中最先进的方法。

第五章 总结与展望

5.1 工作总结

本文针对标号噪声背景下图像数据的清洗和特征学习问题进行了深入的研究。基于改进的自编码网络学习对标号噪声鲁棒的特征空间并完成标号噪声检测、清洗以及在测试集中的分类等主要任务。

本文首先交代了所研究问题的一般背景、目的以及研究意义，并进一步详细介绍了标号噪声处理问题的相关概念和主要研究工作。其次介绍了本文中所涉及到的核心技术，分别是标号噪声处理技术、深度自编码网络以及生成对抗网络等经典的模型与方法。

标签噪声是对大规模数据集进行监督学习时的常见现象。通过异常检测方法处理这个现象是最近提出的一种处理这一问题的方法，它将每个类的异常值当作具有标号噪声的潜在数据点并在训练前去除它们。然而，这种方法可能会导致较高的假阳性率并降低预测性能。因此，结合异常检测和重构误差最小化的优点，提出了一种新的、有效的方法来解决这一问题。主要思想是添加额外的第二个步骤(即重建误差最小化机制)来验证异常检测的结果是否为真正的含标号噪声数据，以减少因丢弃那些不适合底层数据分布但标号正确的数据点的风险。特别的，首先通过一个基于鲁棒的深度自编码网络的异常检测算法在每个类中找到异常值。通过这个检测算法，不仅得到了候选的错误标号数据，而且还得到了一组学习良好的深度自编码网络。然后将基于重建误差最小化的方法应用于这些异常值，以进一步过滤和重新标记错误标签的数据。在 MNIST 手写数字数据集的实验结果表明，该方法可以显著降低异常检测的假阳性率，提高在数据清洗和分类任务中的性能。

进一步地，为了解决因标号噪声的存在而使有效样本数量减少对自编码网络特征学习过程带来的影响，提出了基于鲁棒的类专属自编码的特征学习方法。首先通过验证性实验验证有效样本数量对特征学习以及分类任务的影响，进而探究基于信息最大化生成对抗网络的数据增广技术的可用性。另一方面，通过重建误差的角度分析标号噪声对其分布的影响，并提出基于重要性加权的优化策略，该策略可以进一步减少标号噪声对自编码网络的特征学习的影响。最后，结合上述的数据增广和重要性加权两个策略，并根据上一章提到的重建误差最小化的思想，建立了鲁棒的类专属自编码网络的特征学习框架。在经典的 MNIST 手写数字数据集和 Caltech-10 图像数据集的实验分析表明，所提出的策略和方法均能在一定程度上提高自编码网络特征学习的能力。同时，在不同的标号噪声水平下，其训练数据的清洗任务以及测试集中的分类任务均比当前最先进的方法获得了可比甚至更好的性能。

5.2 未来展望

本文提出的基于异常检测技术和重建误差最小化的数据清洗方法以及鲁棒的类专属自编码网络特征学习模型仍有很多可以优化和改进的地方。未来可能的研究重点与热点主要有以下几个方面：

（1）异常检测技术

本文虽然提出了利用重建误差最小化的思想将异常检测技术所得到的候选标号噪声数据进行进一步筛选且在检测性能上相比只使用异常检测技术效果要好，然而，由于该方法基于异常检测技术的检测结果，因此其最终的性能受到了一定的限制，更多的取决于异常检测技术能否把绝大多数标号噪声数据检测出来。另一方面，如何充分分析并利用异常检测结果也是一个很重要的问题。因此，更好的异常检测技术等待着我们去研究与发现。

（2）标号噪声下的特征学习方法

本文提出的鲁棒的类专属自编码网络特征学习框架是将多种策略结合以最大化减少标号噪声带来的不利影响。相比其中利用到的数据增广技术和重要性加权优化技术具有一定的普适性，其特征学习基本模型反而可以进行更加充分的探索，如将自编码网络扩展到深度自编码网络或卷积形式的自编码网络模型。通过提高基模型的特征学习能力并应用有效的对标号噪声鲁棒的技术和方法是解决这一问题的关键。

（3）研究拓展

针对在大规模数据集中存在标号噪声现象这一实际问题，其研究范围非常广泛，研究内容非常丰富。相比构造包含完全随机的标号噪声数据集，今后也可以针对真正的标号噪声数据集（如通过搜索引擎搜索获得的数据集或通过众包过程获得的数据集）进行更加有挑战性、应用性的研究工作。

参考文献

[1] Xing J, Li K, Hu W, et al. Diagnosing deep learning models for high accuracy age estimation from a single image[J]. Pattern Recognition, 2017, 66.

[2] Majumder N, Poria S, Gelbukh A, et al. Deep Learning-Based Document Modeling for Personality Detection from Text[J]. IEEE Intelligent Systems, 2017, 32(2):74-79.

[3] Matthews I, Matthews I, Matthews I, et al. A deep learning approach for generalized speech animation[J]. Acm Transactions on Graphics, 2017, 36(4):93.

[4] Krishna R A, Hata K, Chen S, et al. Embracing error to enable rapid crowdsourcing[C]//Proceedings of the 2016 CHI conference on human factors in computing systems. ACM, 2016: 3167-3179.

[5] Li W, Wang L, Li W, et al. Webvision database: Visual learning and understanding from web data[J]. arXiv preprint arXiv:1708.02862, 2017.

[6] Pechenizkiy M, Tsymbal A, Puuronen S, et al. Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction[C]// IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, 2006:708-713.

[7] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, 2009: 248-255.

[8] Wang R Y, Storey V C, Firth C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge & Data Engineering, 1995 (4): 623-640.

[9] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases[J]. AI magazine, 1996, 17(3): 37.

[10] Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE transactions on neural networks and learning systems, 2014, 25(5): 845-869.

[11] Smets P. Imperfect information: Imprecision and uncertainty[M]//Uncertainty management in information systems. Springer, Boston, MA, 1997: 225-254.

[12] Hickey R J. Noise modelling and evaluating learning from examples[J]. Artificial Intelligence, 1996, 82(1-2):157-179.

[13] Brazdil P, Clark P. Learning from imperfect data[M]//Machine Learning, Meta-Reasoning and Logics. Springer, Boston, MA, 1990: 207-232.

[14] Dawid A P, Skene A M. Maximum likelihood estimation of observer error-rates using the EM algorithm[J]. Applied statistics, 1979: 20-28.

[15] Snow R, O'Connor B, Jurafsky D, et al. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks[C]//Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008: 254-263.

[16] Ipeirotis P G, Provost F, Wang J. Quality management on Amazon Mechanical Turk[C]// ACM SIGKDD Workshop on Human Computation. ACM, 2010:64-67.

[17] Sculley D, Cormack G V. Filtering Email Spam in the Presence of Noisy User Feedback[C]// CEAS 2008 - The Fifth Conference on Email and Anti-Spam, 21-22 August 2008, Mountain View, California, USA. DBLP, 2008.

[18] Rantalainen M, Holmes C C. Accounting for control mislabeling in case–control biomarker studies[J]. Journal of proteome research, 2011, 10(12): 5562-5567.

[19] Angluin D, Laird P. Learning from noisy examples[J]. Machine Learning, 1988, 2(4):343-370.

[20] Lachenbruch P A. Discriminant Analysis When the Initial Samples Are Misclassified II: Non-Random Misclassification Models[J]. Technometrics, 1974, 16(3):419-424.

[21] Nettleton D F, Orriols-Puig A, Fornells A. A study of the effect of different types of noise on the precision of supervised learning techniques[J]. Artificial Intelligence Review, 2010, 33(4):275-306.

[22] Hanner R, Becker S, Ivanova N V, et al. FISH-BOL and seafood identification: geographically dispersed case studies reveal systemic market substitution across Canada.[J]. Mitochondrial Dna, 2011, 22 Suppl 1(Supp 1):106.

[23] Freund Y, Schapire R, Abe N. A short introduction to boosting[J]. Journal-Japanese Society For Artificial Intelligence, 1999, 14(771-780): 1612.

[24] Quinlan J R. Induction of decision trees[M]. Kluwer Academic Publishers, 1986.

[25] Brodley C E, Friedl M A. Identifying mislabeled training data[J]. Journal of Artificial Intelligence Research, 2011, 11(1):131--167.

[26] Libralon G L, Lorena A C. Pre-processing for noise detection in gene expression classification data[J]. Journal of the Brazilian Computer Society, 2009, 15(1):3-11.

[27] Bross I. Misclassification in 2 X 2 Tables[J]. Biometrics, 1954, 10(4):478-486.

[28] Hout A V D. Randomized Response, Statistical Disclosure Control and Misclassification: A Review[J]. International Statistical Review, 2002, 70(2):269-288.

[29] Zhang W, Rekaya R, Bertrand K. A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer[M]. Oxford University Press, 2006.

[30] Shanab A A, Khoshgoftaar T M, Wald R. Robustness of Threshold-Based Feature Rankers with Data Sampling on Noisy and Imbalanced Data[C]//FLAIRS Conference. 2012.

[31] Brodley C E, Friedl M A. Identifying and eliminating mislabeled training instances[C]// Thirteenth National Conference on Artificial Intelligence. AAAI Press, 1996:799-805.

[32] Verbaeten S, Assche A V. Ensemble Methods for Noise Elimination in Classification Problems[M]// Multiple Classifier Systems. Springer Berlin Heidelberg, 2003:317-325.

[33] Gaba A, Winkler R L. Implications of Errors in Survey Data: A Bayesian Model[J]. Management Science, 1992, 38(7):913-925.

[34] Manwani N, Sastry P S. Noise Tolerance Under Risk Minimization[J]. IEEE Transactions on Cybernetics, 2013, 43(3):1146-1151.

[35] Thathachar M A L, Sastry P S. Networks of Learning Automata: Techniques for Online Stochastic Optimization[J]. Kluwer Academic Publishers, 2004.

[36] Sastry P S, Nagendra G D, Manwani N. A Team of Continuous-Action Learning Automata for Noise-Tolerant Learning of Half-Spaces[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2009, 40(1):19-28.

[37] Beigman E, Klebanov B B. Learning with Annotation Noise.[C]// ACL 2009, Proceedings of the, Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing of the Afnlp, 2-7 August 2009, Singapore. DBLP, 2009:280-287.

[38] Abellán J, Masegosa A R. Bagging decision trees on data sets with classification noise[C]// International Conference on Foundations of Information and Knowledge Systems. Springer-Verlag, 2010:248-265.

[39] Abellán J, Masegosa A R. An Experimental Study about Simple Decision Trees for Bagging Ensemble on Datasets with Classification Noise[C]// European Conference on Symbolic and Quantitative Approaches To Reasoning with Uncertainty. Springer-Verlag, 2009:446-456.

[40] Dietterich T G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization[J]. Machine Learning, 2000, 40(2):139-157.

[41] Jeatrakul P, Wong K W, Fung C C. Data cleaning for classification using misclassification analysis[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2010, 14(3): 297-302.

[42] Pruengkarn R, Wong K W, Fung C C. Data cleaning using complementary fuzzy support vector machine technique[C]//International Conference on Neural Information Processing. Springer, Cham, 2016: 160-167.

[43] Fefilatyev S, Shreve M, Kramer K, et al. Label-noise reduction with support vector machines[C]//Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012: 3504-3508.

[44] Ekambaram R, Fefilatyev S, Shreve M, et al. Active cleaning of label noise[J]. Pattern Recognition, 2016, 51: 463-480.

[45] Teng C M. A Comparison of Noise Handling Techniques[C]//FLAIRS Conference. 2001: 269-273.

[46] Koplowitz J, Brown T A. On the relation of performance to editing in nearest neighbor rules[J]. Pattern Recognition, 1981, 13(3): 251-255.

[47] Wang D, Tan X. Robust Distance Metric Learning in the Presence of Label Noise[C]//AAAI. 2014: 1321-1327.

[48] Ganapathiraju A, Picone J. Support vector machines for automatic data cleanup[C]//Sixth International Conference on Spoken Language Processing. 2000.

[49] Zhang T. An introduction to support vector machines and other kernel-based learning methods[J]. AI Magazine, 2001, 22(2): 103.

[50] Kowalczyk A, Smola A J, Williamson R C. Kernel machines and boolean functions[C]//Advances in Neural Information Processing Systems. 2002: 439-446.

[51] Domingo C, Watanabe O. MadaBoost: A modification of AdaBoost[C]//COLT. 2000: 180-189.

[52] Oza N C. Boosting with averaged weight vectors[C]//International Workshop on Multiple Classifier Systems. Springer, Berlin, Heidelberg, 2003: 15-24.

[53] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533.

[54] Kamimura R, Nakanishi S. Feature detectors by autoencoders: decomposition of input patterns into atomic features by neural networks[J]. Neural Processing Letters, 1995, 2(6): 17-22.

[55] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1096-1103.

[56] Makhzani A, Frey B. K-sparse autoencoders[J]. arXiv preprint arXiv:1312.5663, 2013.

[57] Rifai S, Vincent P, Muller X, et al. Contractive auto-encoders: Explicit invariance during feature extraction[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. Omnipress, 2011: 833-840.

[58] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[59] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of machine learning research, 2010, 11(Dec): 3371-3408.

[60] Maria J, Amaro J, Falcao G, et al. Stacked autoencoders using low-power accelerated architectures for object recognition in autonomous systems[J]. Neural Processing Letters, 2016, 43(2): 445-458.

[61] Gupta K, Majumdar A. Imposing Class-Wise Feature Similarity in Stacked Autoencoders by Nuclear Norm Regularization[J]. Neural Processing Letters, 2017: 1-15.

[62] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[63] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International Conference on Machine Learning. 2017: 214-223.

[64] Liu M Y, Tuzel O. Coupled generative adversarial networks[C]//Advances in neural information processing systems. 2016: 469-477.

[65] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[66] Gauthier J. Conditional generative adversarial nets for convolutional face generation[J]. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014, 2014(5): 2.

[67] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans[J]. arXiv preprint arXiv:1610.09585, 2016.

[68] Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]//Advances in neural information processing systems. 2016: 2172-2180.

[69] Xiong H, Pandey G, Steinbach M, et al. Enhancing data analysis with noise removal[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(3): 304-319.

[70] Lukashevich H, Nowak S, Dunker P. Using one-class SVM outliers detection for verification of collaboratively tagged image training sets[C]//Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. IEEE, 2009: 682-685.

[71] Hawkins D M. Identification of outliers[M]. London: Chapman and Hall, 1980.

[72] Schölkopf B, Smola A J. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. MIT press, 2002.

[73] Chalapathy R, Menon A K, Chawla S. Anomaly Detection using One-Class Neural Networks[J]. arXiv preprint arXiv:1802.06360, 2018.

[74] Zhou C, Paffenroth R C. Anomaly detection with robust deep autoencoders[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 665-674.

[75] Boyd S, Vandenberghe L. Convex optimization[M]. Cambridge university press, 2004.

[76] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.

[77] Boyle J P, Dykstra R L. A method for finding projections onto the intersection of convex sets in Hilbert spaces[M]//Advances in order restricted statistical inference. Springer, New York, NY, 1986: 28-47.

[78] Abadi M, Barham P, Chen J, et al. Tensorflow: a system for large-scale machine learning[C]//OSDI. 2016, 16: 265-283.

[79] Liu F T, Ting K M, Zhou Z H. Isolation forest[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 413-422.

[80] Rebbapragada U D. Strategic targeting of outliers for expert review[D]. Tufts University, 2010.

[81] Wang H, Nie F, Huang H. Robust distance metric learning via simultaneous l1-norm minimization and maximization[C]//International Conference on Machine Learning. 2014: 1836-1844.

[82] Yang L, Jin R, Sukthankar R. Bayesian active distance metric learning[J]. arXiv preprint arXiv:1206.5283, 2012.

[83] Vidal R , Ma Y , Sastry S . Generalized principal component analysis (GPCA)[J]. IEEE Trans Pattern Anal Mach Intell, 2005, 27(12):1945-1959.

[84] D. Wang and X. Tan, “Unsupervised feature learning with C-SVDDNet,” Pattern Recognit., vol. 60, pp. 473–485, Dec. 2016.

致 谢

时间飞逝，一不留神，硕士的学习生涯也将告一段落。尽管在南航已经呆了近七个年头，细细回想起来，在南航的每一天都是很幸福的。同时，相较七年前的我，现在的我在做人处事、专业知识上都有了不小的进步。在这篇论文的完成之际，我谨向曾经帮助和关心过我的老师、同学和亲友们表示感谢。

首先，我要感谢的是授予我专业知识和学习方法的老师们，尤其是我的研究生导师谭晓阳教授。他对学术的热爱深深的影响了我，让我真正体会到机器学习和计算机视觉这一领域的博大精深，让我感受到学习并运用知识的快乐。他渊博的专业知识、浓厚的学术热情和严谨的治学态度让我知道做学术的真正内涵，在学业和研究中为我指明了方向，培养了我们良好的学习习惯和学术态度。感谢谭晓阳对我的谆谆教诲和悉心指导，您付出的努力和耐心我都铭记于心。感谢 PARNEC 实验室的陈松灿教授、张道强教授、刘学军教授等，是你们营造的浓厚学术氛围，让我能心无旁骛的畅游在学术的海洋，让我有更高的学术视野。祝愿各位老师工作顺利，阖家欢乐。

其次，我想要感谢陪伴我的同学和朋友们。感谢舍友周杨溟、廉震、何康亚、韦翔宇、胡梦磊、周剑刚、程浩，无论是生活娱乐还是学术交流都让我受益匪浅，宿舍里简单快乐的日子总是让人难以忘怀。感谢王冬、王宇辉、金鑫、谭晓松、蔡雅薇、刘程、孙强、张文和宋歌师兄，是你们让我很快融入到实验室这个大家庭并在学术上给予我耐心且专业的帮助。感谢同门魏文革和陈泳杰同学，感谢谢烟平、潘陈听、王炳璇师弟，无论是学术中的解疑答惑还是生活中的关心爱护，无论是聚餐时的幽默谈笑还是狼人杀时的睿智表演，都是最珍贵的回忆。祝愿各位能够过上自己想要的生活。

最后，我要感谢我的父母和家人，感谢你们无微不至的关怀和爱护。无论成功或者失败，你们都永远温暖着我，鼓励着我。我一直相信，家是我所有奋斗路上力量的源泉，是我前行路上的明灯。是父母默默的支持让我无忧无虑的度过了美丽的求学时光，并且让我选择未来喜欢的路。感谢我的女朋友杜婧涵，是你让我在困难的日子里选择坚持，是你让我的每一天都快乐美好，希望未来的每一天里都能有你。

再次感谢陪伴和帮助过我的所有老师、同学、朋友和家人！谢谢你们！

在学期间的研究成果及发表的学术论文

攻读硕士学位期间发表（录用）论文情况

1. **Weining Zhang**, Dong Wang, Xiaoyang Tan. Data Cleaning and Classification in the Presence of Label Noise with Class-Specific Autoencoder[C] //International Symposium on Neural Networks. Springer, Cham, 2018:256-264.（LNCS 录用）

2. **Weining Zhang**, Dong Wang, Xiaoyang Tan. Robust Class-Specific Autoencoder for Data Cleaning and Classification in the Presence of Label Noise[J]. Neural Processing Letters.（SCI 录用）

3. **Weining Zhang**, Xiaoyang Tan. Combining Outlier Detection and Reconstruction Error Minimization for Label Noise Reduction[C]. //International Conference on Big Data & Smart Computing. 2019.（EI 会议录用）