分类号_____ 密　　级_____

U D C_____ 学校代码_____10689_____

雲南財經大學

YUNNAN UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

基于标签不完全数据的弱监督学习方法

Weakly Supervised Learning Methods with Incomplete Label Data

姓　　名：黄梅

导　　师（职称）：杨智基 （副教授）

申 请 学 位 类 别：学术硕士

专　　业：概率论与数理统计

研　究　方　向：统计机器学习

学 院 （中 心、 所）：统计与数学学院

论文完成时间：2024 年 5 月 27 日

摘要

弱监督是指在训练过程中使用相对较为不精确或不完备的标签信息来指导模型学习。相对于监督学习，弱监督学习更加灵活，并且更符合实际场景中标注数据的获取难度。本文针对以下两种数据情形下的弱监督问题展开研究，分别是少量数据标签已知是负类和少量数据标签已知但不止一个类别。

对于第一种数据情形，深度支持向量描述（Deep Support Vector Data Description，Deep SVDD）是常用的一种基于支持向量机的深度模型，它是一种有效的异常检测方法，特别适用于具有大量正常样本和相对较少异常样本的情况。但是 Deep SVDD 仅使用正类数据建模，没有利用所有的数据信息，这也导致它的分类边界无法体现出间隔最大化的思想。此外，它的求解算法不够精确，使用正类样本经神经网络映射后的均值作为超球中心，然后通过分位数回归估计超球半径，这样得到的分类器参数值是不够精确的。为解决上述问题，本文提出了可解释小球大间隔网络（Interpretable Deep Small Sphere and Large Margin Network，ID-SSLMN）。模型的主要思想是首先利用神经网络将可获得的数据映射到高维空间，然后在高维空间中构建一个超球。这个超球可以将正类样本包裹在球内，将负类样本排除到球外。该模型在训练时加入了少量可获得的负类样本，通过最大化正类样本与负类样本之间的间隔来更加细化分类边界。此外，受可解释神经聚类方法的启发，本文还探索了一种新颖的算法。通过将分类器参数融入到神经网络中，用来解决模型参数问题。通过统一的反向传播来求解网络和分类器的参数。这种方法不仅能同时更加精确求解所有参数，还能让神经网络最后一层参数具有可解释性。本文的算法为基于距离的深度学习方法的参数精确估计提供了新的见解。另外，本文在 2 个模拟数据集，3 个图像数据集和 4 个 UCI 数据集上比较了所提出的方法与其他 7 种方法的曲线下面积（AUC）值。其中 ID-SSLMN 取得了最先进的结果，在 CIFAR10 数据集上的 AUC 值相较于 Deep SVDD 方法平均提升了 22.44%。

对于第二种数据情形，程序性弱监督是较为先进的一类模型，它的关键挑战

是如何有效地聚合不同来源的弱信号。对抗标签学习（Adversial Label Learning，ALL）是一种标签模型和终端模型联合学习的程序性弱监督框架，它的模型性能依赖于分类器模型的参数化，而且该模型需要为不同的数据集寻找合适的误差边界，模型泛化能力较差。本文针对上述情况提出了 L2 对抗标签学习框架（L2 Adversial Label Learning，LALL）。它假设可获得关于数据标签的一些弱信号。模型的主要思想是利用弱信号构建一个可行标签约束空间，然后在这个空间内通过损失最大化学习一个质量最差的标签（对抗标签），再利用对抗标签通过损失最小化学习一个质量最好的分类器。本文使用 L2 损失作为训练的损失函数，然后为约束添加松弛变量，使得模型可以自适应地调整约束边界的大小。此外，本文还考虑弱信号会放弃标记一些样本的情况，这样的设置更符合实际生活可获得的弱信号。最后，本文使用逻辑回归和支持向量机这两种模型作为终端分类器，提出了基于逻辑回归的 L2 对抗标签学习方法（L2 Adversial Label Learning with Logistic Regression，LALL-LR）和基于支持向量机的 L2 对抗标签学习方法（L2 Adversial Label Learning with Support Vector Machine，LALL-SVM）。这两种方法在 7 个数据集上进行了数值实验，其中 LALL-LR 在 MNIST 数据集上相较于原来的 ALL 模型 ACC 值最大提高了 10.45%。

关键字：弱监督学习；标签不完全；神经网络；支持向量机

Weakly Supervised Learning Methods with Incomplete Label Data

Mei Huang
Probability Theory and Mathematical Statistics
Directed by Zhiji Yang

Weak supervision refers to the use of relatively imprecise or incomplete label information to guide the model during the training process. In comparison to supervised learning, weakly supervised learning is more flexible and aligns better with the challenges of acquiring accurately labeled data in practical scenarios. This article investigates weakly supervised learning problems in the following two data scenarios: scenarios where a small amount of data is labeled as negative class, and situations where a small amount of data is labeled, but not limited to a single category.

For the first type of data scenario, Deep Support Vector Data Description (Deep SVDD) is a commonly used deep model based on support vector machines, which is an effective method for anomaly detection, especially suitable for situations with a large number of normal samples and relatively few abnormal samples. However, Deep SVDD only models positive-class data, failing to utilize all the information in the datasets, which also results in its classification boundary not reflecting the idea of maximizing the margin. Additionally, its solving algorithm is not precise enough, using the mean of the positive-class samples mapped through the neural network as the center of the hyper-sphere, and then estimating the radius of the hyper-sphere through quantile regression, which leads to inaccurate values for the classifier parameters. To address these issues, this paper proposes an Interpretable Deep Small Sphere and Large Margin Network (ID-SSLMN). The main idea of the model is to first use a neural network to map the available data to a high-dimensional space, and then construct a hyper-sphere in this high-dimensional space. This hyper-sphere can enclose the positive-class samples inside the sphere and exclude the negative-class samples outside the sphere. The model introduces a small number of available negative-class samples during training to further refine the classification boundary by maximizing the margin between positive-

class and negative-class samples. Furthermore, inspired by interpretable neural clustering methods, this paper also explores a novel algorithm. By incorporating the classifier parameters into the neural network to solve the parameter problem. The parameters of the network and classifier are solved through unified backpropagation. This approach not only allows for more accurate estimation of all parameters simultaneously but also makes the parameters of the last layer of the neural network interpretable. The algorithm proposed in this paper provides new insights for the precise estimation of parameters in distance-based deep learning methods. Additionally, this paper compares the proposed method with seven other methods in terms of Area Under the Curve (AUC) on 2 simulated datasets, 3 image datasets, and 4 UCI datasets. Among them, ID-SSLMN achieved the most advanced results, with an average improvement of 22.44% in AUC value on the CIFAR10 dataset compared to the Deep SVDD method.

For the second type of data scenario, programmatic weak supervision represents a more advanced category of models. Its key challenge lies in effectively aggregating weak signals from different sources. Adversarial Label Learning (ALL) is a framework of programmatic weak supervision that involves joint learning of label models and terminal models. Its model performance relies on the parameterization of the classifier model. Additionally, this model needs to find suitable error boundaries for different datasets, leading to poor generalization ability. In response to these challenges, this paper proposes the L2 Adversarial Label Learning (L2 Adversial Label Learning, LALL) framework. It assumes the availability of some weak signals regarding data labels. The main idea of the model is to construct a feasible label constraint space using weak signals, then learning a worst-quality label (adversarial label) within this space through maximizing loss, and subsequently learning the best-quality classifier by minimizing loss using the adversarial label. The paper uses L2 loss as the training loss function. It then introduces slack variables to the constraints, enabling the model to adaptively adjust the size of the constraint boundary. Furthermore, the paper considers cases where weak signals may abandon labeling some samples, a setting more reflective of weak signals obtainable in real life. Finally, this paper proposes two terminal classifiers using logistic regression and support vector machine (SVM), respectively. These are termed

as L2 Adversial Label Learning with Logistic Regression (LALL-LR) and L2 Adversial Label Learning with Support Vector Machine (LALL-SVM). Both methods are experimented on seven datasets through numerical experiments. Specifically, LALL-LR achieves a maximum improvement of 10.45% in accuracy compared to the original ALL model on the MNIST dataset.

Key words: Weakly supervised learning; Incomplete labels; Neural networks; Support vector machines

目 录

第一章　引言

第一节　　研究背景及意义

统计机器学习涉及多个领域的知识和技术，包括统计学、机器学习、计算机科学、优化理论、信息论和模式识别等，这些领域相互交叉和融合，共同推动了统计机器学习的发展和应用。它主要包括监督学习、非监督学习、强化学习等。在监督学习中，模型通过利用带有标签的训练数据集来学习输入与输出之间的映射关系，然后利用所学到的模式对未标记数据进行预测。其中，感知机（Perceptron）(Rosenblatt, 1958) 是早期用于二元分类的简单神经网络模型。随后，随着支持向量机（Support Vector Machine，SVM）(Cortes 和 Vapnik, 1995) 等方法被引入，使监督学习逐渐成为机器学习的核心研究方向。然而，监督学习需要大量标记的训练数据集，这一过程费时费力且成本高昂。在任务复杂或领域特定时，获取足够数量的标记数据可能成为一个挑战。此外，如果标签数据中存在噪声或错误，监督学习模型可能会学习到不准确的关系，进而导致性能下降。

弱监督学习在训练过程中使用的标签信息相对不完整、不确切或不可靠(Zhou, 2018)。这类学习涵盖了所有训练数据与标签不构成一一对应关系的情况。与监督学习相比，弱监督学习通常不需要大量准确标记的训练数据，因此能够显著降低数据标记的成本。

不完全监督学习指的是模型在训练过程中使用包含大量未标记数据和部分标记数据的情况。在处理标签不完全数据时，通常会忽视或浪费大量未标记的数据。不完全监督学习通过有效利用这些未标记数据，提高数据的利用率和模型的性能。在真实世界中，数据通常不仅仅是海量的，还可能包含大量未知类别或存在噪声。不完全监督学习能够有效应对这些复杂的数据情况。

不完全监督学习面临着处理标签不确定性的挑战，即对未标记数据的标签猜测可能存在不确定性或错误。同时，有效地利用未标记数据进行训练，以提升模型性能，也是一个关键问题。在不同领域之间实现良好的泛化能力，使得模型能够适应新领域的数据，也是不完全监督学习的重要任务。此外，面对数据中的噪声和异常情况，如何使得不完全监督学习模型更具鲁棒性，也是需要解决的问题之一。

总体而言，研究数据标签不完全情况下的不完全监督学习旨在探索如何有效利用部分标记数据和未标记数据。这种方法旨在解决现实场景中标签数据不完整的挑战，从而提高模型性能、降低标记成本，并更好地应对复杂的真实数据情况。

第二节　国内外研究现状

一、监督学习

监督学习的特点在于训练数据集的每个样本都具有准确的标签。其目标是利用这些标记数据训练一个模型，使其能够对新的、未标记的输入数据进行准确的标签预测。在监督学习中，通常存在着判别模型和生成模型两个基本的范畴，它们分别关注于不同的任务和问题 (李航, 2019)。

（一）　判别模型

判别模型的主要任务找到一个函数或模型，能够将输入数据映射到相应的标签或类别。常见的判别模型包括逻辑回归 (Cramer, 2002)、决策树 (Kingsford 和 Salzberg, 2008; Loh, 2011)、SVM (Cortes 和 Vapnik, 1995; Hearst et al., 1998)、K 最近邻 (Peterson, 2009)、随机森林 (Breiman, 2001)、神经网络 (LeCun et al., 1998; Svozil et al., 1997) 等。

SVM 是一种经典的监督学习分类器，其主要目标是找到一个最优超平面，将不同类别的数据分隔开来，主要思想是间隔最大化 (Cortes 和 Vapnik, 1995)。最早的 SVM 版本是线性 SVM，为了能处理非线性问题，引入了核方法，通过映射数据到高维空间，使其在高维空间中线性可分 (Vapnik, 1999)。针对不同类型的数据，可以选择合适的核函数，如线性核、多项式核、高斯核等 (Shawe-Taylor 和 Cristianini, 2004)。标准 SVM 的参数 C 是一个正则化参数，用于控制错误分类的惩罚。v– 支持向量机引入了具有特殊性质的参数 v 来代替 C，它代表了训练集中的支持向量的上限比例，这样更便于参数调优和控制模型的稀疏性 (Schölkopf et al., 2000)。另外还有一些基于 SVM 的改进模型，如双子支持向量机 (Khemchandani 和 Chandra, 2007)、最小二乘支持向量机 (Suykens 和 Vandewalle, 1999)。SVM 可以高效处理高维数据，利用核技巧处理非线性问题。但是，它需要调整核函数和正则化参数 C 来获得最佳性能，而且对于大规模数据集可能需要较长的训练时间。

神经网络 (Gurney, 1997) 是一种受到神经系统启发而设计的计算模型，由大量的人工神经元组成。它通常分为输入层、隐藏层和输出层。隐藏层中通常会引入激活函数（如 Sigmoid、ReLU、Tanh）来进行非线性映射，然后通过前向传播和反向传播进行训练 (Nielsen, 2015)。常见的神经网络结构有：神经网络 (Svozil et al., 1997)、循环神经网络 (Schmidhuber, 1992)、卷积神经网络 (Gu et al., 2018; LeCun et al., 1998) 等。神经网络在解决很多复杂任务上表现出色，但是深度神经网络的训练可能需要大量的时间和计算资源。

判别模型通常关注数据的表征和决策边界，使得它们在高维数据和复杂任务上表现较好。这类模型通常更容易训练和调优，但是它们可能会忽略一些关于数据结构和特征的信息。

（二）　生成模型

生成模型的目标是学习数据的分布，从而能够生成新的具有相似特征的数据样本。这使得它们在图像生成、自然语言处理等任务上表现出色。概率图模型 (Koller 和 Friedman, 2009) 通过图结构表示随机变量之间的依赖关系，是较为古老的生成模型，包括贝叶斯网络 (Friedman et al., 1997) 和马尔可夫随机场 (Cross 和 Jain, 1983) 等。而近几年较为流行的生成模型是变分自编码器 (Kingma 和 Welling, 2014, 2019) 和生成对抗网络 (Goodfellow et al., 2014, 2020)，使用扩散过程的概念来建模数据的扩散生成模型最近在生成模型领域也引起了很大的关注(Dhariwal 和 Nichol, 2021; Nichol 和 Dhariwal, 2021)。另外，流模型 (Dinh et al., 2014; Kingma 和 Dhariwal, 2018) 在生成样本方面也表现出色。

生成模型能够捕捉数据中的隐含结构。但是训练生成模型可能比判别模型更为复杂和耗时。由于没有明确的标准来衡量生成的样本是否符合期望分布，评估生成模型的性能通常更为困难。

二、弱监督学习

弱监督学习是指在训练机器学习模型时，使用的标签信息相对较弱、不完整或噪声较大的情况。依据实际可获得的标签信息种类，可以将弱监督学习分为不完全监督、不确切监督和不准确监督三种 (Zhou, 2018)。

（一）　不准确监督

不准确监督涉及监督信息不完全真实的情况，即标签信息可能会出现错误。一个典型的场景是带有噪声标签的学习 (Frénay 和 Verleysen, 2013)，假设标签受随机噪声影响来进行建模和求解。为进一步提高模型性能，研究者开始求助廉价的标签来源，如远程监督利用外部知识库获取噪声标签 (Hoffmann et al., 2011)。众包标签将任务分配给互联网的人来完成 (Yuen et al., 2011)。另外还有启发式规则 (Awasthi et al., 2020)、特征注释 (Mann 和 McCallum, 2010) 等。这些方法可以给部分样本标注多个噪声标签，所以接下来的任务是将这些标签来源（标签函数）以一种有原则和抽象的方式结合起来。

（二）　不确切监督

不确切监督是指训练数据只给出粗粒度的标签。假设把输入看成很多个包（bag），每个包里有一些数量不一的实例，可以知道包的标签，但是不知道每个实例的标签 (Zhou, 2018)。多实例学习 (Dietterich et al., 1997) 是解决不确切监督问题的常用方法。大多数多实例方法试图使单实例监督学习算法适应多实例表示 (Foulds 和 Frank, 2010)。还有一些方法试图通过表示变换使多实例表示适应于单实例方法 (Zhou, 2006)。

（三）　不完全监督

不完全监督是指训练数据只有一小部分被标注，而其余数据是未标注的，它允许模型在没有完整标签信息的情况下学习并进行预测。解决不完全监督问题的常见方法是主动学习和半监督学习。此外，最近有研究者假设可以获得一些廉价的弱标签来源，提出程序性弱监督。

主动学习假设未标注数据的真实标签可以向“先知”查询，标注成本只与查询次数有关 (Settles, 2009)。主动学习的目标就是最小化查询次数，选择有价值的未标记数据来查询先知。信息量和代表性是两个衡量价值的标准。基于信息量的方法有不确定性抽样（Uncertainty sampling），即训练单个学习器，选择学习器最不确信的样本向先知询问标签信息 (Lewis, 1995)。另一种是投票询问（query by committee），即训练多个学习器，选择各个学习器争议最大的样本向先知询问标签信息 (Abe, 1998; Seung et al., 1992)。基于代表性的方法是采用聚类方法来挖掘未标记数据的聚类结构 (Dasgupta 和 Hsu, 2008; Nguyen 和 Smeulders,

2004)。

半监督学习尝试在不询问人类专家的情况下利用未标记样本 (Xiaojin, 2006; Zhou 和 Li, 2010)。生成方法假设标记数据和未标记数据都是从相同的固有模型生成的 (Miller 和 Uyar, 1996; Nigam et al., 2000)。基于图的方法构造一个图，然后按照一定的标准在图上传递标签信息 (Blum 和 Chawla, 2001; Zhu et al., 2003)。低密度分离方法强制分类边界跨越输入空间中密度较小的区域。最经典的代表是半监督支持向量机 (Joachims, 1999; Li et al., 2013)。协同训练将半监督学习与主动学习相结合，学习多个分类器，并让他们合作开发未标记的数据 (Blum 和 Mitchell, 1998)。

程序性弱监督是一种利用启发性规则、模拟器或程序性生成的标签来进行训练的弱监督学习方法 (Zhang et al., 2022)。标签函数是用户定义的程序，用来编码弱监督源的形式，例如领域专家的知识、领域规则或预训练模型等。不同标签函数之间可能存在相关性，因此指定并考虑适当的独立结构至关重要 (Cachay et al., 2021)。手动指定依赖结构会给研究人员带来额外的负担，因此研究人员试图让模型自动学习依赖结构 (Bach et al., 2017; Varma et al., 2017)。最近，研究人员还探索了自动生成标签函数的可能性 (Varma 和 Ré, 2018)，或者交互生成 (Boecking et al., 2020)。程序性弱监督有两阶段方法和一阶段方法两种类型。Ratner 为两阶段方法开发了标签模型，它首先聚合标签函数的噪声投票产生训练标签，然后用训练标签训练下游任务的终端模型 (Ratner et al., 2017; Ratner et al., 2016)。一阶段方法旨在以端到端的方式训练标签模型和终端模型，使它们能够相互增强。Tonolini et al. (2023)提出从输入和弱标签中联合学习，以捕捉具有潜在空间的输入信号分布。Boecking (2023)将程序性弱监督与生成对抗网络进行了融和。此外，Mazzetto et al. (2021)和 Arachie 和 Huang (2021)将弱监督分类问题表述为约束最小-最大优化问题。

可获得部分标签信息的不完全监督可以分为两种情况，分别是少量样本标签已知是负类和少量样本标签已知且不止一个类别。Wu 和 Ye (2009)提出的小球大边界方法（Small Sphere and Large Margin，SSLM）是解决第一种情况的一个经典模型，他训练一个超球，将正类样本包裹在球内，负类样本排除在球外，已知类别的少量负类样本被用来细化分类边界。这与支持向量描述方法（Support Vector Data Description，SVDD）的思想类似，区别在于 SVDD 是一个无监督模

型，即训练中未使用负类样本 (Tax 和 Duin, 2004)。而对于少量样本标签已知类别的情况，Karamanolakis et al. (2021)利用所有可用数据，构造了一个半监督学习目标，学习一个端到端的模型。Mazzetto et al. (2021)是使用对抗训练的方式来交替更新分类器和标签模型。对抗标签学习（Adversial Lable Learning，ALL）也是一个对抗训练的过程，不过它的少量已知标签样本不在模型中使用，而是用于实验中生成标签函数 (Arachie 和 Huang, 2021)。

三、存在的问题

目前关于少量样本标签已知是负类的文献研究较少，但是这在实际生活中是一个常见的问题，比如检测样品中的异常，训练集包括大量的正常样本和少量的异常样本。SSLM 方法是解决这类问题的有效方法，但是它对超参数敏感，需要谨慎设置超参数，例如核函数类型、核函数参数、惩罚因子等。这些参数的选择可能会影响模型的性能，并且对不同数据集可能需要不同的调整。另外，SSLM 是一个浅层模型，对于高维数据集，它的计算复杂度可能会很高，因为涉及到计算支持向量和决策边界。对于大型数据集，这可能导致训练时间很长，或者需要大量的计算资源。把 SSLM 方法拓展到深度学习框架可以缓解这些局限性。SVDD 没有考虑少量负类标签已知的情况，但是 Ruff et al. (2018)将浅层的 SVDD 拓展到了深度学习领域，提出深度支持向量描述方法（Deep Support Vector Data Description，Deep SVDD），为本文 SSLM 的拓展提供了思路。然而，Deep SVDD 方法依然没有考虑少量负类样本已知的情况，而且它的分类器参数是通过分位数回归估计的，得到一个近似值。

使用程序性弱监督来处理少量样本标签已知类别的情况，它的关键挑战是如何将不同来源的弱监督源有效地聚合起来。两阶段方法中标签模型和终端模型的训练是分开的，即终端模型的训练结果不会反馈给标签模型。这使得模型训练缺乏灵活性。一阶段模型 ALL 可以克服这一缺陷，它的标签模型和终端模型是联合训练的。但是 ALL 模型的性能依赖于分类器模型的参数化，线性模型适用于具有简单特征的数据集，而面对复杂的文本数据或图像数据时，要考虑使用更为复杂的模型。而且该模型需要寻找合适的期望误差边界。过紧的边界会过度约束优化，导致无法找到解决方案，而过松的边界会限制优化，使对手过于强大，找到的优化方案不能展现良好的性能。

第三节　本文研究内容

本文考虑了不完全监督的两种数据情况：可获得的少量样本标签为负类和少量样本标签已知但不止一个类别。

对于可获得的少量样本的标签为负的情况，我们常见的就是处理不平衡二分类数据，可以获得大量的正类样本以少量的负类样本。但是目前的大多数方法只关注正类数据，并旨在建立对正类数据的描述。然而，数量很少的负类样本也可以用来提高我们模型的性能。基于上述问题，本文提出了可解释性深度小球大间隔网络（Interpretable Deep Small Sphere and Large Margin Network，ID-SSLMN）。ID-SSLMN 使用正类样本和少量负类样本训练一个超球，它将正类样本包裹在球内，负类样本排除在球外。具体工作是先使用神经网络将数据点映射到高维空间，然后在高维空间构建超球。接下来，本文开发了新的算法可以同时求解超球参数神经网络参数，即让神经网络和异常检测器统一训练，得到的超球参数的解相对 Deep SVDD 用分位数估计得到的解会更加精确。具体工作是将异常检测器的参数嵌入到神经网络最后一层神经元参数中。因此，针对异常检测问题，本文的贡献是：

- 更充分利用数据信息。在一分类 Deep SVDD 的基础上，考虑处理含少量负类样本的数据，这使得模型能够更充分利用数据信息。原来的 Deep SVDD 可退化成 ID-SSLMN 的一种特殊情况。
- 分类决策边界得到进一步优化。受浅层 SSLM 模型的启发，进一步改进 Deep SVDD 的分类边界，提出的新模型 ID-SSLMN 实现了“大间隔”的思想，能更好地检测异常样本。
- 建立具有可解释性的深度异常检测网络框架。本文增强了新模型中神经网络结构的可解释性。新网络中的最后一层神经元参数可解释为分类球体的中心和半径。
- 构建更高效、准确的统一求解算法。将原来交替求解策略改进为统一反向传播求解算法。新算法统一、同时求解特征提取的网络参数和异常检测器的参数。这样能够得到参数的更精准的最优解。

处理少量样本标签已知但不止一个类别时，假设可以获得一些来自弱分类器的噪声标签。这些弱分类器可能是启发式规则，或是领域的专家知识，也可以

是少量的标记样本训练的一个简单分类器。弱分类器给出的标签是存在错误的。假设可获得多个弱分类器且它们之间的决策是独立的，那么对每个样本来说，可以获得多个相互独立的弱标签。利用这些易获得的弱标签训练模型，使模型能够给未标记的样本或标记错误的样本相对准确的标签，本文提出了 L2 对抗标签学习。这是基于 ALL 的思想提出的，即用对抗训练的方式同时训练标签模型和终端模型。具体来说在约束的标签空间内，最大化损失以获得质量最差的标签，然后利用最差的标签通过最小化损失训练一个性能最好的分类器。由于弱监督源缺乏先验知识，不能对所有样本进行标记，本文考虑了弱监督源可能会放弃标记一些样本的情况。由于约束范围对模型效果有直接的影响，本文添加松弛变量来自适应调节约束范围。此外，本文将 SVM 模型引入程序性弱监督，发挥 SVM 分类效果好、可利用核函数进行非线性映射等优点。因此，针对程序性弱监督问题，本文的贡献如下：

- 在对抗学习的基础上，使用 L2 损失作为模型的损失函数，考虑弱分类器放弃标记一些样本的情况，并且为约束添加松弛变量，以提高模型性能。
- 将支持向量机模型拓展至弱监督学习领域，提出基于支持向量机的 L2 对抗标签学习模型。

第四节　文章结构

第一章介绍本文的研究背景和意义，简述监督学习和弱监督学习目前的研究现状，主要介绍了监督学习中的支持向量机和神经网络以及弱监督学习中的不完全监督方法。概括本文的主要研究内容是研究不完全监督学习的两种情况，即少量数据标签为负以及少量数据标签已知但不止一类。

第二章主要介绍了与本文相关的一些工作。SSLM 是处理含有少量负类标记样本情况的经典方法，Deep SVDD 是把浅层的 SVDD 拓展至深度学习框架的范例，它为本文将 SSLM 拓展至深度学习领域提供了思路。另外，本文还介绍了可解释神经聚类方法，它将分类器参数嵌入到神经网络参数，本文借鉴这样的思想构建了高效、统一的模型求解算法。ALL 是使用弱标签通过对抗训练的方式统一训练分类器的弱监督框架。

第三章提出了可解释的深度小球大间隔网络。加入少量负类样本参与训练，更加细化分类边界。将分类器参数嵌入到神经网络中，相对 Deep SVDD 更加精

确求解模型参数，且使模型更具可解释性。最后在 9 个数据集上验证了模型的效果。

第四章提出了 L2 对抗标签模型。修改原始 ALL 模型的损失函数，为约束添加松弛变量，将支持向量机模型拓展至弱监督学习领域。最后在 6 个数据集上验证了模型效果。

第五章对本文进行了总结并对未来的工作提出展望。本文针对第一章介绍的两种数据情况分别提出了一个模型，并进行了数值实验。根据数值实验结果，分析了本文提出的模型存在的问题，也为未来的工作方向提供了思路。

第二章　相关工作

小球大间隔方法（SSLM）是一种异常检测方法，是加入少量异常样本参与训练的不完全监督方法。但这是一种浅层方法，在大数据时代背景下，它处理高维、大规模数据的计算成本很大。深度支持向量描述方法（Deep SVDD）是一种深度的异常检测方法，它虽然是一种无监督方法，但是为我们将 SSLM 方法拓展至深度学习领域提供了思路。虽然 Deep SVDD 解决了浅层模型无法处理高维数据的问题，但是它无法精确求解分类器参数，而且神经网络的参数不具有可解释性。可解释神经聚类为我们将 SSLM 拓展至深度且精确求解分类器参数提供了可借鉴的思路。而且它使神经网络的参数具有可解释性。对抗标签学习框架（ALL）是处理少量样本标签已知的一种程序性弱监督框架，本文在这个框架下通过修改损失函数提出了效果更好的 L2 对抗标签模型。

第一节　小球大间隔方法

SSLM 结合了一类异常检测和二元分类算法的思想。它旨在构造一个紧凑的球，将正常样本包裹在球内 (Wu 和 Ye, 2009)。少量的异常样本被用来细化边界，使异常样本与球表面之间的间隔尽可能大。因此，SSLM 的目标是最小化球半径 R，同时最大化异常样本与球之间的间隔。具体来说, 给定一个数据集 $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$，其中有 m 个正常样本 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ 和 $n-m$ 个异常样本 $\{\boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_n\}$，$n-m$ 是远远小于 m 的。SSLM 的思想可以总结成如下式子：

$$
\begin{aligned}
\min_{R,\boldsymbol{c},\rho,\boldsymbol{\xi}} \quad & R^2 - v\rho^2 + \frac{1}{v_1 m}\sum_{i=1}^{m}\xi_i + \frac{1}{v_2(n-m)}\sum_{j=m+1}^{n}\xi_j \\
\text{s.t.} \quad & \| \phi(\boldsymbol{x}_i) - \boldsymbol{c} \|^2 \leq R^2 + \xi_i, \ \ 1 \leq i \leq m \\
& \| \phi(\boldsymbol{x}_j) - \boldsymbol{c} \|^2 \geq R^2 + \rho^2 - \xi_j, \ \ m < j \leq n \\
& \xi_k \geq 0, \ \ 1 \leq k \leq n,
\end{aligned}
\tag{2.1}
$$

其中，参数 $\boldsymbol{c}$ 和 R 分别表示球心和半径。$\rho^2 > 0$ 是球体表面与异常样本之间的间隔。$\boldsymbol{\xi} = [\xi_1, \cdots, \xi_n]^\top$ 是松弛变量向量。超参数 v, v_1, v_2 是三个取值在 $(0, 1]$ 的常数。上式所用的范数为 $L2$ 范数，即刻画 $\phi(\boldsymbol{x}_i)$ 到球心 $\boldsymbol{c}$ 的欧氏距离。对于一个训练样本 $\boldsymbol{x}_i (1 \leq i \leq n)$，如果它相应的松弛变量 $\xi_i > 0$，则把该样本称为边界误差。于是三个超参数 v, v_1, v_2 有这样的性质：$(v+1)v_1$ （或 vv_2）既是边界误

差个数与正常样本（或异常样本）的比值的上界，也是支持向量个数与正常样本（或异常样本）个数的比值的下界。

SSLM 在训练模型时加入少量异常样本，可以通过最大化正常样本与异常样本之间的间隔更加细化分类边界，提高模型性能。由于其对偶问题与 v-SVM 类似，因此可以使用 v-SVM 求解器 (Chang 和 Lin, 2001) 快速求解。它可以通过核方法对数据进行映射。然而，对于大规模和高维数据，求解二次规划问题的计算开销很大。大规模核函数矩阵难以存储。

第二节　深度支持向量数据描述

假设有输入空间 $\mathcal{X} \subseteq \mathbb{R}^d$ 和输出空间 $\mathcal{F} \subseteq \mathbb{R}^p$。用 $\phi(\cdot;\mathcal{W}) : \mathcal{X} \to \mathcal{F}$ 表示 $L \in \mathbb{N}$ 层神经网络从空间 $\mathcal{X}$ 映射到空间 $\mathcal{F}$，其中 $\mathcal{W} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L\}$ 且 $\boldsymbol{W}^l$ 表示第 l 层网络权重。Deep SVDD 的模型框架如图2.1所示，图的左侧部分是输入始空间中正常样本的散点图，中间部分表示神经网络映射，右图是映射后的空间散点图。输入空间中只有正常数据，通过网络 $\phi(\cdot;\mathcal{W})$ 映射，输出空间 $\mathcal{F}$ 中的数据被包裹在一个超球体中 (Ruff et al., 2018)。这样，正常数据就落在超球内，而异常数据则尽可能远离超球。

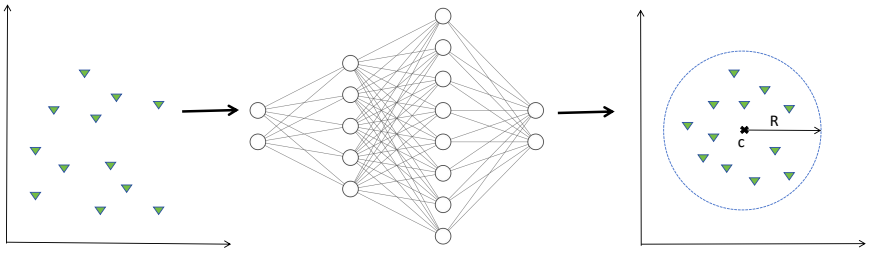


图 2.1 Deep SVDD 在训练集上的建模框架。

给定数据集 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$，可以通过交替训练网络参数 $\mathcal{W}$ 和最小化超球体积来学习最佳模型。软边界 Deep SVDD 目标函数定义如下:

$$\begin{aligned} &\min_{R,\boldsymbol{c},\mathcal{W}} R^2 + \frac{1}{vn}\sum_{i=1}^{n} \max\left\{0, \|\phi\left(\boldsymbol{x}_i;\mathcal{W}\right) - \boldsymbol{c}\|^2 - R^2\right\} \\ &+ \frac{\lambda}{2}\sum_{\ell=1}^{L} \left\|\boldsymbol{W}^{\ell}\right\|_F^2, \end{aligned} \tag{2.2}$$

其中 $\boldsymbol{c}$ 和 R 分别是超球的中心和半径。超参数 $v \in (0, 1]$ 控制模型中异常样本的比例，这与 SVDD 中的 v 属性类似。最后一项是网络参数 $\mathcal{W}$ 的权重正则化，即

使用 F 范数控制模型的复杂度，超参数 $\lambda > 0$。

Deep SVDD 将传统的 SVDD 模型扩展到深度学习框架，具有出色的数据表达能力和良好的分类结果。但是，如果限制偏置项和激活函数，神经网络的参数就会优化为 0，从而导致超球坍塌。而且一些异常点没有得到充分利用。此外，该方法使用分位数方法来求解参数 R 和 $\boldsymbol{c}$，无法获得参数的精确解。

第三节　可解释神经聚类方法

最近，高维数据的聚类也是一项备受关注的任务。深度聚类方法一方面利用深度学习处理高维数据，另一方面利用聚类方法使神经网络具有可解释性。可解释神经聚类（inTerpretable nEuraL cLustering，TELL）的主要思想是将传统的 K 均值聚类重塑为可微分的 K 均值聚类，并在此基础上建立神经层 (Peng et al., 2022)。对于给定的数据集 $\mathcal{D}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$，TELL 的目标函数定义如下：

$$\min \sum_{i=1}^{n} \sum_{j=1}^{k} \mathcal{I}_j(\boldsymbol{x}_i) \left\| \boldsymbol{x}_i - \boldsymbol{\Omega}_j \right\|_2^2, \tag{2.3}$$

其中 $\boldsymbol{\Omega}_j$ 表示第 j 个聚类中心，$\mathcal{I}_j(\boldsymbol{x}_i)$ 是一个示性函数，表示样本 $\boldsymbol{x}_i$ 是否属于第 j 个聚类中心。为将分类变量转为连续变量，可以使用带有温度系数的 softmax 函数对 $\mathcal{I}_j(\boldsymbol{x}_i)$ 进行放松，此时 $\mathcal{I}_j$ 表示样本 $\boldsymbol{x}_i$ 与第 j 个聚类中心的接近程度。

将式 (2.3) 的右边扩展为：

$$\left\| \boldsymbol{x}_i - \boldsymbol{\Omega}_j \right\|_2^2 = \| \boldsymbol{x}_i \|_2^2 - 2\boldsymbol{\Omega}_j^{\top} \boldsymbol{x}_i + \left\| \boldsymbol{\Omega}_j \right\|_2^2. \tag{2.4}$$

然后，定义

$$\boldsymbol{W}_j = 2\boldsymbol{\Omega}_j,\ b_j = -\left\| \boldsymbol{\Omega}_j \right\|_2^2,\ \| \boldsymbol{x}_i \|_2^2 = \beta_i \geq 0, \tag{2.5}$$

其中 $\boldsymbol{W}_j$ 是神经网络参数 $\mathcal{W}$ 的第 j 列，b_j 是网络偏置项 $\boldsymbol{b}$ 的第 j 个分量。β_i 是与数据点 $\boldsymbol{x}_i$ 的长度相对应的一个非负常数。

因此，给定一个样本 $\boldsymbol{x}$，TELL 的目标函数可以重写为

$$\mathcal{L} = \sum_j \mathcal{L}_j = \sum_j \mathcal{I}_j(\boldsymbol{x})(-\boldsymbol{W}_j^{\top} \boldsymbol{x} - b_j + \beta). \tag{2.6}$$

TELL 巧妙地将 K 均值的参数作为权重矩阵的一部分嵌入到神经网络中，并使用反向传播算法来训练神经网络。在这个过程中，神经网络会同时更新其网络参数和嵌入的 K 均值参数。这种方法的好处是可以让神经网络从 K 均值聚类中获取信息，从而增强模型的可解释性。不过，这种方法是一种无监督方法。在本文中，我们利用这一思想提出了一种有监督的异常检测模型，即把超球的中心和半径嵌入到神经网络中，通过统一的反向传播精确求解模型参数。

第四节　对抗标签学习框架

对抗标签学习 (Adversarial Label Learning, ALL) (Arachie 和 Huang, 2021) 是一个二分类弱监督框架，它的目标是在给出一些弱监督信号的情况下为数据返回准确的训练标签。这些标签的估计应该考虑到弱监督信号之间的相关性，为实现这一目标，首先使用弱信号的期望误差来定义数据的可能标签的约束空间，从这个空间采样的任何向量都可以用作训练标签。具体来说，考虑可以访问一组训练样本 $\mathcal{D}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$，这些样本的真实标签 $\{y_1, \cdots, y_n\} \in \{0,1\}^n$ 是无法获得的。但是可以通过启发式规则、领域专家知识等获得一些弱分类器。用这些分类器给未标记样本打上标签，于是便可以得到一组各种来源的弱监督信号 $\{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_s\}$，其中每个弱信号是数据的软标签，即 $\boldsymbol{q}_i \in [0,1]^n$，这些软标签表示的是样本属于正类的概率估计。在获得弱监督信号的同时，还可以接收到弱信号的期望错误率边界 $\boldsymbol{b} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_s]$。在实践中，弱信号的错误率边界被估计或视为超参数。根据期望误差的定义有：

$$b_i \geq \mathbb{E}_{\tilde{\boldsymbol{y}} \sim \boldsymbol{q}_i}\Big[\frac{1}{n}\sum_{j=1}^{n}[\tilde{y}_j \neq y_j]\Big], \tag{2.7}$$

它可以等价地表示为

$$b_i \geq \frac{1}{n}\big(\boldsymbol{q}_i^{\top}(1-\boldsymbol{y}) + (1-\boldsymbol{q}_i)^{\top}\boldsymbol{y}\big). \tag{2.8}$$

如果学习标签概率的当前估计为 $\boldsymbol{p} \in [0,1]^n$，ALL 训练一个分类模型 $f_{\boldsymbol{\theta}}$，它读取数据作为输入并输出标签概率，即有 $[f_{\boldsymbol{\theta}}(x_j)]_{j=1}^n = \boldsymbol{p}$，用 $\boldsymbol{p}(\boldsymbol{\theta})$ 表示 $[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)]_{j=1}^n$。我们将离散标签空间放宽到独立概率标签空间，使得 $\hat{y}_j \in [0,1]^n$ 表示样本 $\boldsymbol{x}_j$ 的真实标签 y_j 为正的概率。用类概率向量 $\hat{\boldsymbol{y}}$ 表示对抗标签。ALL 的主要思想是训

练一个模型，使其在最坏的条件下表现良好，把最坏条件下的标签称为对抗标签，所以学习的目标是最小化相对对抗标签的期望误差，优化以下问题：

$$\begin{aligned} \min_{\boldsymbol{\theta}} \max_{\hat{\boldsymbol{y}}\in[0,1]^n} \quad & \frac{1}{n}\left(\boldsymbol{p}(\boldsymbol{\theta})^\top(1-\hat{\boldsymbol{y}})+(1-\boldsymbol{p}(\boldsymbol{\theta}))^\top\hat{\boldsymbol{y}}\right) \\ \text{s.t.} \quad & b_i \geq \frac{1}{n}\left(\boldsymbol{q}_i^\top(1-\hat{\boldsymbol{y}})+(1-\boldsymbol{q}_i)^\top\hat{\boldsymbol{y}}\right),\ \forall i \in \{1,\cdots,s\}, \end{aligned} \tag{2.9}$$

其中 $\hat{\boldsymbol{y}}$ 是对抗标签，它表示约束范围内质量最差的标签。$\boldsymbol{p}(\boldsymbol{\theta})$ 是分类器模型，在实验时使用逻辑回归模型来作为终端模型。b_i 是第 i 个弱信号的错误率边界。共有 s 个弱信号。

第五节　本章小节

本章将与本文研究内容相关性较大的工作进行了简要的介绍。包括带有少量负类样本参与训练的不完全监督方法 SSLM。针对 SSLM 不适合处理大规模数据，本文旨在将 SSLM 拓展至深度学习框架。借鉴 Deep SVDD 将浅层 SVDD 模型拓展至深度的思想，ID-SSLMN 将输入特征映射到高维空间，再学习一个分类超球。而 Deep SVDD 的一个缺陷是无法精确求解超球参数。于是引入了 TELL 的思想，将分类器参数嵌入到神经网络中，统一求解分类器和神经网络的参数。此外，程序性弱监督方法可以解决不完全监督的另一种情况，即少量样本标签已知但不止一个类别。ALL 是一个重要的一阶段程序性弱监督框架。但是它的性能取决于合适的期望误差边界 b 以及分类器的参数化。本文基于 ALL 对抗学习的思想，从改进约束边界以及调整模型损失函数出发，以期望提高模型的性能。

第三章　可解释的深度小球大间隔网络

目前大多数异常检测方法只使用正常的训练数据。然而，在现实世界的许多情况下，存在一些可用的异常样本。虽然这些样本数量很少，但却蕴含着丰富的有益分类的信息。此外，传统的 Deep SVDD 模型还有一个缺陷，那就是可能出现一种被称为“球体坍塌”的现象，即学习网络会将所有样本映射到同一个点上。因此，本文考虑在训练过程中添加少量可获取的异常样本，以降低球体坍塌的风险。此外，基于 SSLM 的思想，我们还在模型中添加了参数 ρ 以细化分类边界，使模型在面对现实中的噪声时更加鲁棒。最后，受 TELL 的启发，我们提出了一种可以更精确求解超球半径和中心的算法，并使神经网络参数具有可解释性，提出了 ID-SSLMN 模型。

第一节　模型的提出

在各种浅层方法中，SSLM 表现出了卓越的性能。然而，它处理大规模数据的能力有限。我们将 SSLM 扩展到深度学习框架，以利用它强大的数据表达能力和并行计算能力，从而提高模型的性能。我们将利用神经网络把正常样本和异常样本映射到一个高维空间。随后，我们在这个高维空间中构建一个超球体，它可以将正常数据包裹在球内，而将异常数据排除在球外。ID-SSLMN 模型框架如图所示3.1。绿色点代表正常样本，红色点代表少量异常样本。左侧部分是训练数据的在原始空间上的散点图，中间部分表示神经网络映射，右侧部分是被神经网络映射后的散点图。ID-SSLMN 学习了一个以 $\boldsymbol{c}$ 为中心、以 R 为半径的超球体，这个超球体包围了其中的正常样本，并排除了其外的异常样本。ρ 代表异常样本与超球表面之间的间隔。

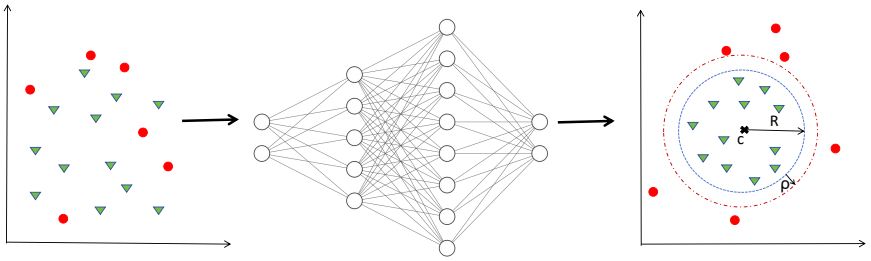


图 3.1　可解释的深度小球大间隔模型在训练集上的模型框架示意图。绿色三角形表示正常样本，红色圆点表示异常样本。

Chang et al. (2013)发现支持向量描述（Support Vector ata Description，SVDD）的原问题是一个关于 R 的非凸优化问题，其强对偶性无法保证。因此，理论上不可能保证求解 SVDD 的对偶问题就能准确地得到原始问题的解。他建议用 $\overline{R}$ 代替原始问题中的 R^2，其中 $\overline{R}$ 表示超球半径的平方。这保证了 SVDD 的目标是关于 R 的凸优化问题，本文借鉴了这一思想。

ID-SSLMN 的目标是最小化超球面的体积，同时最大化超球面与异常样本之间的距离。假设有 n 个数据点，其中 m 个正常样本 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ 和 $n-m$ 个异常样本 $\{\boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_n\}$。用 $\phi(\cdot;\mathcal{W}) : \mathcal{X} \to \mathcal{F}$ 表示数据点从输入空间 $\mathcal{X}$ 经过神经网络 $\phi()$ 到输出空间 $\mathcal{F}$ 的映射。那么，上述想法可以表述为以下优化问题：

$$\begin{aligned} \min_{\overline{R},c,\rho,\mathcal{W}} \quad & \overline{R} - v\rho + \frac{1}{v_1 m}\sum_{i=1}^{m}\max\left\{0, \| \phi(\boldsymbol{x}_i;\mathcal{W}) - c \|^2 - \overline{R}\right\} \\ & + \frac{1}{v_2(n-m)}\sum_{i=m+1}^{n}\max\left\{0, \overline{R} + \rho - \|\phi(\boldsymbol{x}_i;\mathcal{W}) - \boldsymbol{c}\|^2\right\} \\ & + \frac{\lambda}{2}\sum_{\ell=1}^{L}\| \boldsymbol{W}^{\ell} \|_F^2 \\ s.t. \quad & \overline{R} > 0,\ \rho > 0. \end{aligned} \tag{3.1}$$

超参数 $v, v_1, v_2 \in (0, 1]$ 控制着超球体积和违反边界之间的权衡。$\rho > 0$ 是超球表面与异常点之间的距离。$\overline{R}$ 是超球半径的平方。$\boldsymbol{c}$ 是超球的中心。$\mathcal{W} = [\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L]$ 表示神经网络权重，L 是网络层数。在目标函数公式 (3.1) 中，第一项表示超球体积最小化。第二项表示最大化超球面与异常样本之间的距离。第三项是神经网络映射后对位于球外的正常样本点的惩罚项。第四项是神经网络映射后对位于球内的异常样本点的惩罚项。最后一项是网络权重衰减正则化，超参数 $\lambda > 0$。

模型的异常得分定义如下：

$$s(\boldsymbol{x}) = \| \phi(\boldsymbol{x};\mathcal{W}^*) - \boldsymbol{c}^* \|^2 - \overline{R}^*. \tag{3.2}$$

其中，W^*、$\boldsymbol{c}^*$ 和 $\overline{R}^*$ 为模型的最优解。如果测试样本的异常得分大于 0，则表示样本位于超球之外，即样本异常。否则，就是正常数据。

Deep SVDD 用神经网络映射后的样本点的均值来作为超球中心，利用分位

数估计来求解超球半径，是一种求解参数近似值的方法，因此无法获得其精确值。此外，这是一种两阶段方法，神经网络参数和超球的参数交替更新。TELL将基于距离的模型参数嵌入神经网络中，受此启发，我们将超球的参数嵌入到神经网络中，这样就能够更精确求解超球的半径和中心。它通过反向传播统一优化所有参数来实现这一目标。其主要思路是在神经网络的原始输出层之后增加一层，专门用于求解球体的半径和中心。我们将 $s(\boldsymbol{x})$ 展开为

$$\begin{aligned} s(\boldsymbol{x}) =& \| \phi(\boldsymbol{x};\mathcal{W}) - \boldsymbol{c} \|^2 - \overline{R} \\ =& \boldsymbol{c}^\top \boldsymbol{c} - 2\boldsymbol{c}^\top \phi(\boldsymbol{x};\mathcal{W}) - \overline{R} + \| \phi(\boldsymbol{x};\mathcal{W}) \|^2 . \end{aligned} \tag{3.3}$$

我们定义 $\boldsymbol{c}^\top \boldsymbol{c} - 2\boldsymbol{c}^\top \phi(\boldsymbol{x};\mathcal{W}) - \overline{R}$ 为最后一层神经元，即神经网络的新输出。则有

$$g(\boldsymbol{x};\boldsymbol{w},b,\mathcal{W}) = \boldsymbol{w}^\top \phi(\boldsymbol{x};\mathcal{W}) + b. \tag{3.4}$$

其中，$g()$ 是神经网络新的输出，式 (3.4) 里的参数定义如下：

$$\boldsymbol{w} = -2\boldsymbol{c}, \;\; b = \boldsymbol{c}^\top \boldsymbol{c} - \overline{R}. \tag{3.5}$$

神经网络的输出是一个标量。$\boldsymbol{w}$ 是最后一层的权重向量，b 是偏置项。通过这样的定义，我们在网络参数和超球参数之间建立了联系，从而达到将超球参数嵌入神经网络的目的。

根据公式 (3.5) 中的定义，$\boldsymbol{w}$ 和 b 是内在耦合的，即 $b = \frac{1}{4}\boldsymbol{w}^\top \boldsymbol{w} - \overline{R}$。然而，对于神经网络来说，它们在训练过程中应该是解耦的。因此，有必要对最后一层的权重和偏置进行标准化处理，以防止在优化过程中出现梯度爆炸，导致损失函数无法收敛。具体做法是让超球中心的长度等于 1，即 $\boldsymbol{c}^\top \boldsymbol{c} = 1$。然后有

$$\boldsymbol{c} = -\frac{\boldsymbol{w}}{2}, \;\; \overline{R} = 1 - b. \tag{3.6}$$

所以我们可以将式 (3.1) 重写成

$$\begin{aligned}\min_{\rho,\mathcal{W},\boldsymbol{w},b}\quad & 1-b-v\rho+\frac{1}{v_1 m}\sum_{i=1}^{m}\max\left\{0,\parallel\phi(\boldsymbol{x}_i;\mathcal{W})\parallel^2+g(\boldsymbol{x}_i)\right\}\\ & +\frac{1}{v_2(n-m)}\sum_{i=m+1}^{n}\max\left\{0,\rho-\parallel\phi(\boldsymbol{x}_i;\mathcal{W})\parallel^2-g(\boldsymbol{x}_i)\right\}\\ & +\frac{\lambda}{2}\sum_{l=1}^{L}\parallel\boldsymbol{W}^l\parallel_F^2\\ s.t.\quad & \boldsymbol{w}^\top\boldsymbol{w}=4,\ b<1,\ \rho>0,\end{aligned}\tag{3.7}$$

其中 $\boldsymbol{w}$ 和 b 分别是神经网络最后一层的权重和偏置。$\phi(\boldsymbol{x};\mathcal{W})$ 是神经网络倒数第二层的输出。而 $g(\boldsymbol{x})$ 是神经网络的输出。在 ID-SSLMN 中，目标是通过最大化 b（其中 $b<1$）来使超球体积最小化。最大化 ρ，以最大化异常点与超球表面之间的距离。第三项和第四项分别是对正常样本被排除在超球之外和异常样本被包裹在超球之内的惩罚。最后一项是神经网络参数的正则化。约束条件是基于 $\boldsymbol{c}^\top\boldsymbol{c}=1$ 这一假设，超球半径的平方 $\overline{R}$ 和边界 ρ 都应大于 0。

利用拉格朗日乘法将有约束问题 (3.7) 转化为无约束问题，并得到问题 (3.7) 的拉格朗日函数如下：

$$\begin{aligned}&\mathcal{L}_1(\phi(\boldsymbol{x};\mathcal{W}),w,b)\\ &=1-b-v\rho+\frac{1}{v_1 m}\sum_{i=1}^{m}\max\left\{0,\parallel\phi(\boldsymbol{x}_i;\mathcal{W})\parallel^2+g(\boldsymbol{x}_i)\right\}\\ &+\frac{1}{v_2(n-m)}\sum_{i=m+1}^{n}\max\left\{0,\rho-\parallel\phi(\boldsymbol{x}_i;\mathcal{W})\parallel^2-g(\boldsymbol{x}_i)\right\}\\ &+\frac{\lambda}{2}\sum_{l=1}^{L}\parallel\boldsymbol{W}^l\parallel_F^2+\alpha(\boldsymbol{w}^\top\boldsymbol{w}-4)+\beta(b-1)-\gamma\rho,\end{aligned}\tag{3.8}$$

其中，$\alpha\neq 0$、$\beta\geq 0$、$\gamma\geq 0$ 是拉格朗日乘子。我们通过反向传播算法最小化拉格朗日函数 $\mathcal{L}_1$，从而共同确定参数 $\mathcal{W}$、$\boldsymbol{w}$、b 和 ρ。优化策略是 Adam 算法。拉格朗日乘子的更新策略是梯度下降，其梯度为：

$$\frac{\partial\mathcal{L}_1}{\partial\alpha}=\boldsymbol{w}^\top\boldsymbol{w}-4,\tag{3.9}$$

$$\frac{\partial\mathcal{L}_1}{\partial\beta}=b-1,\tag{3.10}$$

$$\frac{\partial \mathcal{L}_1}{\partial \gamma} = -\rho. \tag{3.11}$$

对于给定的新测试点 $\boldsymbol{x}$，我们计算神经网络映射后倒数第二层的输出值到超球中心的距离，从而得出异常得分。

$$s(\boldsymbol{x}) = \left\| \phi(\boldsymbol{x}; \mathcal{W}^*) + \boldsymbol{w}^{*\top} \boldsymbol{w}^* \right\|^2 + b^* - 1, \tag{3.12}$$

其中 $\mathcal{W}^*$、$\boldsymbol{w}^*$ 和 b^* 是通过最小化式 (3.8) 得到的最优参数。如果计算出的 $s(\boldsymbol{x})$ 值大于 0，则样本位于球外，被视为异常数据点。否则，则视为正常数据点。

算法 1: ID-SSLMN 的训练过程

输入: 正常样本训练集 $D_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$，异常样本训练集 $D_2 = \{\boldsymbol{x}_{m+1}, x_{m+2}, \ldots, \boldsymbol{x}_n\}, n >> m$, 验证集 $D_3 = \{\boldsymbol{x}_k 1, \boldsymbol{x}_k 2, \ldots, \boldsymbol{x}_{ks}\}$

输出: 神经网络参数 $\boldsymbol{\theta}$，$\boldsymbol{\theta} = \{\mathcal{W}, \boldsymbol{w}, b, \rho\}$

参数: $\lambda \geq 0, v, v_1, v_2 \geq 0$，学习率 r，最大 epochs 数 N，训练数据的批量数 (batches)M

随机初始化网络参数 $\boldsymbol{\theta}_0$，拉格朗日乘子 α, β 和 γ;
for *epoch = 1 to N* **do**
  for *batch = 1 to M* **do**
    /* 前向传播 */
    计算训练损失：$\mathcal{L}_1(\boldsymbol{\theta})$;
    /* 反向传播 */
    计算梯度值：$\nabla_{\boldsymbol{\theta}} \mathcal{L}_1(\boldsymbol{\theta})$;
    用 Adam 算法更新参数 $\boldsymbol{\theta}$，学习率为 r;
    从参数 $\boldsymbol{\theta}$ 得到 $\mathcal{W}$、$\boldsymbol{w}$，b 和 ρ;
  end
  if *epoch mod 5 = 0* **then**
    $\alpha_{t+1} = \alpha_t + r(\boldsymbol{w}^\top \boldsymbol{w} - 4)$;
    $\beta_{t+1} = [\beta_t + r(b-1)]_+$;
    $\gamma_{t+1} = [\gamma_t - r\rho]_+$;
  end
  if D_3 上的 $\mathcal{L}_2$ 值连续 *5* 个 *epoch* 都下降 **then**
    提前停止迭代;
  end
end

算法 1 总结了 ID-SSLMN 的训练过程。我们使用参数 $\boldsymbol{\theta}$ 来表示所有神经网络参数，即 $\boldsymbol{\theta} = \{\mathcal{W}, \boldsymbol{w}, b, \rho\}$ 。值得注意的是，在实际应用中，我们还将参数 ρ 设置为可学习的神经网络参数，与神经网络一起初始化。$\overline{R}$ 和 $\boldsymbol{c}$ 通过神经网络的反向传播统一更新，可以得到精确解。此外，拉格朗日乘数每 5 个 epoch 更新

一次（神经网络中，一个 epoch 表示所有训练样本一次前向传播和反向传播的过程）。

在以往基于超球体的深度异常检测方法中，如 Deep SVDD (Ruff et al., 2018)，超球体的半径和中心很难精确计算。我们的算法分解了异常得分公式，并将其中一部分定义为神经网络层，添加到神经网络映射的输出中。这样，我们就能从最后一层的输出中提取网络参数，从而准确确定球体的中心和半径。我们提出的算法可以显著提高模型的性能，尤其是在应用于大规模高维数据集时。

第二节　数值实验

一、实验设计

我们在 2 个合成数据集、3 个图像数据集和 4 个 UCI 数据集上进行了数值实验。数据摘要如下：

- 模拟数据集
  - Two Moon：Python 软件 *sklearn* 包中的 textitmake_moon 函数可以直接生成双月数据集。我们将随机种子设为 200，噪声设为 0.08，总共生成了 1000 个二分类样本。
  - Sprial Data：我们应用螺旋线的极坐标公式生成螺旋数据。随机种子设定为 200。噪声是 0 到 1 区间的随机数乘以 0.2。总共生成了 1000 个二分类数据点。
- 图像数据集
  - MNIST：这是一个用于手写数字识别的数据集，其中每个样本都是 28×28 像素的灰度图像，涵盖 0 到 9 的数字
  - Fashion MNIST：这是一个服装图像数据集，其中每个样本都是 28×28 像素的灰度图像，总共是 10 个不同类别的服装图像。
  - CIFAR10：这是一个彩色图像数据集，所代表的物体与现实世界中的物体非常接近。每个样本都是 32×32 像素的彩色图像，包括 6 类动物图像和 4 类交通工具图像。
- UCI 数据集
  - OBS Network：这是一项基于网络节点行为的分类任务。
  - Cardiotocography：该数据集的任务是根据胎儿心率（FHR）和子宫收

缩（UC）特征的测量值对胎儿形态模式进行分类。

– Breast Cancer：该数据集的任务是根据乳腺肿块细针抽吸术（FNA）数字图像计算出的特征诊断是否患有乳腺癌。
– Credit Card：该数据集的任务是根据一组属性描述来评估是否存在信贷风险。

对于图像数据集，采用卷积神经网络（CNN）进行映射，而对于 UCI 和合成数据集，则采用全连接神经网络（FC）进行映射。我们在表3.1中总结了数据集和所用网络层的详细信息。对于每个数据集，我们选择一个类作为正常类，将所有其他类视为异常类。在训练过程中，我们从异常数据中随机抽取正常样本量的 1/10 个样本，将其添加到正常数据中一起进行训练。测试数据包括所有的类。

表 3.1 数据集的相关信息表

数据集名称	类别数	样本量	样本维度	隐藏层数量
MNIST	10	60000 (train) + 10000 (test)	28*28	CNN×2+FC×1
Fashion MNIST	10	60000 (train) + 10000 (test)	28*28	CNN×2+FC×2
CIFAR10	10	50000 (train) + 10000 (test)	32*32	CNN×3+FC×1
OBS Network	4	1060	21	3
Cardiotocography	10	2126	21	3
Breast Cancer	2	569	30	3
Credit Card	2	1000	20	3
Two Moon	2	1000	2	3
Spiral Data	2	1000	2	3

使用受试者工作特征曲线（Receiver Operating Characteristic，ROC）下方的面积（Area Under the Curve，AUC）值来衡量不同基线的性能。我们考虑了浅层和深度异常检测基线。

- 孤立森林（Isolation Forest，IF）(Liu et al., 2008)：这是一种基于树的方法，它随机选择特征，然后随机选择这些特征的分割值，对数据集进行递归分割。与正常点相比，异常点需要的随机分区更少。因此，异常点是树中路径较短的点。IF 的实验将参数 n-estimators 设为 100，最大样本数设为 256，随机噪声设置为 0.1。
- 核密度估计（KDE）(Latecki et al., 2007)：这是一种估计密度函数的非参数统计方法。它通过比较一个点的局部密度和邻近点的局部密度来确定该点

是正常的还是异常的。KDE 使用的是高斯核，本实验中核带宽参数的调整是通过交叉验证完成的，选参范围是 $\{2^{0.5}, 2^{1}, \ldots, 2^{5}\}$。

- OCSVM (Schölkopf et al., 1999)：这是一种基于 SVM 的异常检测方法。它通过在训练样本和坐标系原点之间建立一个超平面来判断样本是否异常。超参数 ν 通过网格搜索在 $\{0.1, 0.2, \cdots, 0.9\}$ 中选择最优参数进行训练。γ 设置为默认值。
- SVDD (Tax 和 Duin, 2004)：这是另一种单类分类方法。它旨在学习一个包围正常样本并排除异常样本的超球。SVDD 的最佳参数 ν 是通过交叉验证在 $\{0.1, 0.2, \cdots, 0.9\}$ 中选出的。
- SSLM (Wu 和 Ye, 2009)：这种方法是在训练过程中加入少量异常样本，被认为是一种浅层方法。SSLM 的最佳参数 ν、ν_1 和 ν_2 是通过交叉验证在 $\{0.1, 0.2, \cdots, 0.9\}$ 中选出的。
- Autoencoder (AE) (An 和 Cho, 2015)：AE 用于异常检测的想法是，异常样本无法被很好地重建，使用重建后的均方误差作为异常得分。
- Deep SVDD (Ruff et al., 2018)：参数 ν 的选择范围为 $\{0.1, 0.2, \cdots, 0.9\}$。为防止球体坍塌，在进行初始前向传递后，将映射后数据的平均值作为超球体中心 $\boldsymbol{c}$，计算映射后的数据点到超球中心的距离，使用各距离值的分位数回归估计半径 R。

另外，我们增加一个基于 Deep SVDD 的对比方法，即加入少量异常样本训练 Deep SVDD，可以把该方法看作是消融实验的对比，它没有加入参数 ρ 细化分类边界，也没有使用更精确求解参数的算法，我们把该方法命名为基于负类样本的深度支持向量描述（Deep Support Vector Data Description with Negative，Deep SVDN）。三种深度对比模型采用了表3.1中相同的网络结构。图像数据集的参数选择使用了网格搜索，UCI 数据集的参数选择使用了 5 折交叉验证。

对于我们提出的方法 ID-SSLMN，实验设置如下：神经网络优化器选择 Adam 优化器。将初始学习率设置为 0.0001。使用 MultiStepLR 动态调整学习率。最大迭代次数为 200 次。UCI 和合成数据集的批次大小为 50，图像数据集的批次大小为 150。权重衰减超参数 λ 设置为 $\lambda = 5e-6$。超参数 ν、ν_1 和 ν_2 的选择范围为 $\{0.1, 0.2, \cdots, 0.9\}$。在图像数据集上使用网格搜索进行参数选择，而在 UCI 数据集上使用 5 折交叉验证进行参数选择。此外，为了防止神经网络过度拟合，我

们还实施了一种机制，如果验证集损失在 5 个 epoch 内持续减少，则停止训练。这里的验证集是从训练集中随机抽样 30% 的数据得到的。

二、模拟数据上的实验结果

我们的模型旨在训练一个网络，将正常点映射到球体内部，将异常点映射到球体外部。我们使用合成数据进行了实验，并通过可视化数据分类边界验证了我们的想法。

图3.2是 ID-SSLMN 方法在双月数据和螺旋数据上的可视化。子图（a）、（b）、（c）和（d）是双月数据的散点图，而子图（e）、（f）、（g）和（h）是螺旋数据的散点图。子图（a）和（b）分别是双月数据集训练集的原始数据散点图和通过神经网络映射后的散点图。子图（c）和（d）分别表示双月数据集测试集的原始散点图和通过神经网络映射后的散点图。值得注意的是，（b）和（d）中的子图描绘的是使用 Deep SSLM 训练的神经网络倒数第二层的映射结果。子图（b）和（d）中的黑色实线代表模型的决策边界，即分类超球。同样，子图 (e)、(f)、(g) 和 (h) 是螺旋数据训练集和测试集映射前后的散点图。可以看出，我们训练的神经网络能够将正常点映射到一个非常紧密的球体中，同时将异常点排除在球外。

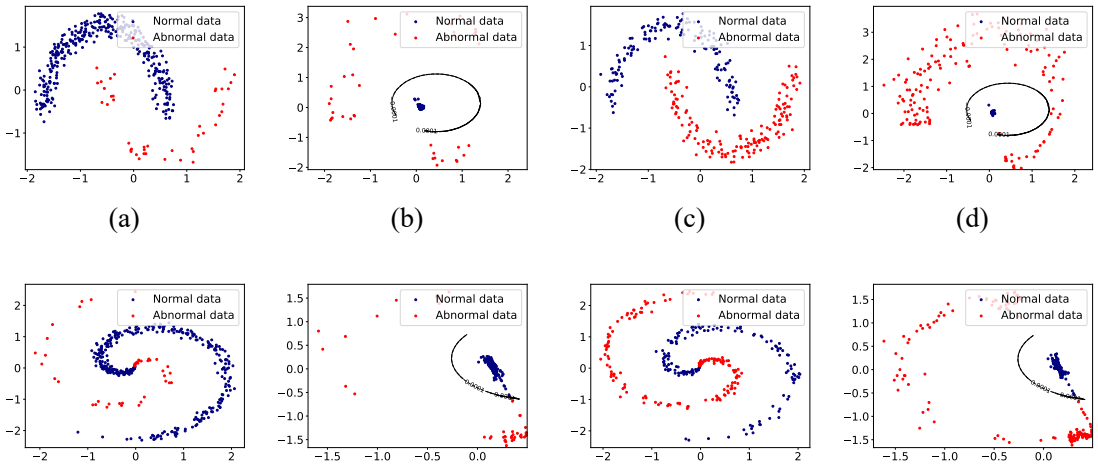
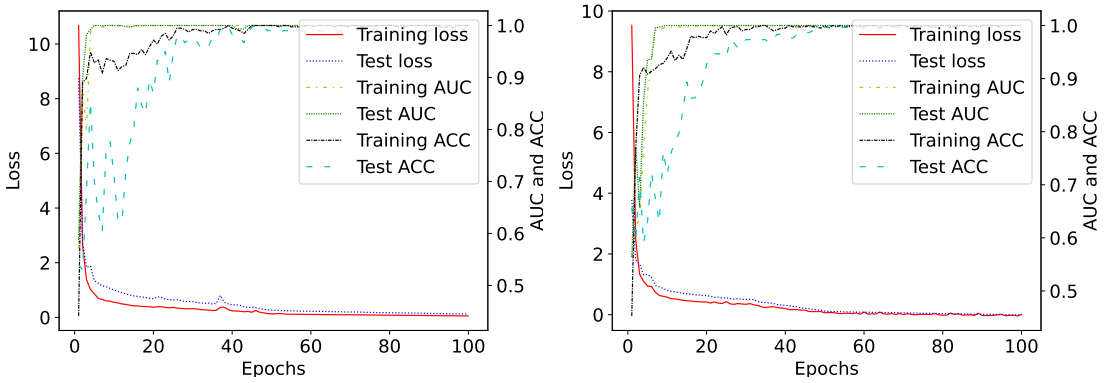


图 3.2 双月数据和螺旋数据上的 ID-SSLMN 的可视化散点图

图3.3显示了 ID-SSLMN 在合成数据上迭代时的各种性能指标，包括准确率（ACC）、AUC 值和损失值。可以看出，迭代 40 次后，模型达到稳定，模型损失显著减少。不过，模型总迭代次数超过了 40 次，因为我们的实验设置为当损失持续减少 5 次时停止迭代。



(a) Two Moon　　(b) Spiral Data

图 3.3 双月数据和螺旋数据上 ID-SSLMN 迭代过程中的各性能指标值

三、真实数据上的实验结果

该实验的评估标准是 AUC 值。由于 MNIST、Fashion MNIST 和 CIFAR10 数据集都由 10 类数据组成，因此每种方法都对所有 10 类数据进行了测试，最终结果如下。

从表3.2、表3.3和表3.4中的结果可以看出，在这 3 个数据集上，深度异常检测方法的性能都明显优于传统异常检测方法，说明深度学习在异常检测中具有更强的表现力。特别是在 MNIST 数据集上，ID-SSLMN 方法的平均 AUC 值比 Deep SVDD 方法提高了 4.92%。Deep SVND 方法的表现也具有竞争力，它在 7 个类别中的结果排名第 2，仅次于 ID-SSLMN 方法。此外，CIFAR10 的结果也证明了 ID-SSLMN 方法的强大优势。与传统的浅层方法和 Deep SVDD 方法相比，ID-SSLMN 方法的 AUC 值明显更高。相比之下，在 Fashion MNIST 数据集上的结果并没有显示出很强的优势。不过，ID-SSLMN 在 10 个类别中的 6 个类别上仍然表现出色。从这 3 个图像数据集上的实验可以看出，我们提出的 ID-SSLMN 方法在处理图像异常检测问题上具有显著优势。这可能是因为我们的方法可以精确计算超球的中心和半径，从而得到更精确的分类边界。

OBS 网络、心脏造影、乳腺癌和信用卡这四个数据集均来自 UCI 网站，这些数据集规模小、维度低。在这些数据集的实验中，我们采用了 5 折交叉验证来选择最佳参数。对于传统的浅层方法，我们计算了 5 折交叉验证中获得的 AUC 值的平均值和方差。至于深度模型，我们记录了使用最佳参数重复 5 次实验所获得的 AUC 值的平均值和方差。

表 3.2 每种方法在 MNIST 上的 AUC（%）结果

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
0	90.33	96.28	99.01	90.39	93.88	97.47	98.00	94.37	**99.90**
1	97.46	92.54	99.56	98.34	98.61	96.93	99.70	79.67	**99.94**
2	78.96	85.56	95.78	53.33	85.49	80.06	91.70	96.02	**99.82**
3	82.12	87.17	93.96	72.44	83.67	86.94	91.90	96.24	**99.62**
4	80.35	82.34	95.79	84.50	89.04	83.73	94.90	98.14	**99.83**
5	74.24	73.05	91.57	59.91	74.10	79.22	88.50	96.36	**99.73**
6	84.92	91.27	98.09	70.66	90.70	94.96	98.30	99.02	**99.35**
7	85.57	86.66	96.01	70.67	87.76	90.61	94.60	95.52	**99.34**
8	81.43	89.16	93.76	78.56	85.06	86.51	93.90	97.14	**99.56**
9	84.15	88.59	96.34	81.57	85.32	93.54	96.50	96.70	**99.02**

表 3.3 每种方法在 Fashion MNIST 上的 AUC（%）结果

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
T-shirt	91.25	91.57	91.87	82.81	86.71	88.70	87.46	92.76	**97.50**
Trouser	97.75	98.90	99.02	83.00	96.03	96.58	96.98	97.68	**99.66**
Pullover	87.33	88.82	**89.16**	86.09	84.99	85.81	84.86	88.92	88.57
Dress	93.55	93.94	**94.01**	80.66	92.05	88.55	90.16	93.99	87.61
Coat	89.92	90.08	**90.68**	85.94	86.60	84.13	87.07	88.49	84.64
Sandal	92.58	90.70	91.78	78.26	84.20	88.64	89.60	95.08	**96.76**
Shirt	79.56	82.40	82.08	80.26	77.07	80.65	79.51	**82.53**	80.24
Sneaker	98.27	98.60	98.68	96.00	98.17	96.34	97.52	96.84	**99.34**
Bag	87.28	88.58	89.87	64.67	67.64	88.71	92.04	93.92	**99.76**
Ankle boot	98.06	96.79	98.03	95.86	88.72	92.53	97.45	98.23	**99.65**

表 3.4 每种方法在 CIFAR10 上的 AUC（%）结果

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
Airplane	75.26	64.03	67.50	62.47	77.58	63.93	61.70	66.77	**90.16**
Automobile	58.87	61.83	62.23	42.22	61.42	60.58	65.90	74.74	**92.46**
Bird	60.17	49.68	51.61	64.98	68.56	51.19	50.80	63.39	**80.75**
Cat	55.10	58.58	56.52	49.34	63.31	54.66	59.10	67.70	**81.81**
Deer	63.15	63.27	59.48	74.13	76.51	55.35	60.90	60.55	**84.21**
Dog	65.10	64.73	59.98	50.04	65.15	58.25	65.70	73.32	**86.61**
Frog	69.42	70.53	63.66	69.93	77.90	62.57	67.70	71.18	**87.74**
Horse	64.72	63.65	64.82	52.16	64.56	63.42	67.30	72.43	**86.74**
Ship	79.86	77.39	80.16	66.44	81.92	79.07	75.90	79.29	**92.80**
Truck	72.58	74.87	74.28	49.93	71.94	70.73	73.10	72.70	**89.22**

表 3.5 OBS Network 上每种方法的 AUC 平均值（% ）

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
NB-No Block	79.89 (±3.21)	79.09 (±3.36)	71.82 (±2.36)	57.39 (±3.00)	81.81 (±2.24)	79.40 (±1.94)	51.26 (±7.49)	59.78 (±4.93)	**95.83 (±0.63)**
Block	88.90 (±4.55)	95.79 (±0.89)	93.49 (±1.46)	92.11 (±1.23)	99.43 (±0.33)	80.46 (±1.71)	76.12 (±14.06)	97.40 (±0.72)	**99.86 (±0.28)**
No Block	92.90 (±4.42)	97.86 (±0.59)	96.73 (±3.59)	95.65 (±1.83)	99.96 (±0.05)	86.74 (±4.46)	84.14 (±7.96)	98.45 (±0.56)	**100.00 (±0.00)**
NB-Wait	80.62 (±3.48)	84.90 (±3.64)	83.23 (±3.99)	55.40 (±3.59)	76.77 (±5.73)	83.75 (±2.13)	56.56 (±10.99)	61.01 (±6.46)	**96.18 (±0.59)**

表 3.6 Cardiotocography 上每种方法的 AUC 平均值（% ）

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
1	**88.08 (±1.53)**	85.90 (±1.62)	85.75 (±1.90)	84.12 (±2.91)	85.98 (±1.31)	86.87 (±1.15)	79.52 (±3.73)	86.35 (±3.96)	79.14 (±4.36)
2	84.13 (±1.46)	81.82 (±2.10)	87.64 (±2.59)	86.00 (±1.30)	90.32 (±1.93)	78.19 (±1.82)	74.76 (±3.33)	**98.98 (±0.20)**	88.32 (±3.28)
3	91.57 (±2.27)	90.83 (±3.49)	90.36 (±1.75)	90.47 (±2.79)	90.27 (±3.76)	82.39 (±6.73)	81.66 (±4.43)	85.70 (±1.51)	**96.42 (±0.71)**
4	88.02 (±4.44)	91.98 (±4.02)	93.26 (±2.05)	93.69 (±1.64)	95.61 (±1.07)	79.39 (±6.05)	88.10 (±4.41)	**96.02 (±0.82)**	95.78 (±2.66)
5	86.43 (±3.89)	86.79 (±1.87)	87.79 (±3.59)	88.03 (±1.87)	90.26 (±1.34)	86.15 (±2.66)	85.33 (±3.24)	89.23 (±0.52)	**91.69 (±1.85)**
6	83.83 (±3.70)	81.71 (±4.38)	89.13 (±1.97)	89.28 (±1.07)	90.44 (±3.05)	76.62 (±4.22)	68.57 (±1.56)	93.77 (±1.49)	**94.13 (±1.83)**
7	84.36 (±3.76)	84.37 (±2.98)	87.46 (±0.82)	87.52 (±3.04)	93.86 (±1.67)	80.72 (±3.14)	79.94 (±3.78)	97.32 (±0.70)	**98.21 (±0.43)**
8	98.06 (±0.96)	97.24 (±2.90)	95.58 (±2.51)	96.57 (±1.49)	**99.57 (±0.17)**	52.49 (±6.05)	96.09 (±1.96)	99.13 (±0.43)	99.05 (±0.48)
9	93.80 (±2.61)	94.93 (±1.76)	89.72 (±5.77)	90.52 (±3.96)	93.85 (±2.52)	74.61 (±4.21)	94.71 (±2.02)	97.58 (±0.83)	**98.06 (±0.23)**
10	93.17 (±2.43)	92.11 (±1.70)	91.57 (±1.71)	91.42 (±1.70)	92.45 (±2.02)	91.31 (±1.90)	89.90 (±2.09)	94.37 (±0.50)	**94.88 (±0.66)**

OBS 网络数据集由 4 个类别组成，但这是一个各类别样本量不平衡的数据集。其中，NB-No Block 和 NB-Wait 类的样本数量很少。表3.5中的实验结果表明，ID-SSLMN 在处理少量正常样本和多个异常类样本的数据时表现较好。

表3.6列出了九种方法在心脏造影数据集上的 AUC 值。该数据集有 10 个类别，ID-SSLMN 在其中 6 个类别中表现良好。Deep SVDN 方法在大多数类别中的表现并不理想，但与之前的传统方法相比，它的 AUC 值仍然有所提高。这表明，对于 ID-SSLMN 方法来说，在训练过程中加入少量异常样本可以提高模型性能。此外，ID-SSLMN 还能通过更精确求解参数进一步提高模型性能的稳定性。

乳腺癌数据集和信用卡数据集都包含两类数据，其中一类被视为正常类，另一类代表异常类。表3.7和表3.8显示了这两个数据集的实验结果。ID-SSLMN 仅在一个实验中表现良好。其次表现较好的是 SSLM。此外，Deep SVND 的性能较差，与 Deep SVDD 方法相比仅有轻微改进。这表明，我们提出的方法更适用于具有多个异常类别的异常检测。

表 3.7 Breast Cancer 上每种方法的 AUC 平均值（% ）

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
Benign	96.29 (±1.28)	91.91 (±2.43)	95.13 (±1.62)	94.42 (±1.81)	**98.47 (±0.81)**	89.86 (±1.62)	61.67 (±8.09)	83.75 (±8.07)	95.76 (± 0.91)
Malignant	87.68 (±2.68)	84.35 (±5.09)	87.39 (±4.90)	86.57 (±4.53)	98.36 (±1.04)	62.35 (±6.21)	67.34 (±10.38)	91.99 (±4.71)	**99.08 (± 0.08)**

表 3.8 Credit Card 上每种方法的 AUC 平均值（%）

Normal class	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
Bad	57.23 (±4.82)	55.21 (±4.47)	53.88 (±4.03)	53.85 (±5.01)	67.06 (±4.74)	55.76 (±4.13)	56.06 (±3.37)	63.97 (±1.80)	**73.06 (±0.54)**
Good	60.93 (±3.91)	60.66 (±1.48)	58.69 (±3.93)	58.94 (±2.12)	**74.56 (±3.27)**	59.32 (±4.90)	56.10 (±6.76)	63.35 (±2.44)	71.90 (±1.33)

四、弗里德曼检验

虽然我们提出的 ID-SSLMN 方法在图像数据集上表现出显著优势，但它在低维 UCI 数据集上的优势并不明显。为了使我们的实验分析更有说服力，我们对 9 种方法在每个数据集上的性能进行了统计分析。基于秩的弗里德曼检验是用于检验多种算法之间是否存在显著差异的常用工具。这是一种简单、非参数但安全的检验方法。根据上一子节中每种算法在每个数据集上的测试结果，我们对每种算法在每个数据集上的性能进行排序，并计算平均排序值。我们分两部分计算：图像数据和 UCI 数据。得到的排名结果汇总在表3.9和表3.10中。

假设对 N 数据集上的 k 算法进行比较。弗里德曼检验统计量定义为：

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]. \tag{3.13}$$

这里的 R_j 表示第 i 个算法的平均排名值，即 $R_j = \frac{1}{N}\sum_i r_i^j$，$r_i^j$ 表示第 j 个算法在第 i 个数据集中的性能排名值。当 k 和 N 都很大时，公式 (3.13) 近似于自由度为 $k-1$ 的卡方分布。然而，上述最初的弗里德曼检验过于保守，这意味着不容易拒绝原假设。现在常用的统计量如下：

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}, \tag{3.14}$$

其中 τ_F 遵循自由度为 $k-1$ 和 $(k-1)(N-1)$ 的 F 分布。

对于图像数据，根据公式（3.13）和公式（3.14），我们可以得到 $\tau_{\chi^2}^1 = 111.25$ 和 $\tau_F^1 = 25.06$。其中，τ_F^1 遵循自由度为 (8, 232) 的 F 分布。在显著性水平为 $\alpha = 0.1$ 时，F(8, 232) 的临界值为 1.70；在显著性水平为 $\alpha = 0.05$ 时，临界值为 1.98；在显著性水平为 $\alpha = 0.025$ 时，临界值为 2.25。可以看出，我们得到的统计量值远远大于临界值。因此，原假设被拒绝，9 种方法之间的差异被认为是显著的。除此之外，表3.9还显示，我们提出的 ID-SSLMN 方法的平均排名低于其他算法，这说明我们的方法是有效的。

表 3.9　9 种方法在图像数据集上的平均 AUC 排名

Datasets	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
MNIST									
0	9	5	2	8	7	4	3	6	1
1	6	8	3	5	4	7	2	9	1
2	8	5	3	9	6	7	4	2	1
3	8	5	3	9	7	6	4	2	1
4	9	8	3	6	5	7	4	2	1
5	6	8	3	9	7	5	4	2	1
6	8	6	4	9	7	5	3	2	1
7	8	7	2	9	6	5	4	3	1
8	8	5	4	9	7	6	3	2	1
9	8	6	4	9	7	5	3	2	1
Fashion MNIST									
T-shirt	5	4	3	9	8	6	7	2	1
Trouser	4	3	2	9	8	7	6	5	1
Pullover	5	3	1	6	8	7	9	2	4
Dress	4	3	1	9	5	7	6	2	8
Coat	3	2	1	7	6	9	5	4	8
Sandal	3	5	4	9	8	7	6	2	1
Shirt	7	2	3	5	9	4	8	1	6
Sneaker	4	3	2	9	5	8	6	7	1
Bag	7	6	4	9	8	5	3	2	1
Ankle boot	3	6	4	7	9	8	5	2	1
CIFAR10									
Airplane	3	6	4	8	2	7	9	5	1
Automobile	8	5	4	9	6	7	3	2	1
Bird	5	9	6	3	2	7	8	4	1
Cat	7	5	6	9	3	8	4	2	1
Deer	5	4	8	3	2	9	6	7	1
Dog	5	6	7	9	4	8	3	2	1
Frog	6	4	8	5	2	9	7	3	1
Horse	6	7	4	9	6	8	3	2	1
Ship	4	7	3	9	2	6	8	5	1
Truck	6	2	3	9	7	8	4	5	1
Average rank	5.90	5.17	3.63	7.8	5.77	6.73	5.00	3.27	1.73

表 3.10　9 种方法在 UCI 数据集上的平均 AUC 排名

Datasets	IF	KDE	OCSVM	SVDD	SSLM	AE	Deep SVDD	Deep SVDN	ID-SSLMN
OBS Network									
NB-No Block	3	5	6	8	2	4	9	7	1
Block	7	4	5	6	2	8	9	3	1
No Block	7	4	5	6	2	8	9	3	1
NB-Wait	5	2	4	9	6	3	8	7	1
Cardiotocography									
1	1	5	6	7	4	2	8	3	9
2	6	7	4	5	2	9	8	1	3
3	2	3	5	4	6	8	9	7	1
4	8	6	5	4	3	9	7	1	2
5	7	6	5	4	2	8	9	3	1
6	6	7	5	4	3	8	9	2	1
7	7	6	5	4	3	8	9	2	1
8	4	5	8	6	1	9	7	2	3
9	6	3	8	7	5	9	4	2	1
10	3	5	6	7	4	8	9	2	1
Breast Cancer									
Benign	2	6	4	5	1	7	9	8	3
Malignant	4	7	5	6	2	9	8	3	1
Credit Card									
Bad	4	7	8	9	2	6	5	3	1
Good	4	5	8	7	1	6	9	3	2
Average rank	4.78	5.17	5.67	6.00	2.83	7.17	8.06	3.44	1.89

对于 UCI 数据集，我们得到 $\tau_{\chi^2}^2 = 77.63$ 和 $\tau_F^2 = 19.88$ 。其中，τ_F^2 遵循自由度为 (8, 136) 的 F 分布。在显著性水平为 $\alpha = 0.1$ 时，F(8, 136) 的临界值为 1.72；在显著性水平为 $\alpha = 0.05$ 时，临界值为 2.01；在显著性水平为 $\alpha = 0.025$ 时，临界值为 2.29。可以知道，τ_F^2 远远大于临界值。因此，我们认为算法之间在 UCI 数据集上的差异也是显著的。此外，表3.10中的结果显示，ID-SSLMN 的平均排名最高，其次是 SSLM，而 Deep SVDN 的平均排名第三。这意味着深度 SSLM 在处理高维和低维数据方面都有很强的优势，说明我们在 ID-SSLMN 中引入的参数 ρ 是一种有效的改进。这也意味着我们设计的更精确求解参数的方法是必要且有效的。

第三节　本章小节

首先，本章将传统的不完全监督模型 SSLM 拓展至深度学习框架。根据深度异常检测模型 Deep SVDD 进行了改进。在训练集中添加了少量异常样本以减少球体塌陷的风险。用 $\overline{R}$ 代替 Deep SVDD 原始问题中的 R^2 以保证模型的原问题是一个关于 $\overline{R}$ 凸优化问题。另外，加入参数 ρ 以进一步细化决策边界，提高

异常检测效率。

其次，本章对改进模型提出了新的求解算法 ID-SSLMN。将异常检测器参数嵌入到神经网络最后一层神经元的参数中，这样可以通过反向传播统一求解异常检测器参数和神经网络参数，并且使得神经网络最后一层神经元参数具有实际意义，而且求得的异常检测器参数会相对 Deep SVDD 更加精确。

最后，提出的模型在两个合成数据集和 7 个真实数据集上进行数值实验，并且与目前经典且先进的异常检测方法进行了对比，通过弗里德曼检验验证了所提出方法的有效性。

第四章　基于 L2 损失的对抗标签学习

在本章中，我们考虑处理不完全监督中少量样本标签已知但不止一个类别的情形。基于弱监督学习框架 ALL，我们提出了 L2 对抗标签学习（L2 Adversial Label Learning，LALL）。具体来说，由于可获得的先验知识不可能覆盖所有的样本情况，所以考虑弱监督源可能会放弃标记一些样本的情况。其次，对抗标签学习的约束范围是由期望误差边界 b 来决定，参数的取值对模型效果的影响是非常大的，因此我们为约束条件添加松弛变量，让模型可以自适应地调节约束范围。此外，我们还修改了损失函数，用 L2 损失来优化分类器。最后，分别使用逻辑回归和支持向量机作为终端模型，通过对抗学习的方式来学习网络参数。

第一节　模型的提出

对抗标签模型的核心思想就是在约束的标签空间内，给出一个质量最差的标签（对抗标签），然后把对抗标签当作真实标签训练一个质量最好的监督模型，这两个过程是交替迭代的，图4.1展示了 LALL 的模型框架。对于这个模型来说，可获得的是无标记样本和一些弱监督源，这些弱监督源会给每个样本一个带有噪声的标签（弱信号），弱信号的形式是软标签，即样本属于某一类的概率。如果是二分类，那么每个弱监督源的输出是一个向量，每个分量表示样本属于正类的概率；如果是多分类，每个弱监督源的输出是一个矩阵。多个弱监督源的结果聚合在一起形成了样本的弱信号矩阵。根据获得的弱信号矩阵，可以构造一个可行标签约束空间（依据期望误差定义的）。分类器使用 SVM 或逻辑回归（Logistic Regression，LR）作为终端分类器。若给定一个 SVM 分类器，通过最大化损失函数在可行空间内找到一个对抗标签，然后再用对抗标签通过最小化损失函数更新 SVM 的参数，学习一个表现良好的分类器模型。接下来可以继续使用更新的 SVM 寻找下一个对抗标签，直至 SVM 的参数收敛获得最优分类器。同理也可使用 LR 分类器。

对于二分类任务，可以获得的弱信号是标记样本属于正类的概率，并且放弃估计所有其他类中的成员的资格。在某些情况下，如自然语言处理，弱监督源并不能完全捕捉所有的单词，这就导致弱监督源会放弃标记一些样本，这些样本的弱信号就是空值，即对每个样本，可以获得 m 个关于样本标签估计的弱信

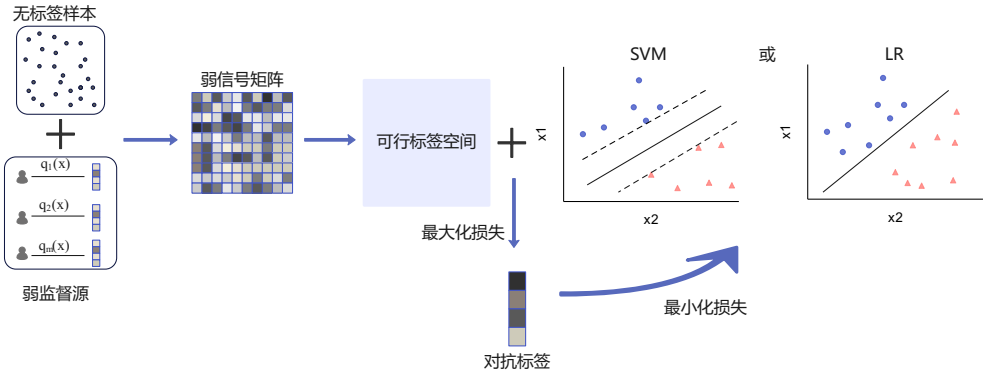


图 4.1 LALL 的模型框架图

号 $\{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_m\}$，其中每个弱信号为 $\boldsymbol{q}_i \in [\emptyset, 0, 1]^n$。因此，定义弱信号的期望误差为

$$\begin{aligned} b_i \geq & \frac{1}{n_i}\left(\mathbb{1}_{(q\neq\emptyset)}\boldsymbol{q}_i^\top(1-\boldsymbol{y}) + \mathbb{1}_{(q\neq\emptyset)}(1-\boldsymbol{q}_i)^\top\boldsymbol{y}\right) \\ \geq & \frac{1}{n_i}\left(\mathbb{1}_{(q\neq\emptyset)}(1-2\boldsymbol{q}_i)^\top\boldsymbol{y} + \boldsymbol{q}_i^\top\mathbb{1}_{(q\neq\emptyset)}\right), \end{aligned} \tag{4.1}$$

其中 $\boldsymbol{y}$ 是弱信号 q_i 标记的真实标签。$n_i = \sum \mathbb{1}_{(q_i\neq\emptyset)}$，且 $\mathbb{1}_{(q_i\neq\emptyset)}$ 是一个指示函数，在弱信号标记的样本中返回 1 表示不放弃标记该样本。将弱信号的期望误差表示成 $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{\epsilon}$ 形式的线性方程组，定义每一行 $\boldsymbol{A}$ 为

$$\boldsymbol{A}_i = \mathbb{1}_{(q_i\neq\emptyset)}(1-2\boldsymbol{q}_i). \tag{4.2}$$

式 (4.2) 是弱信号 $\boldsymbol{q}$ 的线性变换。向量 $\boldsymbol{\epsilon}$ 的每一个分量 ϵ_i 是期望误差与弱信号和之间的差，即

$$\epsilon_i = n_i b_i - \boldsymbol{q}_i^\top\mathbb{1}_{(q\neq\emptyset)}. \tag{4.3}$$

期望误差的定义与真实标签 $\boldsymbol{y}$ 有关，但在本文的假设中，真实标签无法准确获得，于是使用对抗标签 $\hat{\boldsymbol{y}}$ 替代真实标签，即可行标签空间会随对抗标签的变化而变化。重新定义可行标签的约束空间

$$\boldsymbol{A}\hat{\boldsymbol{y}} \leq \boldsymbol{\epsilon}. \tag{4.4}$$

约束由弱信号 $\boldsymbol{q}$ 和错误率边界 b 决定。如果 b 太紧，则找到的优化方案可能不可行。相反，如果 b 太宽松，那么弱信号不能充分地约束目标，标签估计不会包含来自弱信号的信息。在实验时，我们分别使用弱信号的真实错误率和固

定值来作为错误率边界 b，并使用线性松弛惩罚来自适应地松弛约束，所以有约束条件 $\boldsymbol{A}\hat{\boldsymbol{y}} \leq \boldsymbol{\epsilon}+\boldsymbol{\xi}$，其中 $\boldsymbol{\xi}$ 是非负松弛变量，然后向目标添加松弛惩罚 $C\sum_{i=1}^{m}\xi_i$。

期望误差是根据真实的类标签 (硬标签) 定义的，但是约束空间得到的对抗标签是类概率，所以模型并不适合再使用期望误差来作为损失函数。L2 范数损失可以更恰当地表示分类器输出的概率标签相对于对抗标签的损失。L2 损失在预测值高估或低估时给出相似的损失，这种对称性使得模型对于过高或过低的预测都能受到适度的惩罚。另外，L2 损失可以看作是真实值与预测值之间的欧几里得距离的平方。它的定义如下：

$$\text{Loss} = \| f(X) - y \|_2^2, \tag{4.5}$$

其中 $f(x)$ 是模型的输出，y 是每个数据对应的真实值。由于是平方项，它对异常值较为敏感，即远离真实值的预测值具有较强的惩罚。另外，L2 损失是凸的，这意味着在最小化问题中只有一个全局最小值。这有助于使用梯度下降等优化算法找到全局最小值。

学习目标是拟合分类模型 $f_\theta(\boldsymbol{X})$，以符合弱信号约束。$\hat{\boldsymbol{y}}$ 是一个可行标签，我们依然考虑在最坏的情况下学习一个表现较好的模型，使用 L2 损失作为损失函数，所以基于 L2 损失的对抗标签学习优化以下问题：

$$\begin{aligned} \min_{\theta} \max_{\hat{\boldsymbol{y}}\in[0,1]^n} \quad & \| f_\theta(\boldsymbol{X}) - \hat{\boldsymbol{y}} \|_2^2 + C_1 \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \boldsymbol{A}_i\hat{\boldsymbol{y}} \leq \epsilon_i + \xi_i \ , \ \forall i \in 1, \cdots, m. \end{aligned} \tag{4.6}$$

我们可以把外部优化看作是优化一个原始目标，这个目标是有约束的内部优化的最大值。定义一个函数 $g(\theta)$，可以把式 (4.6) 等价地写成 $\min_\theta g(\theta)$。内部优化是给定一个分类器 f_θ，在标签的约束空间内寻找一个使分类器损失最大的对抗标签 $\hat{\boldsymbol{y}}$。外部优化是在学习到对抗标签的情况下，找到一个损失最小的分类器。

我们使用增广拉格朗日乘子法来优化式 (4.6)，它允许对所有正在优化的变量根据梯度进行更新，收敛时，拉格朗日函数找到一个局部最小值，能充分满足

约束条件。目标的增广拉格朗日形式是

$$L_2(\theta, \hat{\boldsymbol{y}}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \| f_\theta(\boldsymbol{X}) - \hat{\boldsymbol{y}} \|_2^2 + C_1 \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \gamma_i (\boldsymbol{A}_i \hat{\boldsymbol{y}} - \epsilon_i - \xi_i) - \frac{\rho}{2} \sum_{i=1}^{m} \| [\boldsymbol{A}_i \hat{\boldsymbol{y}} - \epsilon_i - \xi_i]_+ \|_2^2, \tag{4.7}$$

其中 $[\cdot]_+$ 是铰链函数，如果输入为正则返回输入，否则返回 0。$\boldsymbol{\gamma}$ 是 Karush-Kuhn-Tucker(KKT) 乘子，惩罚违反约束的行为。松弛变量和拉格朗日乘子都被限制为非负的。我们采用投影梯度上升或下降来更新变量 $\theta, \hat{\boldsymbol{y}}, \boldsymbol{\gamma}, \boldsymbol{\xi}$。分类器参数的梯度如下：

$$\frac{\partial \mathcal{L}_2}{\partial \theta} = 2(\frac{f_\theta}{\partial \theta})^\top (f_\theta - \hat{y}), \tag{4.8}$$

其中 $\frac{f_\theta}{\partial \theta}$ 是分类器 f 的雅可比矩阵。通过分类计算的反向传播，可以计算各种模型的雅可比矩阵。对抗标签的梯度如下：

$$\frac{\partial \mathcal{L}_2}{\partial \hat{\boldsymbol{y}}} = 2(f_\theta - \hat{\boldsymbol{y}}) + \sum_{i=1}^{m} \gamma_i \boldsymbol{A}_i + \rho \sum_{i=1}^{m} \boldsymbol{A}_i [\boldsymbol{A}_i \hat{\boldsymbol{y}} - \epsilon_i - \xi_i]_+, \tag{4.9}$$

$[\cdot]_0^1$ 将标签向量投影到 $[0,1]^n$ 中。每个 KKT 乘子的梯度如下：

$$\frac{\partial \mathcal{L}_2}{\partial \gamma_i} = \boldsymbol{A}_i \hat{\boldsymbol{y}} - \epsilon_i - \xi_i, \tag{4.10}$$

每个松弛变量的更新如下

$$\frac{\partial \mathcal{L}_2}{\partial \xi_i} = C_1 + \gamma_i + \rho [\boldsymbol{A}_i \hat{\boldsymbol{y}} - \epsilon_i - \xi_i]_+. \tag{4.11}$$

第二节　终端模型

终端模型的选择也是影响模型效果的一个重要因素。本文选择了两个分类器模型，分别是逻辑回归分类器和支持向量机分类器。

一、逻辑回归

逻辑回归是二分类线性判别模型。对于数据集 $\boldsymbol{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}, y_i \in \{0, 1\}$，线性模型的数学表达是

$$g(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x} \tag{4.12}$$

其中，$\boldsymbol{\theta} = [\theta_1, \cdots, \theta_n, b]^\top$ 是 $n+1$ 维的增广向量，前 n 维表示权值，b 表示偏置。$\boldsymbol{x} = [x_1, \cdots, x_n, 1]^\top$。在分类问题中，需要预测的目标值是离散的标签。因此，在线性模型基础上引入一个判别函数 $f(g(\boldsymbol{x}; \boldsymbol{\theta}))$，对值域进行收缩。逻辑斯蒂函数（logistic）可以将线性模型的值域放缩到 $[0, 1]$ 之间，它的表达式如下：

$$f(t) = \frac{1}{1 + e^{-t}} \tag{4.13}$$

如果将 g 视作 $\boldsymbol{x}$ 为正类的概率，则 $1 - y$ 是 $\boldsymbol{x}$ 为负类的概率，这样就有：

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}.$$

在 L2 对抗标签学习的设置中 y 是未知的，使用对抗标签 $\hat{y}$ 来替代它，且 $\hat{y}$ 表示样本属于正类的概率。所以有

$$\hat{y} = f(\boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}. \tag{4.14}$$

可求得 f 关于参数 $\boldsymbol{\theta}$ 的偏导数是

$$\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} = \frac{\boldsymbol{\theta} e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}{(1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}})^2} = -\boldsymbol{\theta} \hat{y}(1 - \hat{y}). \tag{4.15}$$

LR 是一个简单但高效的分类模型。使用 LR 作为终端模型，我们提出了基于逻辑回归的 L2 对抗标签学习方法（L2 Adversial Label Learning with Logistic Regression，LALL-LR），LR 的参数可以通过梯度下降算法直接学习，而且它不需要太多的计算资源。算法 2 展示了 LALL-LR 的训练过程。输入无标签数据、弱信号，需要训练得到分类器参数 $\boldsymbol{\theta}$。对抗标签依据均匀分布随机初始化。拉格

朗日乘子、弱信号的误差边界初始化是 0 向量。松弛因子初始化为 1 向量。然后根据式 (4.8)、(4.9)、(4.10)、(4.11) 用梯度下降算法依次更新各参数。

算法 2: LALL-LR 的训练过程

输入: 无标签数据 $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$，弱信号 $\boldsymbol{q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_m]$
输出: 分类器参数 $\boldsymbol{\theta}$
参数: 增广拉格朗日参数 ρ，学习率 α_t，最大迭代次数 S

随机初始化对抗标签 $\hat{\boldsymbol{y}}$(均匀分布)，拉格朗日乘子 $\boldsymbol{\gamma}$(0 向量)，松弛因子 $\boldsymbol{\xi}$(1 向量)，误差边界 $\boldsymbol{\epsilon}$(0 向量);
根据式 (4.2) 和式 (4.3) 写出矩阵 $\boldsymbol{A}, \boldsymbol{\epsilon}$;
while θ 不收敛且 $s < S$ **do**
    $\boldsymbol{\theta}^s \leftarrow \boldsymbol{\theta}^{s-1} - \alpha_t \frac{\partial \mathcal{L}_2}{\partial \boldsymbol{\theta}}$;
    $\hat{\boldsymbol{y}}^s \leftarrow \left[\hat{\boldsymbol{y}}^{s-1} + \alpha_t \frac{\partial \mathcal{L}_2}{\partial \hat{\boldsymbol{y}}}\right]_0^1$;
    $\gamma_i^s \leftarrow \left[\gamma_i^{s-1} - \rho \frac{\partial \mathcal{L}_2}{\partial \gamma_i}\right]_+$;
    $\xi_i^s \leftarrow \left[\xi_i^{s-1} + \alpha_t \frac{\partial \mathcal{L}_2}{\partial \xi_i}\right]_+$;
    判断收敛;
end

二、支持向量机

对于浅层模型来说，支持向量机解决高维特征的分类问题很有效，在特征维度大于样本数时依然有很好的效果，它可以利用大量的核函数来灵活的解决各类非线性分类问题。因此，在小数据集上，我们使用经典的 C 支持向量机来作为终端分类器。

SVM 旨在通过最大化两类样本之间的间隔和最小化结构风险，找到一个最优的分类超平面。假设有一个数据集 $\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}, y_i \in \{+1, -1\}$，最优分类超平面通过求解以下二次规划问题得到

$$\begin{aligned} \min_{\boldsymbol{\theta}, b, \boldsymbol{\xi}} \ & \frac{1}{2} \| \boldsymbol{w} \|^2 + C_2 \sum_{i=1}^{n} \eta_i \\ \text{s.t.} \ & y_i(\boldsymbol{\theta}^\top \boldsymbol{x}_i + b) \geq 1 - \eta_i \\ & \eta_i \geq 0, i = 1, \cdots, n \end{aligned} \tag{4.16}$$

使用 SVM 作为终端分类器，在训练时使用对抗标签 $\hat{\boldsymbol{y}}$ 代替真实标签 $\boldsymbol{y}$，即

对于终端分类器优化以下问题：

$$\begin{aligned} \min_{\boldsymbol{\theta},b,\boldsymbol{\eta}} \ & \frac{1}{2} \parallel \boldsymbol{\theta} \parallel^2 + C_2 \sum_{i=1}^{n} \eta_i \\ \text{s.t.} \ & \hat{y}_i(\boldsymbol{\theta}^\top \boldsymbol{x}_i + b) \geq 1 - \eta_i, \\ & \eta_i \geq 0, i = 1, \cdots, n. \end{aligned} \tag{4.17}$$

问题 (4.17) 的拉格朗日函数为

$$\begin{aligned} L(\boldsymbol{\theta}, b, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = & \frac{1}{2} \parallel \boldsymbol{\theta} \parallel^2 + C_2 \sum_{i=1}^{n} \eta_i \\ & - \sum_{i=1}^{n} \alpha_i [\hat{y}_i(\boldsymbol{\theta}^\top \boldsymbol{x}_i + b) - 1 + \eta_i] - \sum_{i=1}^{n} \mu_i \eta_i. \end{aligned} \tag{4.18}$$

这里 $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^\top$ 和 $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_m]^\top$ 分别式两组不等式约束的拉格朗日乘子。

将 L 对各变量求梯度并令其为 0，可得

$$\boldsymbol{\theta} = \sum_{i=1}^{n} \alpha_i \hat{y}_i \boldsymbol{x}_i, \tag{4.19}$$

$$\sum_{i=1}^{n} \alpha_i \hat{y}_i = 0, \tag{4.20}$$

$$C_2 - \alpha_i - \mu_i = 0. \tag{4.21}$$

将式 (4.19)、(4.20)、(4.21) 带回原问题 (4.17)，可得到原问题的对偶问题是：

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \ & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \hat{y}_i \hat{y}_j \boldsymbol{x}_i^\top \boldsymbol{x}_j - \mathbf{1}^\top \boldsymbol{\alpha} \\ \text{s.t.} \ & \sum_{i=1}^{n} \alpha_i \hat{y}_i = 0, \\ & 0 \leq \alpha_i \leq C_2, \ i = 1, 2, \cdots, n. \end{aligned} \tag{4.22}$$

式 (4.22) 可求得最优解 $\boldsymbol{\alpha}^* = [\alpha_1^*, \alpha_2^*, \cdots, \alpha_n^*]^\top$。于是得到分类器参数的最优

解为

$$\begin{aligned}\boldsymbol{\theta}^* &= \sum_{i=1}^{n} \alpha_i^* \hat{y}_i \boldsymbol{x}_i, \\ b^* &= \hat{y}_j - \sum_{i=1}^{n} \hat{y}_i \alpha_i^* (\boldsymbol{x}_i^\top \boldsymbol{x}_j).\end{aligned} \tag{4.23}$$

其中 j 表示 $\boldsymbol{\alpha}^*$ 中的第 j 个分量可以满足 $0 < \alpha_j^* < C_2$。

对于一个测试样本，SVM 判别分数的表达式为

$$f(\boldsymbol{\theta}^*, b^*) = \boldsymbol{\theta}^{*\top} \boldsymbol{x} + b^*. \tag{4.24}$$

若 $f(\boldsymbol{w}, b)$ 的值大于 0，则认为该样本属于正类，否则就是负类。最终的模型性能是由式 (4.24) 的结果来评价的。本文还涉及到使用核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \cdot \phi(\boldsymbol{x}_j)$ 把非线性可分问题转换成线性可分问题。

算法 3: ALL-SVM 的训练过程

输入: 无标签数据 $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$，弱信号 $\boldsymbol{q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_m]$
输出: 分类器参数 $\boldsymbol{w}$ 和 b
参数: 增广拉格朗日参数 ρ，学习率 α_t，最大迭代次数 S

随机初始化对抗标签 $\hat{\boldsymbol{y}}$(均匀分布)，拉格朗日乘子 $\boldsymbol{\gamma}$(0 向量)，松弛因子 $\boldsymbol{\xi}$(1 向量)，误差边界 $\boldsymbol{\epsilon}$(0 向量);
根据式 (4.2) 和式 (4.3) 写出矩阵 $\boldsymbol{A}, \boldsymbol{\epsilon}$;
while θ 不收敛且 $s < S$ **do**
    由式 (4.23) 得到 SVM 的最优解 $\boldsymbol{w}^*$ 和 b^*;
    $f = \boldsymbol{w}^{*\top} \boldsymbol{x} + b^*$，并根据式 (4.25) 进行概率校准;
    $\hat{\boldsymbol{y}}^s \leftarrow \left[\hat{\boldsymbol{y}}^{s-1} + \alpha_t \frac{\partial \mathcal{L}_2}{\partial \hat{\boldsymbol{y}}}\right]_0^1$;
    $\gamma_i^s \leftarrow \left[\gamma_i^{s-1} - \rho \frac{\partial \mathcal{L}_2}{\partial \gamma_i}\right]_+$;
    $\xi_i^s \leftarrow \left[\xi_i^{s-1} + \alpha_t \frac{\partial \mathcal{L}_2}{\partial \xi_i}\right]_+$;
    判断收敛;
end

算法 3 是基于支持向量机的 L2 对抗标签学习方法（L2 Adversial Label Learning with Support Vector Machine，LALL-SVM）的训练过程。为了促进快速收敛到局部均衡，我们先对 $\hat{\boldsymbol{y}}, \boldsymbol{\gamma}$ 和 $\boldsymbol{\xi}$ 进行初始化。然后使用 $\hat{\boldsymbol{y}}$ 优化分类器 f 的参数。但是 SVM 的输出结果是样本到决策边界的距离分数，并不是表示样本属于某一类的概率，而对抗标签的学习过程需要分类器输出的是样本属于正类的概率，于

是我们引入逻辑回归对 SVM 的判别分数进行概率校准，即有

$$p(y_i = 1|f_i) = \frac{1}{\exp(A f_i + B)}. \tag{4.25}$$

校准器的训练就是确定参数 A 和 B，参数是通过极大似然估计得到的。这个操作就会使得 SVM 输出样本属于正类的概率 p。接下来根据 p 用梯度下降法更新其他参数。

第三节　数值实验

一、实验设置

在实践中，可获得的弱监督源以噪声指标或简单标注函数的形式提供。这种弱监督给出样本属于正类的概率。由于我们对实验中使用的数据集没有明确的领域知识，因此我们通过在数据子集上训练简单的一维分类器来生成弱信号。用于训练弱信号的数据子集称为弱监督数据集。我们选择弱监督数据集的一个特征，然后只使用该特征训练一维的逻辑回归模型，把它作为一个弱监督源，对整个数据集进行预测，得到数据集的一个弱信号。我们选择弱监督特征是基于我们对哪些特征可以合理地作为目标类的指标的非专家理解。对于未提供特征描述的数据集，我们使用第一个特征，中间特征和最后一个特征来训练一维逻辑回归模型获得弱监督源。对于 Fashion MNIST 数据集，我们使用沿垂直中心线的 1/4、中心和 3/4 位置的像素值（如图4.2所示）来构建弱信号。

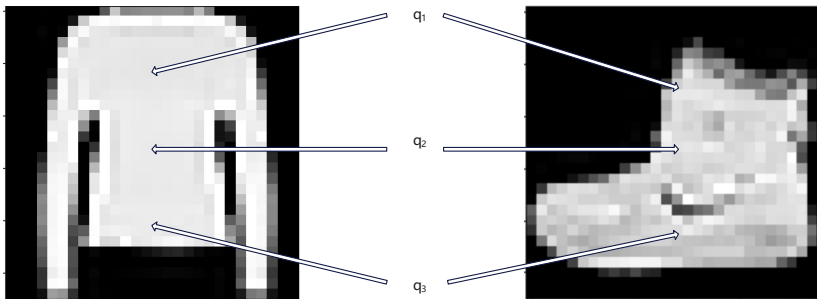


图 4.2 Fashion MNIST 数据集用来构建弱信号的特征

本章模型在 7 个数据集上进行了数值实验，包括 1 个图像数据集和 6 个 UCI 数据集。我们随机分割实验数据集，30% 用作弱监督数据，生成弱信号，40% 用作模型的训练数据，30% 用作模型的测试数据。在每个实验中，我们考虑三种不同的弱信号。以下是关于这 5 个数据集的简要概述。

- Fashion MNIST (FM)：每个样本都是 28×28 的灰色图像，这些图像分为 10 类服装类型，每个类包括 6000 个训练样本和 1000 个测试样本。我们考虑了三对二元分类：dress/sneaker(DvK), sandals/ankle boots(SvA), coats/bags(CvB)。
- Breast Cancer：这个数据集包括 569 个样本，其任务是诊断乳腺细胞是否来自乳腺癌的恶性 (阳性) 或良性 (阴性) 病例。使用细胞核的平均半径、半径标准误差和最坏半径作为训练弱监督模型的三个特征。
- OBS Network：该数据集是基于网络节点的行为来检测网络，判断是否应阻止其潜在的恶意行为。我们使用 the percentage of flood node, average packet drop rate, utilized handwidth 作为训练弱信号的特征。原始数据集包含四个类，我们选择了样本量最多的两个类，总共 795 个样本来做实验。
- Cardiotocography (Cardio)：该数据集的任务是使用产科专家分类的心脏生育图上的子宫收缩特征对胎儿心率进行分类。原始数据集包含 10 个类，我们选择最常见的两类，共 963 个样本进行试验。使用 accelerations per second, mean value of long-term variability, histogram median 作为特征来训练弱信号。
- Wine Quality：该数据集的任务是使用葡萄酒的物理和化学属性对葡萄酒的质量进行分类。原始数据集包含 7 个类，我们选择样本最多的两个类进行实验，共 3634 个样本。使用 fixed acidity、density 和 pH 三个特征来训练 3 个弱信号。
- Ionosphere：根据给定的电离层中的自由电子的雷达回波预测大气结构。g 表示好，b 表示坏。共有 351 个样本。由于每个特征没有确切含义，随机选择第 0、10、25 三个特征来训练弱信号。
- Dry Bean：该数据的主要任务是分析干豆的种类和质量。原始数据集包含 7 个类别，本文选择样本量最多的两个类进行实验，共 5573 个样本。选用 Perimeter、EquivDiameter、ShapeFactor3 三个特征训练弱信号。

使用分类准确率（Accuracy，ACC）来评估不同模型的分类效果，它是分类准确的样本数占该类样本总数的比例。我们将所提出的模型与以下方法进行了对比：

- Snorkel(Ratner et al., 2017)：这是一个用于弱监督学习的开源系统，它提供了一种灵活的方式来利用大量的、不完全标记的数据进行模型训练。它允

许用户通过规则、外部知识和其他源头的标记来生成训练数据，从而提高模型性能。

- ALL(Arachie 和 Huang, 2021)：这是一个一阶段的程序性弱监督模型，使用对抗的方式训练弱信号和分类器。使用弱监督数据集上每个弱信号的真实错误率作为误差边界 $\{b_1, \cdots, b_m\}$。
- Best Weak：使用一维的逻辑回归分类器对测试集进行预测的结果，预测时也只使用一个维度。选取 3 个弱信号中预测性能最好的结果。
- Average Weak：将三个弱信号在测试集上的预测结果求算数平均，得到一个关于测试集每个样本的概率预测。这个基线方法可以表示三个弱信号的平均质量。

关于模型的超参数设置，设置增广拉格朗日参数 $\rho = 2.5$，学习率 $\alpha_t = 0.0001$，最大迭代次数未 5000 次。分类器 SVM 的参数 C 通过网格搜索在 $\{0.05, 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$ 中选择，使用高斯核函数进行非线性映射，高斯核参数 γ 设置为数据集特征数的倒数。

二、实证分析

表 4.1 各弱监督方法在各数据集上的 ACC 值（%）

Datasets	Best Weak	Avarage Weak	Snorkel	ALL	LALL-SVM	LALL-LR
FM_CB	79.90	79.15	60.05	72.70	73.60	**83.15**
FM_DK	96.40	94.15	86.50	94.70	86.95	**98.30**
FM_SA	82.55	87.20	84.75	90.10	81.50	**91.35**
Cardio	**94.46**	**94.46**	60.21	69.55	68.86	83.04
Breast Cancer	92.40	90.64	90.64	91.81	89.47	**92.40**
OBS	69.87	69.87	69.87	**72.38**	69.04	71.13
Wine Quality	56.46	57.97	52.91	**60.25**	52.41	51.14
Ionosphere	75.47	75.47	64.15	73.58	70.75	**77.36**
Dry Bean	93.06	90.97	77.87	90.91	**94.08**	93.48

本章实验的评价指标是分类准确率，表4.1是各数据集在每种方法上的分类结果，从表中可以看出，LALL-LR 取得了比较好的效果，尤其是在图像数据集上表现良好，这说明我们对损失和约束空间的改进是可以提升模型效果的。但是 LALL-SVM 的效果不佳，即将分类器由逻辑回归改为支持向量机对弱监督模型没有很大的帮助。分析原因发现，模型在使用逻辑回归做分类器时，训练分类器的损失与训练对抗标签的损失是同一个，都是期望误差作为损失函数。而使用

SVM 做分类器时，我们用期望误差损失训练对抗标签，用铰链损失训练 SVM。损失的不统一可能会导致训练得到的对抗标签不是对 SVM 来说质量最差的标签。

第四节　　本章小节

首先，本章对程序性弱监督模型 ALL 进行了改进，提出了 LALL 框架。相比于原始 ALL 的期望误差损失，本文使用的 L2 损失可以更好地刻画分类器输出与对抗标签的距离。此外，本文还考虑了弱信号会放弃标记一些样本的情况。另外，本文为约束条件添加了松弛变量。利用松弛变量可以自适应地调节约束的松紧程度。

其次，本章对 LALL 的终端模型进行选择。选用了简单高效的逻辑回归分类器和可处理高维特征分类问题的支持向量机分类器。分别对两个分类器在本文中的使用原理进行了阐述。

最后，本章在 3 个图像数据集和 4 个 UCI 数据集上对所提出的方法进行了数值实验。对比了四种基线，发现模型对损失函数和约束条件的改进是有效的。但是我们对于使用 SVM 做分类器的尝试还有待改进。

第五章　总结与展望

第一节　　总结

本文主要研究了不完全监督的两种情形。一种是可获得少量标记为负类的样本，另一种是少量样本标签已知但不止一个类别。

针对第一种情形，最常见的就是处理不平衡二分类数据，比如异常检测。传统的异常检测方法如 OCSVM、SVDD 或者是一些深度方法 Deep SVDD 等，都是无监督方法，即训练时只使用正类样本，模型训练没有用到标签信息。在实际生活中，是可以获得一些负类样本的。尽管它们的数量很少，但是可以用来更加细化模型分类边界，提高模型的性能。SSLM 方法考虑了可获得少量负类样本的情况，是一种典型的不完全监督方法。但是 SSLM 是浅层模型，面对高维且大规模的数据，它的计算成本很高，而且数据表达能力有限。深度学习模型对于大规模和高维数据的适应性很强。它们在处理图像、语音、文本等复杂数据类型方面表现出色，能够捕捉数据中的复杂关系，而且可以更好地泛化到新的、未见过的数据。因此，本文在第三章中将 SSLM 模型拓展至深度学习框架中。Deep SVDD 是 SVDD 拓展至深度模型的范例，它为本文深度 SSLM 模型的提出提供了思路。因此，本文的建模思想是利用神经网络将数据映射到高维空间，然后在高维空间中构建一个超球，使得正类样本被包裹在球内，负类样本被排除在球外。少量的负类样本被用来细化分类边界，使得模型在面对真实世界的噪声数据时可以更加稳健。

Deep SVDD 还有一个缺陷，就是超球的中心取值为神经网络输出的均值，超球半径是通过分位数估计来近似计算的。这导致模型得到是近似解，而不是精确解。TELL 将聚类中心嵌入到神经网络中的做法为求解深度 SSLM 的参数提供了新的方向。因此，基于 TELL 的思想，本文在第三章提出了 ID-SSLMN 方法。它将分类器参数（即超球中心和半径）嵌入到神经网络最后一层参数中，这样可以通过反向传播算法统一地更精确求解网络参数和分类器参数。此外，为将分类器参数转化为神经网络参数的操作也使得神经网络最后一层参数具有可解释性。

为验证 ID-SSLMN 方法的有效性，本文第三章在 9 个数据集上进行了数值

实验。通过在两个人造数据双月和螺旋数据上在模型训练前后的数据可视化，证明了模型达到了我们建模的初始目标，即训练的网络和球体可以将正类样本映射到球内，将负类样本映射到球外。另外在 7 个真实数据上验证了所提出的 ID-SSLMN 与其他八种异常检测方法的效果，并对所有的实验结果做了费德曼检验，结果证明各方法在真实数据集上的结果是有显著差异的，且本文提出的 ID-SSLMN 在处理高维和低维数据方面都有很强的优势。说明本文将 SSLM 拓展至深度学习领域是一种有效的改进，且本文设计的更精确求解分类器的算法也是有效的。

本文使用程序性弱监督方法来解决第二种情形，即已知少量样本标签但不止一类的问题。程序性弱监督包括两阶段方法和一阶段方法两种。两阶段方法使用标签模型来聚合弱信号，获得一个伪标签。然后把伪标签当作真实标签训练一个监督模型（终端模型）。在两阶段方法的训练中，标签模型的输出是终端模型的输入，但是终端模型的训练结果不反馈给标签模型，这使得标签模型的训练变得非常重要，它直接决定整体的模型效果。而一阶段模型联合训练标签模型和终端模型，终端模型的训练可以影响标签模型的训练，一般来说它们是交替迭代的。这在一定程度上缓解了模型的计算压力。ALL 框架是经典的一阶段程序性弱监督方法。它通过对抗训练的方式，在约束范围内寻找一个质量最差的标签，然后利用这个标签来训练一个质量最好的分类器。这为我们解决数据标签不完整问题提供了思路。但是 ALL 的模型效果依赖于约束中期望误差边界的设置。过大的边界会使得约束过于宽松，限制模型优化，找到的优化方案不能展现良好的性能。过紧的边界会过度约束，从而无法找到合适的优化方案。

本文的第四章改进了 ALL 模型，提出了 LALL-LR 和 LALL-SVM。具体来说，本文对首先对 ALL 的损失函数进行了改进，因为原始期望误差损失是基于真实的标签（硬标签）刻画的，而本文所求得的对抗标签是软标签，表示样本属于正类的概率。相比于期望误差损失，L2 损失从几何角度来说可以更好地刻画两个概率值之间的距离。其次，本文为约束添加松弛变量，这样可以自适应地调节约束边界的大小。另外，本文考虑了一些弱信号会放弃标记样本的情况，这更接近实际生活中可获得的弱信号的情况。基于以上改进，本文提出了 LALL 框架。接下来，本文使用简单的逻辑回归和可处理高维数据的支持向量机分别参数化分类器，得到 LALL-LR 和 LALL-SVM 两个模型。本文在四个数据集包括

一个图像数据集和三个 UCI 数据集上验证了所提出方法的有效性。对比四条基线，本文提出的 LALL-LR 方法有明显的改进效果。

第二节　　展望

针对本文的数值实验结果，本节提出对于接下来可进行的工作的展望。

对于 ID-SSLMN 模型有以下几点：

- 使用迁移学习：使用预训练模型进行迁移学习，通过在大规模数据上进行寻来你，学习到普适的语言知识和模式。这种知识可以在其他任务上进行微调或迁移，无需从零开始训练模型，能够大幅减少在特定任务上的数据需求，并提升模型性能。
- 使用注意力机制：注意力机制允许模型在处理输入序列是动态地关注不同部分。这种灵活性和自适应性使得模型能够根据任务的需要调整对不同输入的关注度，提高模型的适应性。另外，由于模型集中注意力于输入序列中最相关的部分，从而已知噪声和荣誉信息的影响。这有助于提高模型对输入的处理效率和准确性。

对于 LALL-SVM 框架有以下几点：

- 搭建神经网络框架：本文所提出的弱监督学习框架以及所应用的分类器都是浅层模型。在大数据时代，为应对高维、大规模的数据，探索深度弱监督学习方法是必要的。它能够自动从数据中学习特征表示，减轻手动特征工程的负担。
- 构建统一的损失模型：目前 LALL-SVM 模型效果不佳的原因经分析是由于训练对抗标签使用 L2 损失，训练 SVM 使用铰链损失。可考虑将模型的整体损失修改为铰链损失。

参考文献

[1] 李航. 统计学习方法[M]. 北京：清华大学出版社, 2019.

[2] Abe N. Query learning strategies using boosting and bagging[C]// International Conference on Machine Learning. 1998: 1-9.

[3] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability[J]. Special lecture on IE, 2015, 2(1): 1-18.

[4] Arachie C, Huang B. A general framework for adversarial label learning [J]. The Journal of Machine Learning Research, 2021, 22(1): 5254-5286.

[5] Awasthi A, Ghosh S, Goyal R, Sarawagi S. Learning from rules generalizing labeled exemplars[J]. arXiv preprint arXiv:2004.06025, 2020.

[6] Bach S H, He B, Ratner A, Ré C. Learning the structure of generative models without labeled data[C]//Proceedings of the 34th International Conference on Machine Learning: vol. 70. 2017: 273-282.

[7] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts[C]//Proceedings of the 18th International Conference on Machine Learning. 2001: 19-26.

[8] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. 1998: 92-100.

[9] Boecking B. Learning with Diverse Forms of Imperfect and Indirect Supervision[D]. Carnegie Mellon University Pittsburgh, 2023.

[10] Boecking B, Neiswanger W, Xing E, Dubrawski A. Interactive weak supervision: Learning useful heuristics for data labeling[J]. arXiv preprint arXiv:2012.06046, 2020.

[11] Breiman L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[12] Cachay S R, Boecking B, Dubrawski A. Dependency structure misspecification in multi-source weak supervision models[J]. arXiv preprint arXiv:2106.10302, 2021.

[13] Chang C C, Lin C J. Training v-support vector classifiers: theory and algorithms[J]. Neural Computation, 2001, 13(9): 2119-2147.

[14] Chang W C, Lee C P, Lin C J. A revisit to support vector data description [J]. Technical Report, Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2013.

[15] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[16] Cramer J S. The origins of logistic regression[J]. Tinbergen Institute, Tinbergen Institute Discussion Papers, 2002.

[17] Cross G R, Jain A K. Markov random field texture models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983(1): 25-39.

[18] Dasgupta S, Hsu D. Hierarchical sampling for active learning[C]// Proceedings of the 25th International Conference on Machine Learning. 2008: 208-215.

[19] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in Neural Information Processing Systems, 2021, 34: 8780-8794.

[20] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial Intelligence, 1997, 89(1-2): 31-71.

[21] Dinh L, Krueger D, Bengio Y. Nice: Non-linear independent components estimation[J]. arXiv preprint arXiv:1410.8516, 2014.

[22] Foulds J, Frank E. A review of multi-instance learning assumptions[J]. The Knowledge Engineering Review, 2010, 25(1): 1-25.

[23] Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE transactions on neural networks and learning systems, 2013, 25(5): 845-869.

[24] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29: 131-163.

[25] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets[J]. Advances in Neural Infor-

mation Processing Systems, 2014, 27.

[26] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[27] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[28] Gurney K. An introduction to neural networks[M]. CRC Press, 1997.

[29] Hearst M A, Dumais S T, Osuna E, Platt J, Scholkopf B. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[30] Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld D S. Knowledge-based weak supervision for information extraction of overlapping relations[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: 541-550.

[31] Joachims T. Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th International Conference on Machine Learning: vol. 99. 1999: 200-209.

[32] Karamanolakis G, Mukherjee S, Zheng G, Awadallah A H. Self-training with weak supervision[J]. arXiv preprint arXiv:2104.05514, 2021.

[33] Khemchandani R, Chandra S. Twin support vector machines for pattern classification[J]. IEEE Transactions on pattern analysis and machine intelligence, 2007, 29(5): 905-910.

[34] Kingma D P, Welling M. Auto-Encoding Variational Bayes[J]. Stat, 2014, 1050: 1.

[35] Kingma D P, Welling M. An introduction to variational autoencoders[J]. Foundations and Trends® in Machine Learning, 2019, 12(4): 307-392.

[36] Kingma D P, Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions[J]. Advances in Neural Information Processing Systems, 2018, 31.

[37] Kingsford C, Salzberg S L. What are decision trees?[J]. Nature Biotech-

nology, 2008, 26(9): 1011-1013.

[38] Koller D, Friedman N. Probabilistic graphical models: principles and techniques[M]. MIT press, 2009.

[39] Latecki L J, Lazarevic A, Pokrajac D. Outlier detection with kernel density functions[C]//International Workshop on Machine Learning and Data Mining in Pattern Recognition. 2007: 61-75.

[40] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[41] Lewis D D. A sequential algorithm for training text classifiers: Corrigendum and additional data[C]//Acm Sigir Forum: vol. 29: 2. 1995: 13-19.

[42] Li Y F, Tsang I W, Kwok J T, Zhou Z H. Convex and scalable weakly labeled SVMs.[J]. Journal of Machine Learning Research, 2013, 14(7).

[43] Liu F T, Ting K M, Zhou Z H. Isolation forest[C]//2008 eighth ieee international conference on data mining. 2008: 413-422.

[44] Loh W Y. Classification and regression trees[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(1): 14-23.

[45] Mann G S, McCallum A. Generalized expectation criteria for semi-supervised learning with weakly labeled data.[J]. Journal of Machine Learning Research, 2010, 11(2).

[46] Mazzetto A, Sam D, Park A, Upfal E, Bach S. Semi-supervised aggregation of dependent weak supervision sources with performance guarantees[C]//International Conference on Artificial Intelligence and Statistics. 2021: 3196-3204.

[47] Miller D J, Uyar H. A mixture of experts classifier with learning based on both labelled and unlabelled data[J]. Advances in Neural Information Processing Systems, 1996, 9.

[48] Nguyen H T, Smeulders A. Active learning using pre-clustering[C]//Proceedings of the 21th International Conference on Machine Learning. 2004: 79.

[49] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models[C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 8162-8171.

[50] Nielsen M A. Neural networks and deep learning: vol. 25[M]. Determination Press San Francisco, CA, USA, 2015.

[51] Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 39: 103-134.

[52] Peng X, Li Y, Tsang I W, Zhu H, Lv J, Zhou J T. XAI beyond classification: Interpretable neural clustering[J]. The Journal of Machine Learning Research, 2022, 23(1): 227-254.

[53] Peterson L E. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.

[54] Ratner A, Bach S H, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid training data creation with weak supervision[C]//Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases: vol. 11: 3. 2017: 269.

[55] Ratner A J, De Sa C M, Wu S, Selsam D, Ré C. Data programming: Creating large training sets, quickly[C]//: vol. 29. 2016.

[56] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain.[J]. Psychological Review, 1958, 65(6): 386.

[57] Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui S A, Binder A, Müller E, Kloft M. Deep one-class classification[C]//Proceedings of th 35th International Conference on Machine Learning: vol. 80. 2018: 4393-4402.

[58] Schmidhuber J. A fixed size storage O(n3) time complexity learning algorithm for fully recurrent continually running networks[J]. Neural Computation, 1992, 4(2): 243-248.

[59] Schölkopf B, Smola A J, Williamson R C, Bartlett P L. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207-1245.

[60] Schölkopf B, Williamson R C, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection[J]. Advances in neural information processing systems, 1999, 12.

[61] Settles B. Active learning literature survey[J]. 2009.

[62] Seung H S, Opper M, Sompolinsky H. Query by committee[C]//Proceedings of the Fifth Annual Workshop on Computational Learning Theory.

1992: 287-294.

[63] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis[M]. Cambridge University Press, 2004.

[64] Suykens J A, Vandewalle J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9: 293-300.

[65] Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks[J]. Chemometrics and Intelligent Laboratory Systems, 1997, 39(1): 43-62.

[66] Tax D M, Duin R P. Support vector data description[J]. Machine learning, 2004, 54: 45-66.

[67] Tonolini F, Aletras N, Jiao Y, Kazai G. Robust weak supervision with variational auto-encoders[C]//Proceedings of the 40th International Conference on Machine Learning: vol. 1432. 2023: 34394-34408.

[68] Vapnik V. The nature of statistical learning theory[M]. Springer Science & Business Media, 1999.

[69] Varma P, He B D, Bajaj P, Khandwala N, Banerjee I, Rubin D, Ré C. Inferring generative model structure with static analysis[J]. Advances in Neural Information Processing Systems, 2017, 30: 239-249.

[70] Varma P, Ré C. Snuba: Automating weak supervision to label training data[C]//Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases: vol. 12: 3. 2018: 223-236.

[71] Wu M, Ye J. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092.

[72] Xiaojin Z. Semi-supervised learning literature survey[J]. 2006.

[73] Yuen M C, King I, Leung K S. A survey of crowdsourcing systems[C]//2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. 2011: 766-773.

[74] Zhang J, Hsieh C Y, Yu Y, Zhang C, Ratner A. A survey on programmatic weak supervision[J]. arXiv preprint arXiv:2202.05433, 2022.

[75] Zhou Z H. Multi-instance learning from supervised view[J]. Journal of Computer Science and Technology, 2006, 21(5): 800-809.

[76] Zhou Z H. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44-53.

[77] Zhou Z H, Li M. Semi-supervised learning by disagreement[J]. Knowledge and Information Systems, 2010, 24: 415-439.

[78] Zhu X, Ghahramani Z, Lafferty J D. Semi-supervised learning using gaussian fields and harmonic functions[C]//Proceedings of the 20th International Conference on Machine Learning. 2003: 912-919.