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

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。