



西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

2016 届

本科毕业论文（设计）

论文题目：判别式主动学习算法改进

学生姓名：秦浩晨

所在学院：统计学院

专 业：经济统计学（金融统计与风险管理实验班）

学 号：41627029

指导教师：李可

成 绩：

西南财经大学本科生毕业论文学术申明示例：

西南财经大学

本科毕业论文原创性及知识产权声明

本人郑重声明：所呈交的毕业论文是本人在导师的指导下取得的成果，论文写作严格遵循学术规范。对本论文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本毕业论文引起的法律结果完全由本人承担。

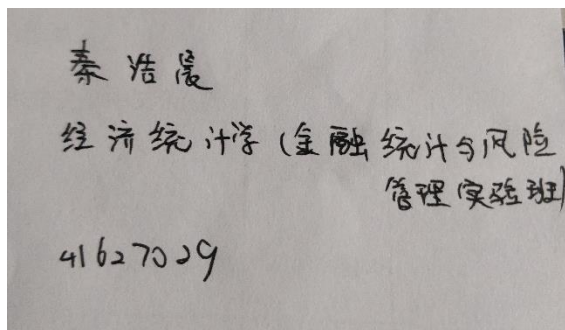
本毕业论文成果归西南财经大学所有。

特此声明

毕业论文作者签名：

作者专业：

作者学号：



秦浩晨
经济统计学(金融统计与风险
管理实验班)
41627029

2020年05月24日

2020年05月

摘要

传统的机器学习在大量有标签的数据上进行训练，在数据量足够大时能获得较高的准确率。然而很多实际任务往往面临有标签数据量十分有限而无标签数据量巨大的问题，针对此问题，主动学习算法从无标签数据中按照一定标准选取信息价值最高的样本子集加入训练集，避免了对整个无标签数据集打标签，减少了人工标记的成本和训练时间。传统的主动学习往往基于不确定性来定义无标签样本的信息价值，并根据该信息价值进行采样，而判别式主动学习将采样过程作为另一个分类任务进行处理，即判别某一无标签样本为已采样或是未采样样本。

该算法与传统主动学习算法相比，进一步提升了预测准确率并降低了所需的采样数。但是该算法将采样分类和任务分类去耦合为两个独立的阶段，也没有考虑迭代后期未采样数据集样本量下降对每次迭代采样数的影响。因此，本文在原始的判别式主动学习基础上，分别加入了采样数自适应机制，以及融合判别式采样和不确定性采样标准。除此之外，由于原始判别式主动学习算法中使用了没有惩罚项的损失函数，因而本文也验证了惩罚项对模型的影响。

本文首先回顾了主动学习的发展状况，对多个方向的主动学习机制进行了简单描述和分析。然后详细介绍了判别式主动学习的算法原理及其与领域自适应联系与区别，并提出算法仍存在的缺点。最后，分别进行实验验证三个新加入的机制对判别式主动学习算法在预测准确率和所需采样数的影响。

本文在 MNIST 和 CIFAR10 数据集上进行实验，实验结果为：加入正则项对模型的准确性和采样数几乎无影响；采样数自适应机制效果不稳定，在不同数据集上的作用不同，可能在减少采样数的同时牺牲预测准确率；而同时考虑判别式和不确定性的融合采样标准明显降低了采样数。

关键词：判别式主动学习；不确定性采样；采样数自适应；融合采样标准

Abstract

Conventional machine learning is rooted on tremendous labeled dataset, and higher prediction accuracy requires big dataset. In practice, however, lots of machine learning scenarios can only utilize a limited labeled dataset but large unlabeled dataset. In those cases, active learning which can actively sample informative data point from unlabeled dataset according to some given criteria, can reduce human labeling cost and model training time, which is a large profit for company and academia.

Conventional active learning algorithms usually choose uncertainty sampling as the criteria to define informativeness of unlabeled data, while discriminative active learning (DAL) conducts sampling as another classification task between unsampled and sampled data (both are in unlabeled dataset). In other words, in the DAL, there are two stages of classification which are uncoupled from each other.

Compared to past active learning algorithms, DAL has been proved to have higher prediction accuracy with less queried (sampled) data. However, since two classification stages in DAL are uncoupled (which might leads to “uninformative” sampling), and DAL does not deal with reduced unsampled data size when iteration process goes on, this paper will add sampling size adaptation and merging DAL with uncertainty sampling in the sample stage. Besides, since original DAL do not use regularization in the loss function, this paper will also test how penalization affects prediction accuracy and required sample size in DAL.

This paper begins with researching over papers in active learning domain, and then makes brief introduction and analysis about several subclasses of active learning. After that, this paper describes procedures of DAL algorithms and its relation to and difference from domain adaptation, and proposes its shortcomings. Then, this paper makes experiments about above three amendments (regularization, sampling size adaptation and merge).

Experimental results show that regularization makes nearly no difference, and sampling size adaptation has unstable effect(can cause different effect on different dataset, and may reduce sampling size with sacrificed prediction accuracy).However, merging DAL with uncertainty sampling can make significant reduction to required sample sizes without contaminating classifier.

Key word: discriminative active learning; uncertainty sampling; sampling size adaptation; merging DAL with uncertainty sampling

目录

| | |
|----------------------|-----------|
| 1. 绪论 | 1 |
| 1.1 研究背景与意义 | 1 |
| 1.2 研究内容与方法 | 2 |
| 1.3 论文结构 | 3 |
| 2. 文献综述 | 5 |
| 2.1 国内外研究现状 | 5 |
| 2.2 文献述评 | 6 |
| 3. 理论基础 | 8 |
| 3.1 主动学习概论 | 8 |
| 3.1.1 基于流的选择性采样 | 8 |
| 3.1.2 基于池的选择性采样 | 9 |
| 3.2 不确定性测量和抽样 | 9 |
| 3.2.1 三种常用的不确定性测量 | 9 |
| 3.2.2 机器学习中的版本空间概念 | 11 |
| 3.2.3 不确定性抽样与版本空间的关系 | 12 |
| 3.3 查询方式 | 12 |
| 3.3.1 异议质询 | 13 |
| 3.3.2 委员会查询 | 14 |
| 3.4 本章小结 | 15 |
| 4. 研究设计 | 16 |
| 4.1 判别式主动学习 | 16 |
| 4.2 改进的判别式主动学习 | 18 |
| 4.3 本章小结 | 19 |
| 5. 实证分析 | 20 |
| 5.1 实验设计 | 20 |

| | |
|------------------|----|
| 5.1.1 实验标准 | 20 |
| 5.1.2 实验数据 | 20 |
| 5.1.3 实验过程 | 21 |
| 5.2 实验结果与分析..... | 21 |
| 5.3 本章小结..... | 24 |
| 6. 总结与展望 | 25 |
| 6.1 全文总结..... | 25 |
| 6.2 未来展望..... | 25 |
| 参考文献..... | 26 |
| 致 谢..... | 28 |

1. 绪论

1.1 研究背景与意义

随着近年来大数据时代的发展，能够获取的数据量愈加呈现“数量大，获取容易但价值密度低”的特点。工业界面临诸多基于大数据的决策问题，传统的基于严谨的数学建模的机制在许多数据量巨大的场景中逐渐被端到端的算法机制所替代，因为后者能更快速地利用大数据的特征做出更准确的预测。因而，相应的机器学习算法以及深度学习也得到普遍发展。

机器学习是人工智能领域研究方向之一，通过使用带标签的数据训练模型预测相同任务下无标签数据的标签值。在很多问题领域中（比如异常检测，置信违约判断，图片分类，语音识别等），常规的机器学习和深度学习算法均可获得非常高的预测准确度，但是在实际问题中，常常只能获取少量的有标签数据和大量无标签数据，并且对数据进行标签的成本巨大。如果继续使用常规的机器学习算法，需要在训练模型之前对所有的无标签数据进行标记。从理论上讲这要求非常高的计算量，从实际上讲则需要极高的成本。因而建模者往往选取整个样本集中的一部分样本来进行标记，但问题在于如何选取样本：如果随机选择样本，就等价于只利用了采样比例部分的信息，因此问题转化为如何识别出样本集中最有价值的样本^[1]。

主动学习为机器学习的分支之一，通过主动“询问”的方式主动选取信息价值高的样本来进行模型训练。通常，使用主动学习可以减少模型训练中使用的样本量，从而减少计算复杂度。在有效的抽样方法和正确的标记下，主动学习算法在降低常规机器学习机制的成本时，也保证了令人满意的预测准确度^[2]。

近年来，主动学习的机制逐渐发展，并用于诸多应用场景。例如在使用计算机辅助医学成像研究时虽然可以从医院联网数据库中获取足够多的图像数据，但把所有的带有或是不带有病变信息的图片都进行标记是不现实的，因此往往需要主动从

未标签数据集中选择在决策边界附近的样本点；在进行 web 网页，音乐和视频推荐时，常规的机器学习要求用户标记出自己感兴趣的内容，但往往很少有用户愿意提供标签，在此情形下使用主动学习将可以降低计算成本。

本文首先对主动学习算法相关的研究进行分类描述和分析，然后介绍判别式主动学习算法的思想，最后对其进行改进，以实现在更少的采样情况下保障足够的准确率水平。

1.2 研究内容与方法

本文的研究内容为对判别式主动学习算法进行改进，研究方法包括通过文献综述和研究相关理论为算法改进提供基础、进行算法改进的研究设计和基于图像分类的实验设计。本小结先简要说明判别式主动学习算法的思想和本文的三个改进方向，然后简要说明本文使用的实验标准、数据和环境配置，具体内容将在第4,5章节“研究设计”和“实验设计与分析”进行详细描述。

本文旨在改进判别式主动学习算法：判别式主动学习的主要思想是在训练任务分类器之前，先训练一个采样分类器以判别初步抽取的未标签数据为已采样或是未采样数据，预测概率大，说明该样本越可能来自于未采样数据集因而与已标签数据集的差异越大，为了让模型能够拟合“真实”分布，该样本将被采样。换句话说，每次抽取的需要被标记的样本为采样分类器预测概率最大的样本点。

整个模型训练分为两个阶段：先使用采样分类器选择出信息价值高的样本加入训练集，然后利用合并后的训练集训练任务分类器。而该算法仍存在一定的改进空间，比如使用的分类器（原文中使用的是多层感知器）没有加入惩罚项；每次迭代的采样数为固定值；以及采样和模型训练为两个完全去耦合的阶段，没有考虑模型训练阶段对采样阶段的影响。因此，本文在已有的判别式主动学习算法基础上，从增加惩罚项、加入采样数自适应机制和并用多种采样标准三个方向进行改进。

1. 增加惩罚项：对采样分类器增加弹性网络范数。

2. 采样数自适应：随着迭代次数的增加，未被采样的无标签数据量减少，且剩下的无标签数据信息价值也是相对较低的，因而可以使用自适应机制让采样数随着迭代数的增加而减少。

3. 融合采样标准：判别式主动学习中，每次采样的标准仅依赖于采样分类器的预测概率值，没有考虑与任务分类器之间的关系。但是采样分类器使用的未采样数据集与已采样数据集（均为无标签数据集）的分布差异，与任务分类器使用的正标签数据集和负标签数据集（均为有标签的训练集）的分布差异可能是不尽相同的，因而采样时应该同时考虑两种分布差异的信息来选择信息价值高的未采样无标签数据。换句话说，每次迭代中的采样数据点，应该同时满足采样分类器预测概率值大，以及位于任务分类器的分类边界（在多分类场景中，则为信息熵最大）。

由于本文的研究目的在于提升预测准确率，或是保证预测准确率不降低的同时进一步减少判别式主动学习的采样数以减少模型训练时间，因此研究方式为通过比较判别式主动学习和三种不同改进的训练时间和预测准确率来验证改进方式的有效性。本文使用 TensorFlow 中的 Keras 多层感知器模型接口，实验数据集为 MNIST 和 CIFAR10，训练环境为 python3.6+单卡 RTX2070。

1.3 论文结构

基于上述分析，本文的大致结构如下图所示：

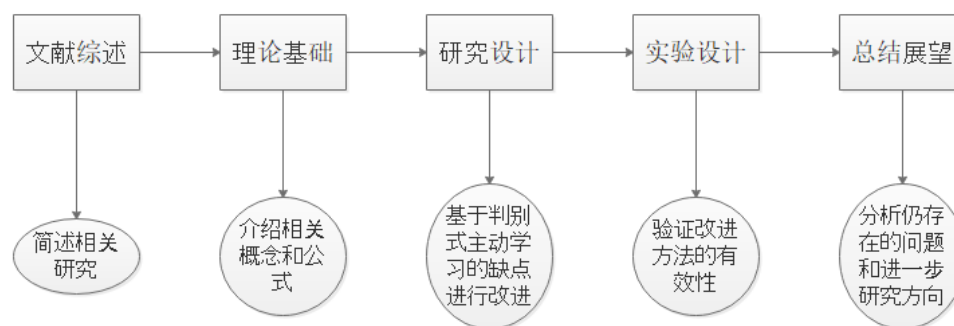


图 1-1 论文流程

第2章节对相关研究现状进行描述和分析，第3章节对主动学习相关重要理论进行补充介绍，第4章节针对判别式主动学习的缺点改进设计，并在第5章设计实验进行验证，最后在第6章总结全文并提出未来进一步研究方向。

2. 文献综述

主动学习的关键在于从无标签样本中有效提取信息度高的样本以减少模型训练的时间，因此近年来国内外的研究集中于如何实现有效的采样算法以保证在较少的标签数据中模型也能取得较高的预测准确率。本章节首先简要陈述关于采样的研究内容，然后对其中的不足之处进行分析，给判别式主动学习算法思想的引入和对比分析提供基础。

2.1 国内外研究现状

起初的主动学习主要基于“委员会”查询：委员会查询使用迭代查询的思想，在每一次迭代中随机选取版本空间 V 中的两个假设(Seung et al., 1992)^[3]。在之后的研究中，委员会查询逐渐采取另一种渠道，即构建一个委员会 C （假设空间的子空间）。在实现细节上，有多种不同的方式来构成 C ：在无噪声数据的情形中，可以直接从定义良好的版本空间 V 中进行随机抽样以构成 C (Freund et al., 1997)^[4]，该算法尽管没有理论证明其准确性，但由于假定了数据是无噪声或是噪声可以忽略不计，最终的模型训练结果也是非常好的。此外，为改进单纯的随机抽样方式，McCallum and Nigam (1998)使用贝叶斯方法基于后验分布 $P(\theta|L)$ ，并使用狄利克雷分布作为模型参数抽样的分布函数。在序列数据中，Dagan and Engelson (1995)^[5]使用基于正态分布族的隐马尔科夫模型对模型参数进行采样。基于抽样分布的思想，Freund and Schapire(1997)^[6]使用提升方法（boosting）对序列假设进行采样，使得学习者能够更擅长对分类边界附近的样本点进行准确分类。基于模型集成的思想，Breiman(1996)^[6]使用基于袋的查询（query by bagging），对有标签样本集进行重采样以训练一个委员会模型集。在此基础上，Muslea 等人(2000)^[7]通过划分特征空间来对委员会模型集进行区域划分成条件独立的子集合，以提高对模型集进行选择时的速度。

近年来, 主动学习则更多地集中于不确定性查询。诸多研究已证明不确定性查询中的多种不同的启发式算法都能有效实现主动学习的抽样任务^[8]。Yang 和 Loog 在 2018 年使用逻辑回归模型, 验证了不同的启发式方法都可以有效进行采样^[9], 其中基于不确定性抽样的方式效果最好。在将主动学习的机制应用于深度学习神经网络中时, Gal 等人在 2017^[10]年提出使用基于蒙特卡洛抛弃 (MC dropout) 来提高不确定性测量, 并且分别使用了基于熵和基于变分率的查询方式。从本质上来讲, 该算法使用蒙特卡洛抛弃而不是随机抛弃, 实现了主动学习机制在深度学习中对参数点进行主动选择, 在牺牲一定计算时间的情况下获得了准确性的提高。Ducoffe 和 Precioso 在 2018 年^[11]使用对抗的机制来实现深度神经网络中的主动学习, 其思想在于将最接近的对抗样本作为选取的样本, 充分利用了对抗学习的数据增强和噪声处理优势, 提高了模型的稳健性。Huang 等人在 2016 年^[12]提出主动学习者可以采样具有最大的期望梯度范数的样本点, 该算法中的期望表示在后验分布中, 对所有标签所对应的梯度范数进行求期望。与传统的三种不确定性测量相比, 该算法更充分利用了样本的分布信息。

除了上述采样方法之外, 该领域中还有关于基于核心集的方法实现采样的研究。该方法需要首先确定一个数据中心, 然后设定一个距离阈值 ϵ 以选择出该中心数据的所有邻域样本点。通常选取一个合适的阈值 ϵ 是非常困难的, 因此 Sener 和 Savarese 在 2018^[13]提出用整型优化来近似查找最优的 ϵ 。

最后, 本文欲改进的判别式主动学习由 Gissin 等人在 2019 年提出, 其主要思想是将抽样作为另一个分类任务, 在 MNIST, CIFAR10 和 CIFAR100 等数据集的图像分类任务上取得了优于其他基于不确定性的启发式抽样算法的结果。

2.2 文献述评

上述研究中, 基于委员会的查询算法需要建立一个委员会模型集合 C , 尽管能保证抽取的样本能够尽可能代表无标签数据的分布信息, 但由于计算量过大, 在近年来的大数据环境中逐渐被其他的算法所代替。另一方面, 基于不确定性和基于核心

集的抽样算法更为简便，在多个人工智能领域任务中均获得了较好的效果，判别式主动学习以不同于不确定性抽样的方法论进行抽样，在图像识别任务中进一步降低了所需的抽样数。

然而，判别式主动算法是一个两阶段去耦合的分类过程，不确定性采样则在采样阶段就考虑了抽取的样本对任务分类器的影响。两者使用不同的采样标准，对不同的信息进行了表示，前者为是否被采样的分布信息，后者则为正负样本差异的分布信息。本文将两个不同的采样标准进行融合，旨在同时使用两种不同的信息，以尝试对判别式主动学习进行改进。除此之外，本文还对该算法的其他方面进行分析和改进，具体过程将在 4.5 章节进行详细阐述。

3. 理论基础

对主动学习相关研究简要回顾后，本章节以补充的形式，在 3.1 节介绍主动学习的基本思想、算法目的和两种不同的适用场景，在 3.2-3.3 节介绍两种不同的采样实现方式。

3.1 主动学习概论

主动学习的目的是期望机器学习算法能够在更少的样本上获得相同或是更高的预测准确性。一个主动学习算法机制需要实现从无标签数据集中进行有选择的抽样，然后被权威所标签（比如人工标记），最后提供给相应的模型进行训练。

在主动学习框架中，存在一个“学习者”机制。该机制基于预先给定的标准来从整个训练样本中选取最有价值的一部分，使得模型可以在较小的样本集中同样能获得令人满意的预测结果。由于在实际应用中，对数据打标签是一个困难、耗时且昂贵的过程，因此设计一个好的标准来主动选取样本（称为“查询选择算法”）进行训练是很有意义的。通常来讲并不是所有的情形都适合使用主动学习，只有当如下两个条件满足的时候主动学习机制才能带来很大的提升：无标签训练数据量大，容易收集但是难以打标签；有标签数据量有限，但需要足够多的数据才能训练出有效的分类器模型。

根据使用的数据形式，主动学习通常有两种应用场景，即基于流的选择性采样（stream-based selective sampling）和基于池的选择性采样（pool-based selective sampling）。

3.1.1 基于流的选择性采样

该机制适用于流数据环境，学习者可以较轻易地获得足够多无标签数据。因此，学习者可以基于近似真实的分布来进行第一轮筛选，然后再根据一定的标准来进行第二次筛选。两阶段筛选可以保证模型获取带足够有价值的样本，同时也不会导致计算量过大。与查询合成相比，基于流的选择性采样充分利用了数据获取容易，数

据量大但价值密度低的特性，因而往往具有更好的模型性能。然而，如果流数据近似均匀分布，一阶段筛选将毫无所用，导致基于流的算法退化成查询合成算法。基于流的选择性算法在近二十多年的研究中已有效应用于多个任务领域，比如词性标注(Dagan and Engelson,1995)^[16]；信息检索中的排序(Yu,2005)^[17]；

3.1.2 基于池的选择性采样

在现实场景中，除了流数据，数据的获取还存在池数据的形式，即一次性获取大量数据并在一段时间内不再获取新的数据。基于此，学习者在给定的一段时间内将只能在同一个数据池中进行选择，所有数据采样都是基于不变的分布特征。由于池数据量的巨大，学习者往往使用贪婪算法的思想进行采样以降低采样的算法复杂度，比如高维特征下的二分查询算法(Lewis and Gale,1994)^[18]

基于流和基于池的选择性采样的主要区别在于，前者一次查询只获取一个样本，并且每次查询独立；后者对池数据按照一定标准对池中的样本进行排序后，一次查询即可获取多个最有价值的样本。尽管从算法机制角度分析时，基于池的选择性采样的查询效率更高，但很多实际任务中的数据都是流形式，因而两种机制都存在很大的研究价值。考虑到本文实验环境均为一次性的大样本任务，因而下一小节仅对基于池的选择性采样的原理进行详细叙述。

3.2 不确定性测量和抽样

上一小结简要介绍了主动学习常用的两种应用场景，本小结对主动学习的采样理论进行介绍：依次介绍主动学习采样过程中使用的不确定性抽样方式、传统机器学习中的版本空间概念以及两者的理论联系。

3.2.1 三种常用的不确定性测量

主动选择具有价值的样本进行标签是主动学习算法的关键，因此如何定义样本的价值是主动学习的研究关键。在最基础的二分类任务中，往往介于分类边界

（decision boundary）的样本是最具有价值的，即预测的概率值约等于 0.5。在此基础上，可以对多标签场景的复杂任务中的样本价值进行定义：

1. 基于最低置信（least confidence）的不确定性测量：

$$x_{LC}^* = \underset{x}{\operatorname{argmin}} P_{\theta}(\hat{y}|x)$$

其中， $\hat{y} = \underset{y}{\operatorname{argmin}} P_{\theta}(y|x)$ 表示具有最高后验概率的预测值，此标准本质上也是选取最靠近分类边界的样本。该方法的缺点在于仅使用了后验概率最低的哪些样本，而没有利用剩余样本的分布信息或是后验概率信息。

2. 基于间隔（margin）的不确定性测量：

$$x_M^* = \underset{x}{\operatorname{argmin}} [P_{\theta}(\hat{y}_1|x) - P_{\theta}(\hat{y}_2|x)]$$

其中， \hat{y}_1 和 \hat{y}_2 分别是样本集中具有最高和第二高后验概率的样本。该方法的思想在于：一个较弱学习者就能容易地区分两个差别大的样本，而我们期望获得即使差别很小的两个样本也能区分开的强学习者。虽然此方法一定程度上解决了上述基于最低置信方法的缺点，但仍然忽略了大部分数据的分布信息。

3. 基于熵（entropy）的不确定性测量：

$$x_H^* = \underset{x}{\operatorname{argmin}} H_{\theta}(y|x) = \underset{x}{\operatorname{argmax}} - \sum_y P_{\theta}(y|x) \log P_{\theta}(y|x)$$

熵衡量了变量的平均信息量，该方法使用了所有样本的熵值，因此经常被用在机器学习的损失函数中作为不确定性和不纯度的表示。

三种不确定性测量的内在关系和对比如下图所示(横轴为概率值，纵轴为相应的不确定测量的值)：

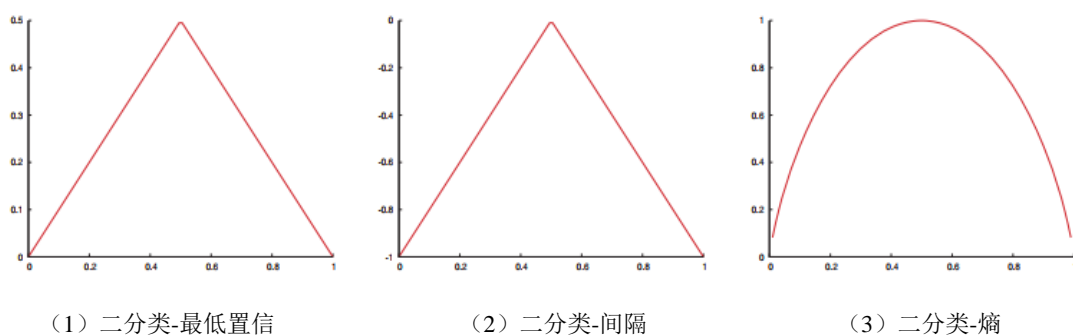


图 2-1 不确定性函数与样本概率的关系

在二分类任务中，三种函数都是同增同减的，均在后验概率为 0.5 获取最大的测量值。另一方面，基于最低置信和基于间隔的策略在二分类任务中是完全等价的，基于熵的策略则会给 0.5 后验概率的临近位置赋予比其他两种策略更好的测量值，因而前两者将会的后验概率的变化更为敏感，即在给定相同的阈值下，前两者更可能拒绝对某个样本进行抽样。从理论上讲，前两者适合用于损失函数为 0-1 损失，而基于熵的策略则适用于对数似然损失函数。

3.2.2 机器学习中的版本空间概念

在机器学习中，“假设”的概念指的是某个特定的用于近似和解释输入样本特征的分布模型，并基于该假设，模型能够新的数据集上进行预测。将假设空间标记为 H ，在机器学习算法中， H 由所有可能的权重参数组合成，而最优假设则是假设空间中能够最大化近似训练集数据特征分布的一个特定权重参数组。

版本空间 V 是假设空间 H 的一个子空间，与训练数据的特征一致，即基于版本空间 V 中的任意一个参数组合，我们都能在训练集上做出准确的预测。因此，版本空间也可以代表能很好解释输入数据的“假设候选集”。通常，我们可以假设模型学习的函数是可分离的，即随着训练集数据量的增大，我们可以逐渐排出假设候选集中一些不再准确预测的参数组合。每次迭代将会减小版本空间的假设个数，最终在足够多的数据支持下，版本空间中剩下的假设能够对真实的目标函数进行近似表示，其原理类似于中心极限定理中的无穷大数据情形下的极限分布近似。

3.2.3 不确定性抽样与版本空间的关系

不确定性抽样通常是一次查询抽取单个样本，而版本空间则是在确定最优的参数组合后选取一整块样本集。从概念上来讲，两种方法似乎是完全不同的，但在特定情形下，不确定性抽样可以作为工具，用于版本空间中的假设候选集迭代减小过程：在每次迭代中，利用不确定性抽样进行单个样本采样，增加总的训练集样本并减小假设候选集，重复上述步骤直到假设候选集足够小并能够准确地近似目标分布。

由于不确定抽样每次选取无标签数据集中最接近分类边界的样本点，因此该样本点可以最大化地识别出当前假设候选集中不能够满足新加入样本点的参数组合。相比较于随机抽样，新加入的样本点很可能无法识别出不满足要求的参数组合。因此，从理论上讲，足够多次的不确定性抽样在为机器学习算法提供样本时，与版本空间的效果是等价的。

然而，使用不确定性抽样以实现版本空间的效果仍然存在一些问题。首先，基于间隔的不确定性抽样只有当版本空间是对称的时候，才能将所有的参数组合二等分。如果参数组合不为近似对称，在使用多次不确定性抽样把版本空间进行划分后，可能会有成团的参数组合（参数组合预测出的标签的分布相近）无法被进一步的不确定性抽样所分割。换句话说，不确定性抽样期望实现近似二分查找的效果，但是当版本空间高度不对称（偏斜）时，多次的不确定性抽样只能进行原地查找；其次，上述分析假设了不确定抽样的效果是对版本空间进行查找，但是没有理论证明保证该查找机制具有稳健性，因此其效果在很大程度上取决于实际问题的特点和获取数据的分布特征。

3.3 查询方式

在早期主动学习研究领域里，除了对不确定性测量进行分析和改进之外，还有另一种机制进行主动抽样，即基于查询的抽样。基于查询的主动学习在更大程度上与机器学习的假设空间进行联系，其主要思想为通过直接搜寻假设空间中的参数组

合，以确定是否需要对数据进行查询。基于查询方式的主动学习通常可以分为异议质询和基于委员会的查询两种方式。

3.3.1 异议质询

异议质询 (query by disagreement) 通常用于流数据任务中，即学习者需要实时地对新加入的单个样本进行判断是否需要查询。在该算法中的某一个时间点保存一个“运行中的”版本空间 V ，如果新加入的样本点使得版本空间 V 中的两个有效假设获得了不同的标签预测值，那么该样本点需要被查询，即需要被打上标签并加入训练集。相反，如果 V 中的所有有效假设都对该样本点做出了一致的预测，那么该样本点将会抛弃，因为它对模型的提升没有显著帮助。

从几何角度上讲，被异议质询的样本构成了一个异议空间 (region of disagreement)，异议质询的主动学习算法将只抽取特征空间落入该空间的样本点。该算法的优点在于非常直接和简便，能够指数级减少需要被标签的数据量：简便的原因在于只用知道 V 中有效假设的标签预测值，而不需要获取分布信息或概率值；指数级减少的原因在于需要所有的假设都一致通过的样本点才能够被采样。

然而，异议质询在实际应用中仍然存在一些缺点。首先，版本空间 V 可能具有非常多甚至是无限个假设，因此无法在内存中保存所有的有效假设，尽管分布式计算能在一定程度上缓解该问题，但过大的计算量使得计算版本空间中所有假设不是首要选择。另一种解决办法是使用隐式表示：在二分类问题中，设定两个学习者模型以代表 V 中的所有可能参数组合，然后将已标签数据集和异议空间的正样本所组成的并集提供给学习者一进行学习，将已标签数据集和异议空间的负样本提供给学习者二进行学习。

$$h_1 = \text{train}(L \cup \langle x, + \rangle)$$

$$h_2 = \text{train}(L \cup \langle x, - \rangle)$$

其中， L 是已标签数据集， train 是一个给定机器学习算法， h_1 和 h_2 分别代表该算法的两个不同参数组合。

3.3.2 委员会查询

委员会查询(query by committee)是对异议质询的改进。通常异议质询需要满足两个假设：首先，需要对版本空间 V 中的所有候选假设进行验证（或者使用近似的两个极端学习者 h_1 和 h_2 ）其次，异议质询是一个基于二分的测量方法，一个样本点要么被假设集的所有假设一致通过并抛弃。要么得到不同的质询结果并作为抽样对象，该算法没有保存介于这两者之间的任何信息。第一个假设将会导致计算量过大或者近似的学习者无法表示整个假设空间，而第二个假设则导致损失很多的样本信息，使得模型训练结果可能存在很大的问题。因此委员会查询放松了该两条假设，使用基于信息熵的标准来进行样本查询，推广了主动学习算法的应用范围。

基于信息熵的一些测量方法包括：投票熵(vote entropy)、软投票熵(soft vote entropy)和基于 KL 散度的信息量衡量。

1. 投票熵：

$$x_{VE}^* = \underset{x}{argmax} - \sum_y \frac{vote_C(y, x)}{|C|} \log \left(\frac{vote_C(y, x)}{|C|} \right)$$

其中 $vote_C(y, x) = \frac{1}{|C|} \sum_{\theta \in C} \mathbf{1}_{\{h_{\theta}(x)=y\}}$ 代表了样本点标签为 y 的投票数，对应的是 0-1 损失。该信息量衡量函数的计算量较小，但损失了分布信息。

2. 软投票熵：

$$x_{SVE}^* = \underset{x}{argmax} - \sum_y P_C(y|x) \log P_C(y|x)$$

其中 $P_C(y|x) = \frac{1}{|C|} \sum_{\theta \in C} P_{\theta}(y|x)$ 代表委员会模型集中标签 y 是正确的且一致的概率。与投票熵相比，软投票熵基于贝叶斯方法的后验概率，利用了更多的分布信息而不是简单地标签信息，在一定程度上能够平滑参数的变化。

3. 基于 KL 散度的信息量衡量：

$$x_{KL}^* = \underset{x}{argmax} \frac{1}{|C|} \sum_{\theta \in C} KL(P_{\theta}(Y|x) || P_C(Y|x))$$

其中 KL 散度为 $KL(P_{\theta}(Y|x)||P_C(Y|x)) = \sum_y P_{\theta}(y|x) \log \left(\frac{P_{\theta}(y|x)}{P_C(y|x)} \right)$ 衡量了某个参数组合 θ 对应的分布于委员会参数集 C 对应的分布的差异值。与软投票熵方法相比，此方法利用了不仅利用了分布信息，而考虑了与委员会分布信息的差异值。但往往 KL 散度的计算量异常昂贵，在使用中应该根据数据量和机器学习主题算法的复杂度进行合理选择。

3.4 本章小结

本章节对主动学习的理论进行了简要介绍，包括主动学习适用的场景和常用的两种获取样本的方式（基于不确定性和基于查询的）。从本质上讲，基于不确定性的抽样和基于查询的方法都是为了从无标签数据中选择出信息量最为丰富的样本，这与判别式主动学习的思想也是一致的，但后者再采样的具体实现细节熵有略微不同。下一章节将引入该算法的主要思想，分析其中的缺点并基于此进行研究设计。

4. 研究设计

从本质上讲，无论是基于不确定性抽样，还是利用即将介绍的判别式主动学习，其目的都是在于获取未标签数据中的一小部分子集来表示整个数据集的分布信息。这一思想与机器学习中的另一话题“领域自适应”有一定的联系和区别。因此，本章节将在 4.1 节详细描述判别式主动学习算法的主要思想，然后分析其与领域自适应算法的相似与不同之处，相似之处在于进一步证明其理论合理性，而不同之处则为本文提供了改进算法的思想；4.2 节详解介绍算法改进的原因和实现细节。

4.1 判别式主动学习

判别式主动学习的主要思想是将机器学习分为两个阶段的分类任务^[19]：在第一个阶段中，先训练一个采样分类器以判别还未被主动选择的未标签数据和已经被选择的未标签数据，将预测概率大的样本（即更可能是未被采样）进行打标签并加入训练集；在第二个阶段中，在训练集上训练分类器，即为传统的被动式机器学习算法的阶段。

过去的主动学习算法在采样阶段（从未标签数据集中选择出信息量大的样本）往往基于给定的信息标准，比如后验概率最大化和间隔最大化等方法。这些方法大多基于信息论的公式来选择出有价值的样本，而判别式主动学习则直接训练一个采样分类器以区分未标签数据集中已经被采样的和未被采样的。对于采样分类器预测出的越接近未被采样类型的样本，其信息与已经被采样的样本有很大程度不同，能为任务分类器提供较多的信息，因而在采样阶段被加入训练集。该算法流程如下所示：

算法一 判别式主动学习

```
1: 初始化  $U, L, K, n$       #  $U, L$  分别为无标签和有标签数据,  $K$  为总采样数,  $n$  为每次迭代最小采样数
2: 初始化  $C_s$  和  $C_m$       #  $C_s$  和  $C_m$  分别为采样分类器和任务分类器
3: for  $i = 1 \dots n$  do
4:    $P = C_s(U, L)$ 
5:   for  $j = 1 \dots \frac{K}{n}$  do
6:     $\hat{x} = \underset{x \in U}{\operatorname{argmax}} P(y = u|x)$ 
7:     $L = L \cup \hat{x}$ 
8:     $U = U \setminus \hat{x}$ 
9:   end for
10: end for
11: 返回  $L, U$ 
```

此思想与领域自适应(domain adaptation)方法有紧密联系和区别之处。领域自适应解决的问题是减少在源分布(source distribution)上训练的模型在目标分布上的预测误差值, 即通过训练一个模型来就可以从源分布得到目标分布。为减少预测误差值, 使用如下损失函数对预测误差进行限定阈值:

$$d_H(D_S, D_T) = 2 \sup_{h \in H} |P_{x \sim D_S}[h(x) = 1] - P_{x \sim D_T}[h(x) = 1]|$$

对于判别式主动学习, 只用将上述损失函数中的源集, 目标集分别替换成已标签数据集和未标签数据集。因此, 其与领域自适应具有等价的损失函数 (如下所示)。

$$d_H(D_S, D_T) = 2 \sup_{h \in H} \left| \frac{1}{|L|} \sum_{x \in L} h(x) - \frac{1}{|U|} \sum_{x \in U} h(x) \right|$$

其中, $D_S = \frac{1}{|L|} \sum_{x \in L} h(x)$, 因此解决该最大化的优化问题等价于找到分类器 H 能够最大程度区分 L 和 U 。换句话说, 解决了领域自适应问题, 也就找到了判别式主动学习中的采样分类器。

4.2 改进的判别式主动学习

尽管判别式主动学习与领域自适应的相似之处在一定程度上说明其采样机制的可行性。但在实现细节上，该两者也存在一定的区别，从而导致算法还存在一定的改进空间。领域自适应中，分类器 H 能够同时最小化源集上的训练误差和目标集上的测试误差。而在判别式主动学习中，先训练采样分类器，然后基于它的预测值选取信息量最大的样本，最后加入到原有已标签数据集训练分类器。换句话说，领域自适应为端到端的一阶段学习过程，而判别式主动学习将无监督过程和监督学习过程分开处理，是一个两阶段的半监督学习机制。

通过上述分析可以看出，作为半监督的主动学习算法，判别式主动学习算法是一种将采样和分类器训练去耦合的机制，这种方式的好处在于损失函数的计算更容易。但同时，这也不可避免地导致了一些问题，即由于在采样阶段没有考虑样本的正负标签信息，导致可能抽取了一些无价值的样本。因此本文在采样阶段，将同时考虑采样样本的两个信息：是否靠近未采样和已采样类别的分类边界，和是否靠近正负标签样本的分类边界。

同时，由于判别式主动学习的采样分类器没有加入惩罚项，在图像分类中会导致产生过多的参数而增加训练时间，本文也会验证加入惩罚项对模型训练时间和预测准确率的影响；另一方面，原文中的判别式主动学习在每次迭代中的采样数为固定值，当迭代数增加后，未采样数据集的总量减小，并且信息价值降低，因而本文使用采样数自适应的机制以减少迭代中的采样数。改进后的判别式主动学习算法如下所示：

算法二：改进的 判别式主动学习

```
1: 初始化  $U, L, K, n$       #  $U, L$  分别为无标签和有标签数据,  $K$  为总采样数,  $n$  为每次迭代最小采样数
2: 初始化  $C_s$  和  $C_m$       #  $C_s$  和  $C_m$  分别为采样分类器和任务分类器
3: for  $i = 1 \dots n$  do
4:    $P_1 = C_s(U, L), P_2 = C_m(U, L)$ 
5:   for  $j = 1 \dots \frac{K}{n}$  do
6:     $m = a * |U|$   # 当前迭代的抽样数
7:     $\hat{x}_1 = \text{sort}(P_1(y = u|x))[0:m]$ 
8:     $\hat{x}_2 = \text{sort}(P_2(y = u|x))[0:m]$ 
9:     $\hat{x} = \hat{x}_1 \cap \hat{x}_2$ 
10:    $L = L \cup \hat{x}$ 
11:    $U = U \setminus \hat{x}$ 
12:  end for
13: end for
14: 返回  $L, U$ 
```

4.3 本章小结

本章依次介绍和分析了判别式主动学习的主要思想, 以及其与传统主动学习以及领域自适应的相似和不同之处, 然后分析了该算法还存在的一些问题。基于此, 下一章节将设计实验以验证改进方法是否有效。

5. 实证分析

为验证上一章节中提出的三个改进方向是否对判别式主动学习算法有提升作用，本章节在 MNIST 和 CIFAR10 数据集（为机器学习图片分类任务常使用的数据集）上设计图片识别的实验进行验证。5.1 节将分别对相关的实验标准、实验数据描述、实验过程和结果进行详细描述和分析；5.2 节展示实验数据结果和分析算法改进的有效性。

5.1 实验设计

本小结首先介绍实验中分析改进方法是否有效的标准，以及简要介绍使用的两个数据集，然后详细描述实验中的模型训练过程。

5.1.1 实验标准

在对主动学习算法进行实验验证时，目前学术界尚未确定统一的标准或方式。本文选取预测准确率和采样样本数（代表训练时间）作为算法优良性的衡量标准：一方面，在原始判别式算法与改进后的算法的预测准确率相等或相近时，后者若能降低所需的采样数，则说明改进机制可以减少算法所需的训练时间和人工标记的成本；另一方面，若在相近的采样数下后者的预测准确率更高，则说明改进的算法能够更有效地选择出信息价值高的样本，并且最终表现为模型可实现更高的准确率。

5.1.2 实验数据

本文在 MNIST 和 CIFAR10 数据集上基于图片分类问题验证算法改进机制的有效性。MNIST 数据集为来自美国国家标准与技术研究的手写数据集，训练集和测试集样本数分别为 50000 和 10000；CIFAR10 相对比 MNIST 内容更丰富，训练难度更高：该数据集由 10 个类的 32*32 彩色图像组成，每个类包含 6000 个图像。训练集和测试集总计分别包含 50000 和 10000 个图像。

5.1.3 实验过程

相关配置设定和预处理：考虑到本文在训练时的资源限制，即仅使用了单卡进行训练，程序将神经网络训练中的批处理量设定为 1，以避免内存和显存溢出；本文测试的模型均为包含四层隐含层的多层感知器，激活函数和 `relu` 函数，最后一层隐含层到输出层使用 `softmax` 得到每一类的预测概率值；考虑到 MNIST 和 CIFAR10 数据的处理难度，前者对应的训练迭代数(`epochs`)设定为 500，后者为 1000，均可成功收敛；实验使用的两个数据集均无缺失值，只需在进行模型训练之前将所有的数据标准化处理到(0,1)，并使用 32 位浮点数进行计算。

初始化阶段：从 MNIST 或 CIFAR10 的训练集中随机选取 20% 作为需要被采样的无标签数据，20% 作为验证集，而剩余部分作为训练集用以训练任务分类器并得到初始化的参数；在被选出的无标签数据中，再随机抽取一小部分作为被采样的数据，与剩余部分一起训练采样分类器得到其初始化参数。

迭代训练阶段：每经过多次全样本的训练后(`epochs`)，先暂时缓存当前模型的参数，并使用采样分类器和任务分类器同时计算当前未采样样本的概率值。采样分类器的样本选取标准为后验概率最大的一批样本点，而任务分类器则选择信息熵最大的样本点（由于本实验中的图像分类非二分类任务，因此不能直接使用预测概率作为采样标准），将两者选取的样本点取交集加入训练集。其中，采样大小在加入采样数自适应时根据当前未采样数据集的大小所决定，否则为固定值 $\frac{K}{n}$ （ K 为设定的总采样数， n 为迭代过程中的采样次数）。

5.2 实验结果与分析

三种改进机制在两个数据集下的实验结果见下图 5-1 和 5-2，具体采样数和测试集预测准确率的值见表 5-1 至 5-4。

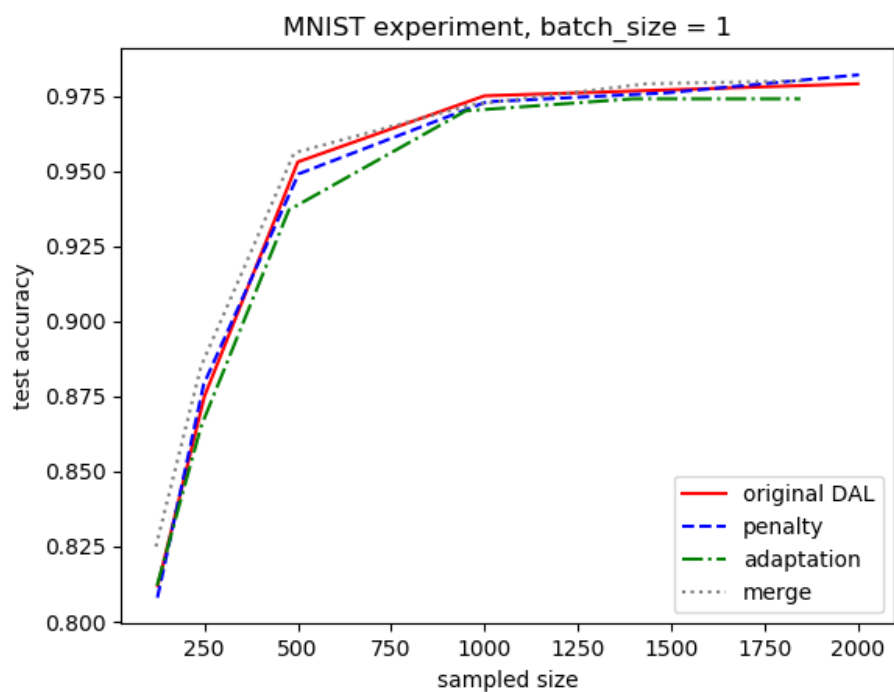


图 5-1 MNIST 实验结果

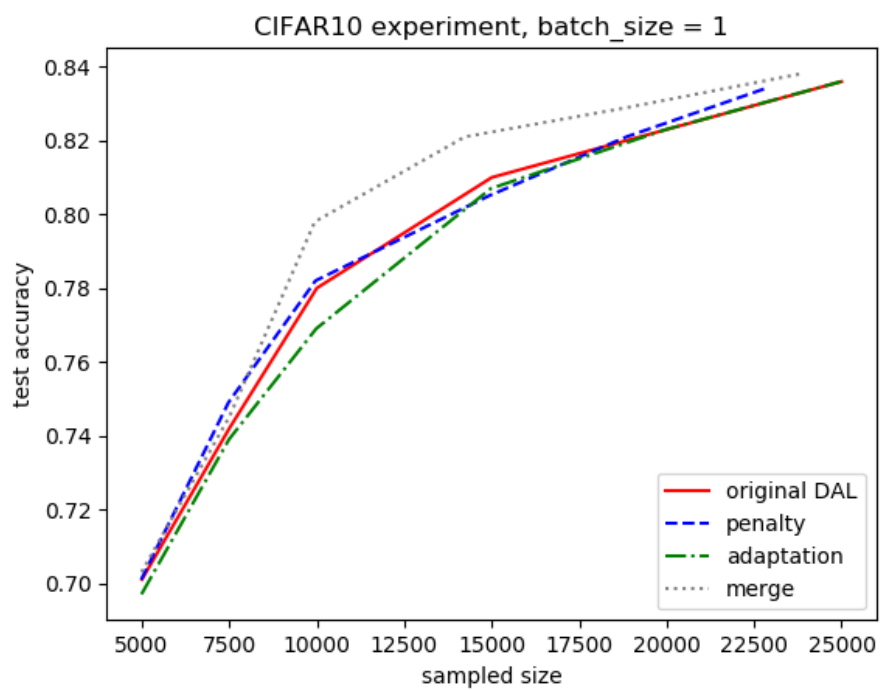


图 5-2 CIFAR10 实验结果

表 5-1 MNIST 采样数

| 算法 | 采样数 | | | | | |
|-------|-----|-----|-----|------|------|------|
| 原始算法 | 125 | 250 | 500 | 1000 | 1500 | 2000 |
| 惩罚项 | 124 | 246 | 502 | 1005 | 1489 | 2003 |
| 采样自适应 | 123 | 241 | 476 | 950 | 1398 | 1846 |
| 融合 | 120 | 241 | 490 | 965 | 1432 | 1856 |

表 5-2 CIFAR10 采样数

| 算法 | 采样数 | | | | | |
|-------|------|------|-------|-------|-------|-------|
| 原始算法 | 5000 | 7500 | 10000 | 15000 | 20000 | 25000 |
| 惩罚项 | 4987 | 7490 | 9982 | 14975 | 19978 | 24986 |
| 采样自适应 | 4981 | 7475 | 9962 | 14915 | 18846 | 22786 |
| 融合 | 4988 | 7493 | 9943 | 14230 | 18904 | 23785 |

表 5-3 MNIST 预测准确率

| 算法 | 预测准确率 | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| 原始算法 | 0.812 | 0.875 | 0.953 | 0.975 | 0.977 | 0.979 |
| 惩罚项 | 0.808 | 0.879 | 0.949 | 0.973 | 0.976 | 0.982 |
| 采样自适应 | 0.812 | 0.865 | 0.937 | 0.970 | 0.974 | 0.974 |
| 融合 | 0.825 | 0.885 | 0.956 | 0.972 | 0.979 | 0.980 |

表 5-4 CIFAR10 预测准确率

| 算法 | 预测准确率 | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| 原始算法 | 0.701 | 0.742 | 0.780 | 0.811 | 0.823 | 0.836 |
| 惩罚项 | 0.697 | 0.739 | 0.769 | 0.807 | 0.823 | 0.836 |
| 采样自适应 | 0.701 | 0.749 | 0.782 | 0.805 | 0.821 | 0.834 |
| 融合 | 0.703 | 0.745 | 0.798 | 0.821 | 0.829 | 0.838 |

上述实验结果表明如下三个结论：

1. 在两个数据集上，融合了两个标准的采样方法均有提升，且该机制对基准（benchmark）较低的 CIFAR10 数据集的提升更明显：当采样大小较小时，能够使用更少的样本获得相近的准确率，或是在采样数相近的情况下得到更高的预测准确率；

2. 采样自适应的方式作用不稳定：在 MNIST 数据集上起到了负作用，而在 CIFAR10 数据集上却能减少采样数和提高准确率，因此该方法的作用是不稳定的，可以认为对原算法没有显著的提升作用；

3. 加入惩罚项的作用不明显：在两个数据集上，加入惩罚项对应的曲线和原始算法对应的曲线几乎重合，因而该机制对采样数和预测准确率几乎无影响。

5.3 本章小结

本章节基于研究设计的三个改进方向在 MNIST 和 CIFAR10 数据集上进行实验验证，结果为融合了判别式采样和不确定性采样的改进方式有效而其他两个改进方式没有明显的提升效果。

6. 总结与展望

6.1 全文总结

本文首先对主动学习相关的主要研究工作进行了综述，表现为基于不确定性采样的算法在多个机器学习的应用领域中（也包括近年来的深度学习）都取得优良的效果；然后，本文详细介绍了主动学习从发展到现在的一些主要研究方向，在此基础上引入了使用另一种完全不同采样机制的判别式主动学习；最后，本文对该算法进行分析和改进，并设计实验进行验证。实验结果表明加入惩罚项作用不明显，采样自适应机制作用不稳定，而融合了不确定性采样和判别式采样的方法则对原始算法有显著提高作用。

6.2 未来展望

本文在对判别式主动算法分析改进并进行实验验证后，得出的结论为融合了判别式和不确定性采样的机制能在一定程度上减少所需的采样数和提升模型的预测准确率，但该机制仍然存在一定问题：仅简单地将任务分类器和采样分类器选择出的样本进行求交集，没有进行严谨的理论论证，可能在不同的任务领域中表现出不稳健的效果。因此，未来的研究可以使用领域自适应的理论基础来验证该方法的可行性和稳健性，并在更多的数据集和机器学习的其他任务上进行实验验证。

参考文献

- [1] 周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013, 39(11).
- [2] 杨文柱,田潇潇,王思乐. 主动学习算法研究进展[J]. 河北大学学报:自然科学版,
- [3] Seung H S, Oppen M, Sompolinsky H. Query by committee[C]. *Proceedings of the fifth annual workshop on Computational learning theory*, 1992: 287-294.
- [4] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]. *European conference on computational learning theory. Springer, Berlin, Heidelberg*, 1995: 23-37.
- [5] Dagan I, Elong S P. Committee-based sampling for training probabilistic classifiers[M]. *Machine Learning Proceedings*, 1995: 150-157.
- [6] Breiman L. Bagging predictors[J]. *Machine learning*, 1996, 24(2): 123-140.
- [7] Muslea I, Minton S, Knoblock C A. Selective sampling with redundant views[C]. *AAAI/IAAI*, 2000: 621-626.
- [8] Settles B. Active learning literature survey[J]. *Computer Sciences Technical Report*, 2014,1648.
- [9] Yang Y, Loog M. A benchmark and comparison of active learning for logistic regression[J]. *Pattern Recognition*, 2018, 83: 401-415.
- [10] Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data[C]. *Proceedings of the 34th International Conference on Machine Learning-Volume*, 2017: 1183-1192.
- [11] Ducoffe M, Precioso F. Adversarial active learning for deep networks: a margin based approach[J]. *arXiv preprint arXiv:1802.09841*, 2018.
- [12] Huang J, Child R, Rao V, et al. Active learning for speech recognition: the power of gradients[J]. *arXiv preprint arXiv:1612.03226*, 2016.
- [13] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach[J]. *arXiv preprint arXiv:1708.00489*, 2017.
- [14] Angluin D. Queries and concept learning[J]. *Machine learning*, 1988, 2(4): 319-342.

- [15] Baum E B, Lang K. Query learning can work poorly when a human oracle is used[C]. *International joint conference on neural networks*, 1992, 8: 8.
- [16] Dagan I, Engelson S P. Committee-based sampling for training probabilistic classifiers[M]. *Machine Learning Proceedings*, 1995: 150-157.
- [17] Yu H. SVM selective sampling for ranking with application to data retrieval[C]. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005: 354-363.
- [18] Lewis D D, Catlett J. Heterogeneous uncertainty sampling for supervised learning[M]. *Machine learning proceedings* 1994. Morgan Kaufmann, 1994: 148-156.
- [19] Gissin D, Shalev-Shwartz S. Discriminative active learning[J]. *arXiv preprint arXiv:1907.06347*, 2019.

致 谢

论文完成之际，短短四年的本科生生活也即将结束，回想在西南财经大学校园成长的点滴，心中充满无限感激与留念。

感谢我的导师在我论文完成过程中提供的宝贵建议和帮助，让我顺利完成了这篇论文。感谢西南财经大学统计学院各位老师的帮助和关心，让我一步步从统计的数学基础走到数学与统计的连接，并实现将统计作为工具进行分析，现在的我懂得并且对用统计的思想解决实际问题产生了兴趣。

感谢西南财经大学同学对我的帮助，让我在四年的大学生活中顺利度过一次次困难，也拥有无数快乐而难忘的经历，感谢你们与我一起成长，一起面对学习和生活的酸甜苦辣。

感谢我的家人和朋友们默默的支持，是他们给予我莫大的精神鼓励。

