分类号 _____ 密级 公开

UDC _____

# 学 位 论 文

## 增广信息学习

（题名和副题名）

朱 越

（作者姓名）

指导教师姓名、职务、职称、学位、单位名称及地址 周志华 教授

南京大学计算机科学与技术系 南京市栖霞区仙林大道 163 号 210023

申请学位级别 博士 专业名称 计算机科学与技术

论文提交日期 2018 年 6 月 20 日 论文答辩日期 2018 年 7 月 25 日

学位授予单位和日期 _____

答辩委员会主席： 陈松灿 教授

评阅人： 杨 明 教授

孙权森 教授

姜 远 教授

戴新宇 教授


南京大學

# 研 究 生 毕 业 论 文

# （申请博士学位）

论 文 题 目 增广信息学习

作 者 姓 名 朱 越

学科、专业方向 计算机科学与技术

指 导 教 师 周志华 教授

研 究 方 向 机器学习与数据挖掘

**2018** 年 **7** 月 **29** 日

学　　号：**DG1333047**

论文答辩日期：**2018** 年 **7** 月 **25** 日

指 导 教 师：　　（签字）

# Learning with Augmented Information

by

**ZHU Yue**

Supervised by

Professor ZHOU Zhi-Hua

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
Computer Science and Technology


Department of Computer Science and Technology
Nanjing University

July 29th, 2018

# 南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：增广信息学习

计算机科学与技术 专业 2013 级博士生姓名：朱 越

指导教师（姓名、职称）：周志华 教授

## 摘 要

传统监督学习通常假设训练数据类别标记恒定、特征信息充分、样本充足。但很多现实的机器学习任务不满足这些假设条件，导致学习效果不尽人意。为此，本文考虑通过引入增广信息 (Augmented Information) 进行学习。增广信息包括传统静态学习中未考虑的额外信息以及动态学习过程中出现的新信息。本文主要工作如下：

1. **提出了一种训练集标记增广学习方法 GLOCAL。**该方法利用标记关系对多标记训练数据中部分缺失的标记进行恢复补全，但无需额外的先验知识来指定标记关系矩阵，而是在优化过程中同时习得全局和局部标记关系。实验验证了本文方法的有效性。
2. **提出了分别用于静态、动态测试集标记增广学习的方法 DMNL 和 MuENL。**DMNL 通过最小化多示例包级损失和聚类正则化项，预测静态测试集中的多个新标记；MuENL 通过特征和预测值训练新标记检测器并建立鲁棒模型，以检测动态新增的标记并对其建模。实验验证了本文方法的有效性。
3. **提出了一种多示例特征增广学习方法 AMIV-lss。**针对数据特征信息不足的学习问题，将额外获取的带噪信息形式化为增广多示例视图 (AMIV) 作为样本的特征增广。AMIV-lss 通过在两个异构视图之间建立公共隐藏语义子空间，减少噪声影响，提升学习性能。实验验证了本文方法的有效性。
4. **提出了一种多视图样本增广学习方法 OPMV。**OPMV 通过对每个样本优化视图一致性约束下的组合目标函数，即可随着新增多视图数据高效更新模型，并能够利用视图之间结构提升学习性能，而无需存储整个数据集，避免从头进行训练。理论和实验验证了本文方法的有效性和高效性。
5. **提出了同时进行标记/特征/样本增广学习的方法 EM3NL。**EM3NL 基于多视图多示例多标记深度卷积神经网络，利用额外文本描述作为图片的补充 (特

征增广)，能够检测新标记 (标记增广)，并可根据动态增加的样本即时更新模型 (样本增广)。实验验证了本文方法的有效性。

**关键词：** 机器学习；增广学习；增广信息；多标记学习；新标记学习；多视图学习；单趟学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Learning with Augmented Information
SPECIALIZATION: Computer Science and Technology
POSTGRADUATE: ZHU Yue
MENTOR: Professor ZHOU Zhi-Hua


## Abstract

Most previous studies on traditional supervised learning usually assume fixed target class labels, sufficient discriminative features, and enough training instances. In real applications, however, those assumptions could not be satisfied. As a result, it is necessary to study learning with augmented information problem, where the augmented information includes additional information that is not considered by traditional static learning paradigms, and new information that is introduced by a dynamic learning procedure. The main studies of this paper are listed as follows:

1. **Propose `GLOCAL` approach to handle training label augmentation problem**. In order to learn with partially observed labels, `GLOCAL` recovers those missing labels (i.e. training label set augmentation), via taking advantage of the low-rank structure of label matrix, exploring and exploiting global and local label correlations, where label correlations are learned during the optimization without manual specification. Experimental results validate the effectiveness of the proposed approach.
2. **Propose `DMNL` and `MuENL` for static and dynamic testing label augmentation problems respectively**. `DMNL` addresses the static testing label augmentation learning via minimizing a bag-level loss with a clustering regularization. `MuENL` detects new labels based on the combinations of features and predictive values, and builds robust models, so as to solve dynamic testing label augmentation problem. Experimental results validate the effectiveness of the proposed approaches.
3. **Propose `AMIV-lss` approach to deal with multi-instance feature augmentation problem**. `AMIV-lss` formulates the augmented feature information as an additional multi-instance view, and establishes a common latent semantic subspace between the two heterogeneous views (one single-instance view and one multi-

instance view), so as to improve the performance. Experimental results validate the effectiveness of the proposed approach.

4. **Propose `OPMV` approach for multi-view instance augmentation problem**. `OPMV` takes advantage of multi-view structure, which forces the prediction on different views of the same object to be similar, updates the model efficiently according to each newly arrived multi-view instance, so as to handle multi-view instance augmentation problem. Both the theoretical and experimental results validate the effectiveness and efficiency of the proposed approach.
5. **Propose `EM3NL` approach for simultaneously learning with augmented labels, features, and instances**. `EM3NL` is based on a deep multi-instance multi-label convolution neuron network. It takes advantage of augmented text descriptions for images (augmented features), detects whether an instance holds a new label(augmented labels), and handles mini-batch model updates (augmented instances). The experimental results validate the effectiveness of the proposed approach.

**keywords:** Machine learning, learning with augmented information, augmented information, multi-label learning, learning with new labels, multi-view learning, one-pass learning

# 目　　次

目　　次 …… v

**1　绪论** …… **1**

1.1　引言 …… 1

1.2　研究现状 …… 4

1.3　有待研究的问题 …… 6

1.4　本文工作 …… 7

**2　训练集标记增广学习** …… **9**

2.1　引言 …… 9

2.2　本文方法 …… 11

2.3　实验测试 …… 16

2.4　本章小结 …… 25

**3　测试集标记增广学习** …… **27**

3.1　引言 …… 27

3.2　本文方法 …… 30

3.3　实验测试 …… 43

3.4　本章小结 …… 55

**4　特征增广学习** …… **57**

4.1　引言 …… 57

4.2　本文方法 …… 58

4.3　实验测试 …… 63

4.4　本章小结 …… 66

**5　样本增广学习** …… **67**

5.1　引言 …… 67

5.2　本文方法 …… 68

5.3 实验测试 …… 74
5.4 本章小结 …… 77

**6 综合增广学习 …… 79**
6.1 引言 …… 79
6.2 本文方法 …… 80
6.3 实验测试 …… 85
6.4 本章小结 …… 88

**7 结束语 …… 89**

**参考文献 …… 91**

**致 谢 …… 103**

**A 攻读博士学位期间的学术成果和获奖情况 …… 105**

# 第一章　绪论

## 1.1　引言

机器学习是人工智能领域的一个重要分支，通过对数据进行处理和分析，使得算法可以利用经验自动改善自身性能。它的主要研究内容是从训练数据中产生模型的算法，使其在测试数据上取得满意的预测性能。学习任务大致分为两大类：如果训练数据有标记信息则为监督学习，否则为无监督学习。本工作主要关注监督学习。

在传统监督学习中，给定包含 $n$ 个有标记样本的训练集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}^n$，其中 $\boldsymbol{x}_i\in\mathcal{X}$ 是 $d$ 维样本空间 $\mathcal{X}$ 上的一个样本，$\boldsymbol{y}_i\in\mathcal{Y}$ 是 $\boldsymbol{x}_i$ 对应的标记。传统监督学习的任务是学到一个从样本空间到标记空间的映射函数 $f:\mathcal{X}\to\mathcal{Y}$，通过 $f(\boldsymbol{x}_{\text{new}})$ 预测训练集中未见过的样本 $\boldsymbol{x}_{\text{new}}$ 的标记。在传统监督学习中，一般假定类别标记恒定且测试集和训练集的标记分布一致、特征信息充分、样本充足且数量固定。

在实际应用中，上述假设经常不能被满足。实际任务中可能遇到测试集的类别标记分布和训练集的类别标记分布不一致的情况。例如多标记分类任务，由于标注者的知识背景、关注点不同，可能造成训练样本上的标记部分缺失。在这种有缺失标记的情况，尽管测试集与训练集的标记集合一样，但是观察到的标记分布与真实标记分布不一致，如果直接应用传统的多标记学习方法效果将不甚理想。进一步地，测试集中还可能存在训练集上从未观察到的新标记，如果不考虑这些新标记，那么在测试过程中，新标记会被当成某个已知标记而无法被预测出，造成学习模型的整体预测性能不尽人意。除了测试集、训练集标记分布不一致，还可能出现训练数据特征判别信息不足的情况。例如在短文本主题分类任务中，由于文本非常简短，甚至可能不包含和主题直接相关的词汇，这将导致短文本相应特征的判别信息不足。如果学习算法仅仅利用这些已有的判别信息较弱的特征进行学习，将很难得到令人满意的主题分类模型。此外，学习任务也常常会面临样本动态增加的场景。例如在 Twitter 上，平均每秒钟用户会发布 5700 条消息；在 YouTube 上，平均每分钟会有总时长达上百小

时的视频被上传；在 Flickr 上，平均每天会产生 168 万张新的照片。对于这些新数据产生非常频繁的场景，如果应用传统监督学习方法，需要经常重新训练模型，存储、计算开销巨大，非常低效。

针对上述传统监督学习难以处理的问题，本文考虑引入增广信息，提出了增广信息学习。增广信息包括传统静态学习方法中未考虑的信息以及动态过程中出现的新信息，例如新的类别标记、额外的特征、动态增加的样本等。从增广信息的类型进行划分，增广信息学习包括：

(1) 训练集标记增广学习，利用样本结构和标记关系补全训练集缺失标记；
(2) 测试集标记增广学习，除已知标记外，还要预测测试集上的新标记；
(3) 特征增广学习，利用额外获取的特征信息提升学习器性能；
(4) 样本增广学习，根据学习过程中动态增加的新样本即时更新模型；
(5) 综合增广学习，同时进行多种形式的增广学习。

其中，训练集标记增广学习利用样本结构和标记关系对缺失标记进行补全，尽可能使训练集上的标记分布与真实标记分布一致，从而训练的模型能够在测试时取得更好的泛化性能；测试集标记增广学习在检测新标记并对其建模的同时，进一步对已知标记模型进行约束，从而取得更好的整体性能；特征增广学习利用额外的特征信息作为原始特征的补充，从而提升学习器性能；样本增广学习针对学习过程中动态增加的新样本即时更新模型，避免从头训练，将大大降低存储和计算开销，提升学习算法效率。综上所述，在学习过程中考虑增广信息可以提升整体学习性能、为复杂环境中的学习（如标记缺失、出现新标记、动态增加新样本等）提供解决方案，并能够高效地进行模型更新。

与增广信息学习最为相关的学习范式包括增量学习 (Incremental Learning) [1–12] 与辅助信息学习 (Previleged Learning) [13–16]。增量学习被广泛应用于那些经常有数据更新的动态学习场景，而无需从头重新训练模型。根据文献 [1]，增量学习的常见设置基本分为三种：即属性增量学习 (A-IL)[2, 3]；类增量学习 (C-IL)[4–6]；以及样本增量学习 (E-IL)[7]。A-IL 主要研究在学习系统训练好之后，如何整合新的特征输入。例如在生物环境研究中，分批投放传感器，传感器收集的数据特征会阶段性增多，A-IL 研究利用这些不断增多的特征更新模型，以避免重新采集数据，从头训练一个新模型。文献 [2] 研究的是有特征演变的数据流学习，其中，随着数据流，旧的特征会消失而新的特征会出现。为了解决这个问题，文献 [2] 恢复消失特征，并在恢复的特征以及现有的特征上训练模型，并通过结合两个模型提升学习性能。C-IL 主要研究学习系统在训练

好之后如何预测测试集中的新类别标记，要求学习系统在考虑新标记的同时尽量不牺牲在已知标记上的性能，且要避免重新训练模型。文献 [4] 通过为每个新类和已知类建立二分类器将已经存在的类别和新类别区分开来，从而将类别标记数从 $n$ 扩展到 $n+1$；文献 [5] 基于最大化间隔原理利用未标记数据使得分类超平面穿过低密度区域，从而识别新类别；文献 [8] 基于完全随机树提出了在数据流上检测新类并对其进行建模的统一框架。E-IL 主要研究当出现新增样本时如何更新模型，而不是从头开始训练模型。在线学习是一种非常重要的处理增量样本的学习范式，能够对新出现的样本及时调整现有模型[9–12]。

增量学习，无论是 A-IL、C-IL、E-IL，都属于特殊的增广信息学习，分别考虑动态过程中新增的特征、标记和样本，且要求增广的形式是一部分一部分逐渐增量。而增广信息学习并不强调增广的形式是增量增广，它不仅包括了动态学习过程中的新增信息，也考虑了静态学习问题中传统方法未考虑的信息。例如特征增广学习中利用的额外特征可以一开始就能获得；在训练集标记增广学习任务中，增广标记是对训练集观察标记矩阵中缺失标记的补全（测试集的标记集合没有增广）。这些静态的增广学习问题是增量学习所没有涉及的。对于动态过程中的测试集标记增广学习问题，传统的 C-IL 主要针对多类别学习，而增广信息学习还考虑了更为复杂的多标记学习设置下的标记增广：在多类别设置下，新标记样本不属于任何已知类，从特征空间就能够将其与已知标记样本分开；而在多标记设置下，新标记样本也可能同时与一些已知标记关联，使得很难将同时有新标记和已知标记的样本与那些没有新标记但是有相同已知标记的样本区分开来。

辅助信息学习由 Vapnnik 提出 [13]，利用额外的特征信息提升学习性能。近些年来，该学习范式引起了大量关注：文献 [14] 证明使用辅助信息可以将学习算法收敛率从 $O(1/\sqrt{n})$ 提升到 $O(1/n)$；文献 [15] 将辅助信息学习从二分类问题推广至排序学习问题；文献 [16] 研究了多标记设置下的辅助信息学习等等。一般地，在辅助信息学习中，给定有标记训练样本 $\boldsymbol{D}=\{(\boldsymbol{x}_i,\boldsymbol{x}_i^*,y_i)\}^n$，其中 $\boldsymbol{x}_i\in\mathcal{X}$ 是原始样本特征，$\boldsymbol{x}_i^*\in\mathcal{X}^*$ 是辅助信息特征，目标是学习一个从原始样本特征空间到标记空间的映射函数：$f:\mathcal{X}\xrightarrow{\mathcal{X}^*}\mathcal{Y}$。从形式上，辅助信息 $\boldsymbol{x}_i^*$ 与原始样本 $\boldsymbol{x}_i$ 一一对应，且只在训练集里可以获得，而测试阶段不可用。从方法上来说，这些辅助信息并不参与决策，而是提供了额外的约束信息。

相较于辅助信息学习，增广信息学习中的增广信息包含的内容更加广泛：它不限于额外获取的增广特征，还包括学习过程中出现的新标记、动态增加的

样本等，这些增广信息种类是辅助信息学习所未涉及的。即便在形式上与辅助信息学习类似的特征增广学习中，额外获得的增广特征向量与原始特征向量的对应关系也不限于一对一，也可以多对一或者多对多；且这些增广的特征可以直接参与学习和决策过程 (训练和预测时都可以利用这些特征)，其作用不限于提供额外的约束信息。

## 1.2 研究现状

在现实应用中，一个样本可能与多个类别标记相关联：例如一张图片可能有多个标注，一首歌曲可能同时属于多个流派，一篇文章可能与多个主题相关联。多标记学习就是研究这种数据的学习范式。很多时候由于标注者的背景知识、关注点不同，造成样本上部分标记缺失，需要对这些缺失标记进行补全，即训练集标记增广。在现有研究中，它又被称为多标记弱标记或缺失标记学习。文献 [17] 假设相似样本上的标记预测应该相似，利用样本相似性进行标记传播，从而对缺失的标记进行补全。文献 [18] 提出了一种快速的低秩矩阵补全方法，且有着很好的理论保证。LEML [19] 是一种高效的基于低秩结构的处理缺失标记的方法，将特征标记映射矩阵分解成两个小矩阵乘积。这些方法或者考虑样本之间的相似性结构 [17]，或者利用标记关系带来的低秩结构 [18, 19]。如果能够显式地探索、利用标记之间的关系，可能会取得更好的结果。

测试集标记增广源于测试集中包含的新标记，需要在测试时对它们作出预测。这些新标记的相关语义可能在给定训练集上已经存在，但是从未被标记出，即静态测试集标记增广问题；也可能是动态过程中数据流上新增样本带来的前所未有的新语义，即动态测试集标记增广问题。对于前者，真实标记集合等于训练集中所有观察到的标记集合与训练集隐藏的新标记集合的并集。虽然同样是存在语义未被标注，在训练集标记增广学习中，训练集标记集合等于测试集标记集合，通过低秩方法可以进行补全 [18, 19]。但是对于这种某些语义完全未被标注的测试集标记增广问题，相当于真实标记矩阵整列缺失，则无法利用低秩结构进行恢复。尽管同样是新增类别标记，类增量学习 [4, 5] 主要是在单示例多类别学习框架下为新类建立模型，即每个对象只属于多个类别中的一个。它们一般只考虑只有一个新类的情况，且这个新类的语义在训练集中并未出现过。因而无法通过这些方法处理静态测试集标记增广问题。

对于数据流中新增样本带来的新标记，真实标记集合不断增广。且在多标记设置下，新增的样本可能同时出现新标记和已知标记。目前在数据流上处理

新类的方法 SENCForest [8] 主要针对多类别学习，即新样本要么属于某个已知类别要么属于一个新类别，无法直接处理多标记设置下的动态测试集标记增广问题。一个简单的策略可以使得它能够处理多标记数据流中出现的新语义，即将每一种可能的标记组合当成是一个独立的类别，从而把多标记问题转化为多类学习问题 [20]。但是这种方法有两个非常严重的问题：(1) 检测到的新类可能并不意味着一个新的标记，而可能是一些在训练集上没有出现过的已知标记的组合；(2) 当标记集合比较大时，将导致可能的标记组合数（即转化为多类学习后的类别数）非常大，有可能一些类只有非常少量的相关样本，使得训练任务变得非常困难。也因此，这种简单的转换在实际问题中并不可行。当把检测新标记问题当做是异常检测问题来处理时，即把异常样本当做是新标记样本，可以利用很多现有的异常检测方法。例如 OC-SVM [21] 为每个已知类训练一个边界，把落到边界外的样本当做是异常样本；iForest [22] 则把那些落到密度稀疏区域的样本当做是异常样本。但在多标记设置下，新标记和已知标记可能同时出现在一个新样本上，使得很难仅通过区分特征，将同时有新标记和已知标记的样本与那些没有新标记但是有相同已知标记的样本区分开来。

许多研究工作期望在学习过程中利用增广特征提高性能，例如多视图学习、利用辅助信息学习。多视图学习用于处理由多个视图描述的数据（如图像和描述图像的文本等），即多个特征集合，不同的视图可以相互看做增广信息源。这里，每个视图是一个特征集合，每个样本在每个视图上都有相应的特征向量示。多视图学习的目标是通过利用多个视图之间的关系提升学习性能或是降低样本复杂度。尤其在利用未标记示例学习问题上，多视图学习被广泛研究。例如最为著名的多视图学习方法 co-training [23]，就是一种半监督学习算法。随后，Wang [24, 25] 证明以 co-training 为首的这类基于分歧的方法其实并不一定要求多个视图具有真实物理意义，只要在多个视图上分类器的多样性足够充分。2007，Zhou [26] 给出了更令人振奋的结果：在合适的多个视图上，仅仅需要一个标记样本，半监督学习就可以取得成功。由此可见，多个视图确实能够对学习提供很大帮助。除此之外，利用多个视图，主动学习在不可分情况下的样本复杂度将会指数级改进 [27]，而且多视图学习可以天然将主动学习和半监督学习结合在一起 [28]。值得指出的是，许多多视图学习通过建立隐藏子空间，使得同一样本对应的多个视图上的特征表示映射到该子空间后距离相近 [29–33]。与多视图学习类似，辅助信息学习包括两个视图：原始视图和辅助信息视图。但是辅助信息只能用于训练，并为优化问题提供额外的约束信息 [13–16]。这些利

用额外特征进行学习的方法中，增广特征向量与原始样本特征向量一一对应，且不考虑增广信息中可能引入噪声。

在动态学习过程中，往往会遇到样本动态增加的情况。尤其是在数据表现形式丰富的互联网时代，多视图数据时刻产生，这要求模型能够根据动态增加的数据高效地进行更新。现有的多视图学习 [23–33] 和辅助信息学习方法 [13–16] 主要是批学习方法，需要存储所有数据，反复扫描，并不能根据动态新增的多视图样本即时更新模型。

## 1.3 有待研究的问题

如第1.1节所述，增广信息学习包括训练集标记增广学习、测试集标记增广学习、特征增广学习、样本增广学习以及综合增广学习。下文将针对每一类增广学习，阐述其亟待解决的问题。

对于训练集标记增广学习，目标为恢复训练集上部分缺失的标记并预测新样本上的标记。现有工作主要利用低秩结构对有缺失标记的标记矩阵进行恢复，这可以看作是隐式地利用了全局标记关系。在很多实际应用中，全局标记关系和局部标记关系可能同时存在。合理利用标记关系对学习性能有着至关重要的作用，是多标记学习的一个核心。但是标记关系往往很难由人指定，这在标记部分缺失、全局和局部标记关系同时存在且各不相同的情况下尤为困难。因此，如何显式地探索并利用全局和局部标记关系进行缺失标记恢复以及测试集标记预测是训练集标记增广学习中一个亟待解决的问题。

测试集标记增广的新标记来自于数据中的新语义，要求学习系统能够预测测试集样本中是否包含这些新标记。新标记对应的语义可能存在于训练集中，但它们在整个训练集中都未被标出来，对应真实标记矩阵整列缺失。这种情况下，无法利用低秩结构对标记矩阵进行恢复。因此，如何在测试集上对多个这种新标记进行预测是一个非常重要而未解决的问题。除此之外，数据流上动态增加的样本可能会带来新标记。由于在多标记的设置下，新增样本中的新标记可能会和已知标记同时出现，难以用现有的异常检测方法进行检测。因此，如何在多标记设置下预测数据流上出现的新标记具有非常重大的研究价值。

对于特征增广学习，引入额外的信息作为增广特征对原始特征进行补充，以提升学习性能。但在实际应用中，这些引入的额外信息可能也包括噪声。例如基于摘要信息的科学文献主题分类任务中，原始特征信息是原文的摘要，考虑利用其参考文献对应的摘要作为增广特征信息。但是并非所有的参考文献的

主题都与分类任务相关，即包含着噪声信息。如果直接将所有参考文献的摘要信息拼在一起作为增广特征，那么它们所引入的噪声可能反而导致学习性能下降。因此，在特征增广学习中，如何利用这种带噪的增广特征进行学习以提升学习性能是一个非常重要的问题。

对于样本增广学习，现有的解决方案主要针对单视图数据。而很多应用场景都牵涉到大规模多视图数据，且样本会不断动态增加。如 YouTube 上，每分钟都会有总计上百小时时长的视频被上传（包括图像、声音、文本等多个视图）。传统多视图学习研究证实利用多视图之间的关系可以提升学习性能。但是它们往往需要反复扫描数据集，具有较高的复杂度，且无法根据新增的样本进行高效更新。如何根据动态增广的多视图样本，利用多个视图之间的关系提升性能，并对模型进行高效、即时更新，避免从头训练，从而尽可能减小计算、存储开销，是样本增广学习中极具研究价值的问题。

本文前期的研究多针对训练集标记增广学习、测试集标记增广学习、特征增广学习、样本增广学习中的某一种进行展开。还有一些情况下，多种不同形式的增广信息可能会同时出现：例如预测测试样本上的新标记（标记增广）问题，可以进一步从环境中获取额外的信息作为增广视图（特征增广）以提升学习器性能，同时要求模型能够根据新出现的多视图样本即时更新并能利用视图之间的关系（样本增广）。因此，如何解决这类同时出现标记增广、特征增广和样本增广的综合增广学习问题，为其设计一个统一的解决方案，是一个具有极高应用价值的研究问题。

## 1.4 本文工作

针对上述问题，本文主要完成了五项工作，分别对应于第二章至第六章。

在第二章中，针对训练集标记增广问题，即补全训练集样本部分缺失的标记，提出了一种多标记关系学习算法 GLOCAL，它能够通过学习隐标记表示及优化标记流形，同时恢复缺失标记，训练分类器，探索与利用全局和局部标记关系。与之前的工作相较，它同时利用全局和局部标记关系，且直接通过数据学习标记拉普拉斯矩阵，而不需要其它关于标记关系的先验知识。除此之外，GLOCAL 为标记完整情况和标记缺失情况的多标记学习提供了一个统一的解决方案。实验结果表明了 GLOCAL 利用全局和局部标记关系及标记矩阵低秩结构增广训练集标记的有效性。

在第三章中，主要研究了两种测试集标记增广学习：静态测试集标记增

广学习（测试集中的新标记对应的语义在训练集出现而从未被标注）和动态测试集标记增广学习（新标记随着数据流动态增加），分别提出了 DMNL 和 MuENL 方法。DMNL 基于多示例多标记框架，将静态测试集标记增广学习问题形式化成一个非负正交约束下的优化问题，通过优化示例包损失项和聚类正则化项，使得已知标记和新标记可以被同时建模。MuENL 同时利用特征与标记预测值构造新标记检测器，并为新标记设计了鲁棒的更新模型。实验结果分别验证了 DMNL 和 MuENL 方法处理静态和动态测试集增广学习问题的有效性。

在第四章中，针对利用带噪增广特征进行学习的问题，将其形式化成增广多示例视图学习问题 (AMIV)。对于基于摘要的科学文献分类应用，将原文摘要作为原始视图，并从网上获取其参考文献的摘要作为增广多示例视图（每篇文章的参考文献组成的多示例包中至少有一个示例与原文主题相关），利用多示例学习框架减少噪声对学习器的影响。提出了 AMIV-lss 方法，通过在单示例视图和多示例视图之间建立公共隐藏语义子空间，从而利用视图之间的结构关系提升学习性能。实验结果表明了 AMIV-lss 方法利用带噪增广特征提升学习性能的有效性。

在第五章中，为了根据动态增广的多视图数据高效、即时更新模型，针对多视图样本增广学习问题，提出了单趟多视图学习 (OPMV) 方法。该方法基于多视图一致性约束下的组合目标函数优化，要求同一对象在不同视图上的预测值相同，以此利用视图之间的结构关系提升学习性能。OPMV 方法的优化在每轮迭代中只涉及一个多视图样本，因此无需将所有数据载入内存，即可根据动态增广的新样本对模型进行高效更新，避免从头训练模型。理论分析证明了 OPMV 方法的收敛速率为 $O(1/\sqrt{T})$，实验结果验证了 OPMV 处理多视图样本增广学习问题的有效性和高效性。

在第六章中，针对同时进行标记、特征、样本增广学习的综合增广学习问题，提出了 EM3NL 方法。EM3NL 将深度卷积神经网络与多视图、多示例多标记学习相结合，设计了一种端到端统一解决方案。它利用增广文本视图（特征增广）提升学习性能，直接输入原始图像和相应文本描述，通过深度卷积网络提取特征表示；在多示例学习框架下优化包级的误分损失和新标记排序关系正则化项，同时对已知标记和新标记进行建模（标记增广）；通过小批量优化方法能够根据新增样本对模型进行高效更新（样本增广）。实验结果验证了 EM3NL 方法处理这种综合增广学习问题的有效性。

# 第二章　训练集标记增广学习

## 2.1　引言

由于人类标记者常常不会标记样本中那些他们不感兴趣、不了解的语义，或者是根据某些算法指导进行标记以减少整体的标记代价 [34, 35]，造成训练数据的标记部分缺失。在训练集中部分样本上缺失的标记，在另一些训练样本上能够观测到，所有训练样本标记的并集等于真实的标记集合。尽管这种情况下训练集标记集合与测试集标记集合相同，但是观察到的标记分布由于部分标记缺失与真实标记分布不同，直接用观察到的标记训练模型，效果将不甚理想。因此需要考虑对缺失标记进行补全，即训练集标记增广。

本章的研究延袭了一种常用设置，即正标记和负标记均可能缺失 [18, 36]；令 $\mathcal{X}$ 表示特征空间，$\mathcal{Y} = \{-1, 1\}^l$ 表示标记空间。真实的标记向量表示为 $\tilde{\boldsymbol{y}} \in \mathcal{Y}$，观察到的标记向量表示为 $\boldsymbol{y}$，其中 $[\boldsymbol{y}]_j = 0$ 表示第 $j$ 个标记未被标记出，否则 $[\boldsymbol{y}]_j = [\tilde{\boldsymbol{y}}]_j$ 与真实标记相等；给定 $n$ 个训练样本 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$，目标是学一个从特征空间到标记空间的映射 $\Psi : \mathcal{X} \to \mathcal{Y}$。由于标记关系的存在，标记矩阵是有着低秩结构的。对于部分标记缺失的多标记训练集增广学习问题，利用这种低秩结构，可以对标记矩阵进行矩阵补全 [18, 37]。一种常用的矩阵补全方法是优化核范数正则化项 [38, 39]，但是核范数优化一般计算开销会比较大[19]。一个更直接的方法是利用矩阵分解，用两个较小矩阵的乘积逼近观察到的标记矩阵 [40, 41]。令 $\tilde{\boldsymbol{Y}} = [\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n] \in \{-1, 1\}^{l \times n}$ 表示真实标记矩阵。令 $\tilde{\boldsymbol{Y}}$ 的秩为 $k < l$，则它可以写作两个小矩阵的乘积

$$\tilde{\boldsymbol{Y}} \simeq \boldsymbol{U}\boldsymbol{V}, \tag{2.1}$$

其中，$\boldsymbol{U} \in \mathbb{R}^{l \times k}$，$\boldsymbol{V} \in \mathbb{R}^{k \times n}$。直观上看，$\boldsymbol{V}$ 表示隐标记，刻画了更高阶的概念。这种表示比原标记空间的标记更为紧致，语义上更为抽象；而 $\boldsymbol{U}$ 则反映了原空间的标记与隐标记关联的程度。更一般地，低秩建模在矩阵补全中起到至关重要的作用，且天然可以解决标记部分缺失问题。令观察到的标记矩阵为 $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \{-1, 0, 1\}^{l \times n}$，$\Omega$ 表示观察到的标记在标记矩阵的坐标集合，即 $\boldsymbol{Y}$ 中非 0 元素的坐标集合。假设观察标记矩阵 $\boldsymbol{Y}$ 与真实标记矩阵 $\tilde{\boldsymbol{Y}}$ 在观察到

的标记上一致 (即不考虑标记噪声)，且 $\tilde{\boldsymbol{Y}}$ 能由 $\boldsymbol{UV}$ 表示。最小化在观察标记上的重构误差，即最小化 $\|\Pi_{\Omega}(\boldsymbol{Y}-\boldsymbol{UV})\|_F^2$。如果 $(i,j)\in\Omega$ 则 $[\Pi_{\Omega}(\boldsymbol{A})]_{i,j}=A_{i,j}$，否则为 0。标记完整的多标记学习任务 (即标记矩阵的所有元素均被观察到) 可以看成一种特殊情况：$\boldsymbol{Y}=\tilde{\boldsymbol{Y}}$，则 $\Pi_{\Omega}(\boldsymbol{Y}-\boldsymbol{UV})=\tilde{\boldsymbol{Y}}-\boldsymbol{UV}$。在得到 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 后，标记矩阵 $\boldsymbol{Y}$ 中 $(i,j)\notin\Omega$ 处的缺失标记可由 $\text{sign}(\boldsymbol{u}_{i,:}\boldsymbol{v}_{:,j})$ 恢复。其中，$\boldsymbol{u}_{i,:}$ 是 $\boldsymbol{U}$ 的第 $i$ 行，$\boldsymbol{v}_{:,j}$ 是 $\boldsymbol{V}$ 的第 $j$ 列。

尽管利用低秩结构进行补全可以认为是隐式地利用了全局标记关系，但若直接显式地利用标记关系则可以更好地恢复缺失的标记 [18]。如何利用标记之间的关系，也是多标记学习的核心问题之一。这些标记关系能够提供很多重要信息。比如如果 “主题公园” 和 “米老鼠” 这两个标记同时出现，那么标记 “迪士尼” 也非常有可能出现；类似地，如果 “蓝天” 和 “白云” 同时出现，那么 “大雾” 则不会出现。现有的工作分别从不同的层面考虑标记之间的关系 [42]：(1) 一阶关系，不考虑任何标记关系，而把多标记学习转变成多个独立的二分类问题，例如 BR [43]，它为每个标记分别独立训练一个分类器；(2) 二阶关系，考虑两两标记对之间的关系，例如 CLR [44] 把多标记学习问题转换成一个逐对标记排序优化问题；(3) 高阶关系，所有的标记关系都会被考虑，例如 CC [45]，把多标记学习问题转化为一个二分类问题串联，并依次将真实标记编码到特征空间。另一种利用全部标记关系的方法是通过学习一个隐标记空间描述更高层的标记语义。这可以由标记矩阵分解的方法得到 [46]。类似地，Jing 等人 [47] 利用字典学习得到标记嵌入；Yeh 等人 [48] 提出了一种深度学习方法来联合学习特征和标记嵌入。这些工作和典型相关分析 (CCA) 非常相关，都是学一个样本和标记的公共同子空间表示。上述工作主要考虑全局标记关系，并假设这些标记关系在所有样本上都适用。但是事实上，有些类别标记关系仅适用于部分样本 [49, 50]。例如在美食杂志上，“苹果” 和 “水果” 相关，但是在科技杂志上，“苹果” 更多地和 “电子设备” 联系在一起。以往的研究工作只关注全局或是局部的标记关系，毫无疑问，将两者同时考虑将会更有效。

利用标记关系的一大难题是如何去指定标记关系，它们通常很难由人给定，尤其当标记非常多的时候。所以在应用中，标记关系通常是从观察数据中估计出来的。有的方法假设标记是有层次结构的，它们通过层次聚类 [51] 或是贝叶斯网络结构学习 [36, 52] 获取这种层次关系。可是在很多应用中，这种层次结构并不存在。例如标记 “沙漠”，“山”，“海”，“日落”，“树” 中并不存在层次关系。另一些方法通过统计训练数据中标记同时出现的频率 [53] 来估计标记间

的关系。但是这种统计方法当有的标记只有非常少量正样本时并不可靠。而在有缺失标记的情况下估计标记关系矩阵变得更加困难，因为缺失标记的存在，观察到的标记分布与真实标记分布不同。之前提到的基于层次聚类或是同现频率的标记关系估计方法将会产生巨大偏差，进一步给学习结果带来负面影响。

针对部分标记缺失的多标记学习问题，本章提出了 GLOCAL (learning with GLObal and loCAL label correlations) 方法，它同时恢复缺失标记，训练线性分类器，发现与利用全局和局部标记关系。它无需额外的先验知识指定标记关系矩阵，而是在优化过程中同时习得。

## 2.2 本文方法

本节提出了 GLOCAL 方法。它在学习低秩标记矩阵分解、训练分类器的同时，探索、利用了全局和局部标记关系。GLOCAL 的成功主要由于以下四个因素：(1) 利用了标记矩阵的低秩结构，从而获得更紧致抽象的隐标记表示，并可以天然解决缺失标记补全问题；(2) 同时利用了全局和局部的标记关系；(3) 自动从数据中直接习得标记关系，而无需人工指定标记关系矩阵；(4) 整合了上述内容，并形式化成一个联合优化问题，可以通过交替优化高效求解。

### 2.2.1 GLOCAL 方法形式化

GLOCAL 的基本模型对标记矩阵进行低秩分解得到隐标记，并同时学得从特征空间到隐标记的映射。这样可以获得更紧致更抽象的隐标记表示。这种表示是低维稠密的实数值，和原来高维稀疏的二值标记空间相比，更容易通过优化得到一个好的映射。除此之外，这种分解能够直接解决缺失标记恢复问题。具体地，应用式 (2.1) 对标记矩阵 $\tilde{\boldsymbol{Y}}$ 进行低秩分解得到 $\boldsymbol{U}$ 和 $\boldsymbol{V}$, 其中 $\boldsymbol{V}$ 表示隐标记，$\boldsymbol{U}$ 反映了原标记与隐标记的关联程度。$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 可以通过最小化重构误差 $\|\tilde{\boldsymbol{Y}} - \boldsymbol{U}\boldsymbol{V}\|_F^2$ 求解。

为了能够映射特征空间到隐标记空间，通过最小化平方损失 $\|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2$ 获得映射矩阵 $\boldsymbol{W} \in \mathbb{R}^{d\times k}$。其中，$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ 是样本矩阵。然后，通过 $\operatorname{sign}(\boldsymbol{f}(\boldsymbol{x}))$ 得到样本 $\boldsymbol{x}$ 的预测标记，其中，$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{W}^\top \boldsymbol{x}$。令 $\boldsymbol{f} = [f_1, \ldots, f_l]^\top$，其中 $f_j(\boldsymbol{x})$ 是对 $\boldsymbol{x}$ 在第 $j$ 个标记上的预测值。把对所有样本的预测 $\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{X}$ 拼接在一起，可得 $\boldsymbol{F}_0 = [\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n)] = \boldsymbol{U}\boldsymbol{W}^\top \boldsymbol{X}$。

将低秩矩阵分解的重构误差与学习样本到隐标记映射的平方损失合并，即

可得 GLOCAL 的基本优化模型：

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \|\Pi_{\Omega}(\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V})\|_F^2+\lambda\|\boldsymbol{V}-\boldsymbol{W}^{\top}\boldsymbol{X}\|_F^2+\lambda_2\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}), \tag{2.2}$$

其中 $\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$ 表示正则化项，$\lambda,\lambda_2$ 表示权衡参数。尽管式 (2.2) 使用了平方损失，但它可以被任何可微损失函数替代。

利用标记关系是多标记学习能够成功的重要环节。这里，通过标记关系对模型正则化。由于全局和局部标记关系可能共存，考虑利用全局和部分标记关系设计流形正则化项。设计全局流形正则化项的基本思想是由样本流形正则化项 [54] 启发而来，鼓励学习器在正相关的标记上的输出更相近，反之亦然，即标记正相关会将相应的学习器输出值拉近，而标记负相关则会将相应学习器输出值推远。

将其形式化：对所有 $n$ 个样本的预测输出表示在 $l\times n$ 矩阵 $\boldsymbol{F}_0$ 中，其中，第 $i$ 行 $\boldsymbol{f}_{i,:}$ 是对第 $i$ 个标记的输出。如果第 $i$ 和 $j$ 标记更加正相关，则 $\boldsymbol{f}_{i,:}$ 应该与 $\boldsymbol{f}_{j,:}$ 更接近，反之亦然。类比样本流形正则化项 [54, 55]，定义标记流形正则化项如式 (2.3) 所示。

$$\sum_{i,j}[\boldsymbol{S}_0]_{i,j}\|\boldsymbol{f}_{i,:}-\boldsymbol{f}_{j,:}\|_2^2, \tag{2.3}$$

其中，$\boldsymbol{S}_0$ 表示 $l\times l$ 全局标记关系矩阵。如果第 $i$ 和 $j$ 标记正相关，则 $[\boldsymbol{S}_0]_{i,j}$ 为正数。通过最小化式 (2.3), $\|\boldsymbol{f}_{i,:}-\boldsymbol{f}_{j,:}\|_2^2$ 将会变得比较小。令 $\boldsymbol{D}_0$ 为对角矩阵，其对角线元素为 $\boldsymbol{S}_0\mathbf{1}$，$\mathbf{1}$ 代表全 1 向量。流形正则化项式 (2.3) 可等价写为 $\mathrm{tr}(\boldsymbol{F}_0^{\top}\boldsymbol{L}_0\boldsymbol{F}_0)$ [56]，其中 $\boldsymbol{L}_0=\boldsymbol{D}_0-\boldsymbol{S}_0$ 是 $\boldsymbol{S}_0$ 的 $l\times l$ 标记拉普拉斯矩阵。

考虑到不同的局部样本上的局部标记关系可能大不相同，引入局部标记流形正则化。假设整个数据集 $\boldsymbol{X}$ 被划分为 $g$ 组：$\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_g\}$，每组 $\boldsymbol{X}_m\in\mathbb{R}^{d\times n_m}$ 包含 $n_m$ 个样本。分组划分可由领域知识得到 (如生物信息学应用中的基因路径[57] 和基因网络[58] )，或是通过聚类得到。令 $\boldsymbol{Y}_m$ 表示标记矩阵 $\boldsymbol{Y}$ 中的子矩阵，对应分组 $\boldsymbol{X}_m$ 的标记；令 $\boldsymbol{S}_m\in\mathbb{R}^{l\times l}$ 表示第 $m$ 组的局部标记关系矩阵。与全局标记关系类似，在局部样本上，仍希望学习器的输出结果在正相关标记上相近，负相关标记上相远，并通过最小化 $\mathrm{tr}(\boldsymbol{F}_m^{\top}\boldsymbol{L}_m\boldsymbol{F}_m)$ 实现这点。其中，$\boldsymbol{L}_m$ 是 $\boldsymbol{S}_m$ 的拉普拉斯矩阵，$\boldsymbol{F}_m=\boldsymbol{U}\boldsymbol{W}^{\top}\boldsymbol{X}_m$ 是第 $m$ 组上的学习器输出。

将全局和局部正则化项加入基本模型 (2.2)，可得式 (2.4)：

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \|\Pi_{\Omega}(\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V})\|_F^2+\lambda\|\boldsymbol{V}-\boldsymbol{W}^{\top}\boldsymbol{X}\|_F^2+\lambda_2\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})+\lambda_3\mathrm{tr}(\boldsymbol{F}_0^{\top}\boldsymbol{L}_0\boldsymbol{F}_0)$$

$$+\sum_{m=1}^{g}\lambda_4\mathrm{tr}(\boldsymbol{F}_m^\top\boldsymbol{L}_m\boldsymbol{F}_m), \tag{2.4}$$

其中，$\lambda,\lambda_2,\lambda_3,\lambda_4$ 是权衡参数。

全局标记关系蕴含于拉普拉斯矩阵 $\boldsymbol{L}_0$ 中，而局部标记关系则蕴含于每个 $\boldsymbol{L}_m$ 中。直观上看，大的局部分组的局部标记关系将对全局标记关系贡献更大。特别地，接下来的引理2.1将展示这点：当余弦距离被用于计算 $\boldsymbol{S}_{ij}$ 时，有 $\boldsymbol{S}_0=\sum_{m=1}^{g}\frac{n_m}{n}\boldsymbol{S}_m$。

**引理 2.1** 令 $[\boldsymbol{S}_0]_{ij}=\frac{\boldsymbol{y}_{i,:}\boldsymbol{y}_{j,:}^\top}{\|\boldsymbol{y}_{i,:}\|\cdot\|\boldsymbol{y}_{j,:}\|}$，$[\boldsymbol{S}_m]_{ij}=\frac{\boldsymbol{y}_{m,i,:}\boldsymbol{y}_{m,j,:}^\top}{\|\boldsymbol{y}_{m,i,:}\|\cdot\|\boldsymbol{y}_{m,j,:}\|}$，其中 $\boldsymbol{y}_{i,:}$ 是 $\boldsymbol{Y}$ 的第 $i$ 行，$\boldsymbol{y}_{m,i,:}$ 是 $\boldsymbol{Y}_m$ 的第 $i$ 行，那么 $\boldsymbol{S}_0=\sum_{m=1}^{g}\frac{n_m}{n}\boldsymbol{S}_m$ 成立。

**证明**: 因为 $\mathcal{Y}\in\{-1,1\}^l$，有 $\|\boldsymbol{y}_{i,:}\|=\sqrt{n}$，对 $\forall i$ 成立；$\|\boldsymbol{y}_{m,i,:}\|=\sqrt{n_m}$，对 $\forall i,m$ 成立。不失一般性，令样本按分组排列，即 $\boldsymbol{y}_{i,:}=[\boldsymbol{y}_{1,i,:},\cdots,\boldsymbol{y}_{m,i,:}]$。则

$$[\boldsymbol{S}_0]_{ij} = \frac{1}{n}[\boldsymbol{y}_{1,i,:},\cdots,\boldsymbol{y}_{m,i,:}][\boldsymbol{y}_{1,j,:},\cdots,\boldsymbol{y}_{m,j,:}]^\top=\frac{1}{n}\sum_{g=1}^{m}\boldsymbol{y}_{g,i,:}\boldsymbol{y}_{g,j,:}^\top=\sum_{g=1}^{m}\frac{n_m}{n}[\boldsymbol{S}_m]_{ij}.$$

□

**命题 2.2** 如果 $\boldsymbol{S}_0=\sum_{m=1}^{g}\beta_m\boldsymbol{S}_m$，则 $\boldsymbol{L}_0=\sum_{m=1}^{g}\beta_m\boldsymbol{L}_m$。

一般而言，当全局标记关系矩阵是局部关系矩阵的线性组合时，命题2.2展示相应的全局标记拉普拉斯矩阵同样也是局部标记拉普拉斯矩阵的线性组合，且组合系数与关系矩阵的组合系数相同。

根据引理2.1和命题2.2，式 (2.4) 可被改写为

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}}\ &\|\Pi_\Omega(\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V})\|_F^2+\lambda\|\boldsymbol{V}-\boldsymbol{W}^\top\boldsymbol{X}\|_F^2+\lambda_2\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\\&+\sum_{m=1}^{g}\left(\frac{\lambda_3 n_m}{n}\mathrm{tr}(\boldsymbol{F}_0^\top\boldsymbol{L}_m\boldsymbol{F}_0)+\lambda_4\mathrm{tr}(\boldsymbol{F}_m^\top\boldsymbol{L}_m\boldsymbol{F}_m)\right).\end{aligned} \tag{2.5}$$

标记流形正则化的成功取决于好的标记关系矩阵 (或等价地，好的标记拉普拉斯矩阵)。在多标记学习中，一个估计标记关系的方法是计算两者之间的余弦距离 [59]。但是这种方法得到的估计偏差很大，因为有些标记在训练集中可能只有非常少量的正样本。而当有标记缺失的情况下，这种方法得到的估计甚至可能是错误的。因为在标记缺失情况下观察到的标记分布和真实的标记分布不

**算法 2.1** GLOCAL

**输入**: 特征矩阵 $\boldsymbol{X}$, 标记矩阵 $\boldsymbol{Y}$, 观察指示矩阵 $\boldsymbol{J}$, 以及小组划分
**输出**: $\boldsymbol{U}, \boldsymbol{W}, \mathcal{Z}$.

初始化 $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \mathcal{Z}$;
**repeat**
  **for** $m = 1, \ldots, g$
    固定 $\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{W}$, 通过优化式 (2.7) 更新 $\boldsymbol{Z}_m$; //学习标记关系
  **end for**
  固定 $\boldsymbol{U}, \boldsymbol{W}, \mathcal{Z}$，通过优化式 (2.8) 更新 $\boldsymbol{V}$; //学习隐标记
  固定 $\boldsymbol{V}, \boldsymbol{W}, \mathcal{Z}$，通过优化式 (2.9) 更新 $\boldsymbol{U}$; //学习原始标记到隐标记的映射
  固定 $\boldsymbol{U}, \boldsymbol{V}, \mathcal{Z}$，通过优化式 (2.10) 更新 $\boldsymbol{W}$; //学习特征到隐标记的映射
**until** 算法收敛;
输出 $\boldsymbol{U}, : bmV, \boldsymbol{W}$，$\mathcal{Z} \equiv \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_g\}$。

同，直接用余弦距离进行估计自然不准。

在本工作中，直接去学标记拉普拉斯矩阵，而非人为指定某个相关性度量或是标记关系矩阵。由于拉普拉斯矩阵是对称正定矩阵，对每个局部组 $m \in \{1, \ldots, g\}$，将标记拉普拉斯矩阵 $\boldsymbol{L}_m$ 分解为 $\boldsymbol{Z}_m \boldsymbol{Z}_m^\top$，其中 $\boldsymbol{Z}_m \in \mathbb{R}^{l \times k}$。为了简化计算，设置 $k$ 与隐标记表示 $\boldsymbol{V}$ 的维度相同。因此，学习拉普拉斯矩阵被转换成为学 $\mathcal{Z} \equiv \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_g\}$。注意到关于 $\boldsymbol{Z}_m$ 的优化可能会导致平凡解：$\boldsymbol{Z}_m = \boldsymbol{0}$。为了避免平凡解，为对角矩阵元素增加约束：$\mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \boldsymbol{1}$。这种约束使得优化可以得到 $\boldsymbol{L}_m$ 对应的归一化拉普拉斯矩阵 [60]。

令 $\boldsymbol{J} = [J_{ij}]$ 为指示矩阵，如果 $(i, j) \in \Omega$ 则 $J_{ij} = 1$，否则 $J_{ij} = 0$。$\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})$ 可以写作 $\boldsymbol{J}$ 和 $\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}$ 的对应元素乘积 $\boldsymbol{J} \circ (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})$。加入拉普拉斯矩阵分解和对角约束，可得优化式 (2.6)：

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\mathcal{Z}} \quad & \|\boldsymbol{J} \circ (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})\|_F^2 + \lambda \|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 + \lambda_2 \mathcal{R}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \\ & + \sum_{m=1}^{g} \left( \frac{\lambda_3 n_m}{n} \mathrm{tr}\,(\boldsymbol{F}_0^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_0) + \lambda_4 \mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_m) \right) \\ \text{s.t.} \quad & \mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \boldsymbol{1}, m = 1, 2, \ldots, g. \end{aligned} \tag{2.6}$$

在后文使用标准的 F 范数的平方作为 $\mathcal{R}$：$\mathcal{R}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) = \|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 + \|\boldsymbol{W}\|_F^2$。

### 2.2.2 优化求解

优化问题 (2.6) 可以通过交替优化求解 (算法 2.1)，从而迭代调整优化变量，找到满意的解。在每一轮迭代中，通过梯度下降法更新 $\{\mathcal{Z}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}\}$ 其中

之一，其它变量保持固定。本来复杂的优化问题就被转换成几个更简单的子问题，从而可以更容易求解。具体实现采用 MANOPT 工具包 [61]，利用带线性搜索的梯度下降法在欧氏空间中更新 $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$，以及在流形上更新 $\mathcal{Z}$。

**更新 $\boldsymbol{Z}_m$ (算法 2.1第 4 行)**

当固定 $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$，优化问题 (2.6) 将简化为式 (2.7)：

$$\begin{aligned} \min_{\boldsymbol{Z}_m} \quad & \frac{\lambda_3 n_m}{n} \mathrm{tr}\left(\boldsymbol{F}_0^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_0\right) + \lambda_4 \mathrm{tr}\left(\boldsymbol{F}_m^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_m\right) \\ \text{s.t.} \quad & \mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \mathbf{1}, \end{aligned} \tag{2.7}$$

$m \in \{1, \ldots, g\}$。由于约束 $\mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \mathbf{1}$，式 (2.7) 没有闭式解，用投影梯度下降来解。目标函数关于 $\boldsymbol{Z}_m$ 的梯度是

$$\nabla_{\boldsymbol{Z}_m} = \frac{\lambda_3 n_m}{n} \boldsymbol{U}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{U}^\top \boldsymbol{Z}_m + \lambda_4 \boldsymbol{U}\boldsymbol{W}^\top \boldsymbol{X}_m \boldsymbol{X}_m^\top \boldsymbol{W}\boldsymbol{U}^\top \boldsymbol{Z}_m.$$

为了使得约束 $\mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \mathbf{1}$ 满足，在每次梯度下降更新后，将 $\boldsymbol{Z}_m$ 的每一行投影到单位范数球：$\boldsymbol{z}_{m,j,:} \leftarrow \boldsymbol{z}_{m,j,:} / \|\boldsymbol{z}_{m,j,:}\|$，其中，$\boldsymbol{z}_{m,j,:}$ 代表 $\boldsymbol{Z}_m$ 的第 $j$ 行。

**更新 $\boldsymbol{V}$ (算法 2.1第 6 行)**

固定各 $\boldsymbol{Z}_m$ 和 $\boldsymbol{U}, \boldsymbol{W}$，优化问题 (2.6) 简化为式 (2.8)：

$$\min_{\boldsymbol{V}} \|\boldsymbol{J} \circ (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})\|_F^2 + \lambda \|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 + \lambda_2 \|\boldsymbol{V}\|_F^2. \tag{2.8}$$

由于 $\boldsymbol{V}$ 的每一列相互独立，$\boldsymbol{V}$ 可以一列一列优化。令 $\boldsymbol{j}_i$ 和 $\boldsymbol{v}_i$ 分别表示 $\boldsymbol{J}$ 和 $\boldsymbol{V}$ 的第 $i$ 列，则 $\boldsymbol{v}_i$ 的优化可以写作：

$$\min_{\boldsymbol{v}_i} \|\mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{y}_i - \mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{U}\boldsymbol{v}_i\|^2 + \lambda \|\boldsymbol{v}_i - \boldsymbol{W}^\top \boldsymbol{x}_i\|^2 + \lambda_2 \|\boldsymbol{v}_i\|^2.$$

令目标函数关于 $\boldsymbol{v}_i$ 的梯度为 0，可以得到 $\boldsymbol{v}_i$ 的闭式解：

$$\boldsymbol{v}_i = \left(\boldsymbol{U}^\top \mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{U} + (\lambda + \lambda_2)\mathbf{I}\right)^{-1} \left(\lambda \boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{U}^\top \mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{y}_i\right).$$

但是这个闭式解对每个 $i$ 都要计算矩阵求逆操作。如果计算开销比较大，可以用梯度下降求解式 (2.8)，其中目标函数关于 $\boldsymbol{V}$ 的梯度为：

$$\nabla_{\boldsymbol{V}} = \boldsymbol{U}^\top (\boldsymbol{J} \circ (\boldsymbol{U}\boldsymbol{V} - \boldsymbol{Y})) + \lambda(\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}) + \lambda_2 \boldsymbol{V}.$$

**更新 $U$ (算法 2.1第 7 行)**

固定 $Z_m$、$V$ 和 $W$，优化问题 (2.6) 简化为式 (2.9)：

$$\min_{U} \|J \circ (Y - UV)\|_F^2 + \lambda_2 \|U\|_F^2 + \sum_{m=1}^{g} \left( \frac{\lambda_3 n_m}{n} \mathrm{tr}(F_0^\top Z_m Z_m^\top F_0) + \lambda_4 \mathrm{tr}(F_m^\top Z_m Z_m^\top F_m) \right). \tag{2.9}$$

它关于 $U$ 的梯度为：

$$\nabla_U = (J \circ (UV - Y))V^\top + \lambda_2 U + \sum_{m=1}^{g} Z_i Z_i^\top U \left( \frac{\lambda_3 n_m}{n} W^\top X_m X_m^\top W + \lambda_4 W^\top X X^\top W \right).$$

**更新 $W$ (2.1第 8 行)**

固定各 $Z_m$ 和 $U, V$，优化问题 (2.6) 简化为式 (2.10)：

$$\min_{W} \lambda \|V - W^\top X\|_F^2 + \lambda_2 \|W\|_F^2 + \sum_{m=1}^{g} \left( \frac{\lambda_3 n_m}{n} \mathrm{tr}(F_0^\top Z_m Z_m^\top F_0) + \lambda_4 \mathrm{tr}(F_m^\top Z_m Z_m^\top F_m) \right). \tag{2.10}$$

它关于 $W$ 的梯度为:

$$\nabla_W = \lambda X (X^\top W - V^\top) + \lambda_2 W + \sum_{m=1}^{g} \left( \frac{\lambda_3 n_m}{n} X X^\top + \lambda_4 X_m X_m^\top \right) W U^\top Z_m Z_m^\top U.$$

## 2.3 实验测试

通过大量在文本和图像数据集上的实验验证了在完整标记情况下和部分标记缺失情况下 GLOCAL 方法的有效性。

### 2.3.1 实验设置

在广泛使用的多标记学习标准数据集上验证 GLOCAL 的有效性。其中，文本数据集包括 9 个 Yahoo 数据集 (Arts, Business, Computers, Education, Entertainment, Health, Recreation, Reference, Society) [62] 和 Enron 数据集 [63]；图像数据集包括 Corel5k [64] 和 Image [65]。每个数据集均由其前三个字母缩写。细节信息如表2.1所示。对每一个数据集，随机选择 60% 样本作为训练集，剩余样本用于测试。为了模拟缺失标记，随机从真实标记矩阵中采样 $\rho$% 个元素作为观察到的标记，剩余部分作为缺失标记。当 $\rho = 100$ 时，即为标记完整情况

**表 2.1** 实验数据集 ( “# 标记/样本” 代表每个样本平均标记数).

| 数据集 | 样本数 | 样本维度 | 标记数 | # 标记/样本 |
|---|---|---|---|---|
| Arts | 5,000 | 462 | 26 | 1.64 |
| Business | 5,000 | 438 | 30 | 1.59 |
| Computers | 5,000 | 681 | 33 | 1.51 |
| Education | 5,000 | 550 | 33 | 1.46 |
| Entertainment | 5,000 | 640 | 21 | 1.42 |
| Health | 5,000 | 612 | 32 | 1.66 |
| Recreation | 5,000 | 606 | 22 | 1.42 |
| Reference | 5,000 | 793 | 33 | 1.17 |
| Society | 5,000 | 636 | 27 | 1.69 |
| Enron | 1,702 | 1,001 | 53 | 3.37 |
| Corel5k | 5,000 | 499 | 374 | 3.52 |
| Image | 2,000 | 294 | 5 | 1.24 |

同时关注算法对未知测试样本预测标记的性能以及恢复训练集中缺失标记的性能。令 $p$ 为测试样本的个数；$\boldsymbol{C}_i^+, \boldsymbol{C}_i^-$ 分别表示第 $i$ 个样本的正、负标记集合；$\boldsymbol{Z}_j^+, \boldsymbol{Z}_j^-$ 分别表示第 $j$ 个标记的正、负样本集合。给定一个输入样本 $\boldsymbol{x}$，令 $\mathrm{rank}_{\boldsymbol{f}}(\boldsymbol{x}, y)$ 表示标记 $y$ 的预测值排序 (从高到底)。用以下常用指标进行评判 [66]：

(1) Ranking loss (Rkl)：认为在多标记学习的预测结果中，相关标记的预测值要比无关标记大，否则就认为这一对标记的相对排序产生了错误。对于样本 $i$ 定义 $\boldsymbol{Q}_i = \{(j', j'') \mid f_{j'}(\boldsymbol{x}_i) \leq f_{j''}(\boldsymbol{x}_i), (j', j'') \in \boldsymbol{C}_i^+ \times \boldsymbol{C}_i^-\}$，则 $\mathrm{Rkl} = \frac{1}{p} \sum_{i=1}^{p} \frac{|\boldsymbol{Q}_i|}{|\boldsymbol{C}_i^+||\boldsymbol{C}_i^-|}$.

(2) Average AUC (Auc): 这个指标表征了平均每个标记的正样本排在负样本之前的比例。对于标记 $j$，定义 $\tilde{\boldsymbol{Q}}_j = \{(i', i'') \mid f_j(\boldsymbol{x}_{i'}) \geq f_j(\boldsymbol{x}_{i''}), (\boldsymbol{x}_{i'}, \boldsymbol{x}_{i''}) \in \boldsymbol{Z}_j^+ \times \boldsymbol{Z}_j^-\}$，则 $\mathrm{Auc} = \frac{1}{l} \sum_{j=1}^{l} \frac{|\tilde{\boldsymbol{Q}}_j|}{|\boldsymbol{Z}_j^+||\boldsymbol{Z}_j^-|}$.

(3) Coverage (Cvg): 它度量的是预测的最后一个正标记所在的位次。$\mathrm{Cvg} = \frac{1}{p} \sum_{i=1}^{p} \max\{\mathrm{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, j) \mid j \in \boldsymbol{C}_i^+\} - 1$.

(4) Average precision (Ap): 计算了排在某个正标记之前的正标记的平均个数。对于样本 $i$，定义 $\hat{\boldsymbol{Q}}_{i,c} = \{j \mid \mathrm{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, j) \leq \mathrm{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, c), j \in \boldsymbol{C}_i^+\}$，则 $\mathrm{Ap} = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{|\boldsymbol{C}_i^+|} \sum_{c \in \boldsymbol{C}_i^+} \frac{|\hat{\boldsymbol{Q}}_{i,c}|}{\mathrm{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, c)}$。

对于 Auc 和 Ap，指标值越大越好；而对于 Rkl 和 Cvg，指标值则是越小越好。

在 GLOCAL 算法中，用 `kmeans` 聚类方法将数据划分成若干局部小组，用式 (2.2) 的解对各个变量进行初始化。实验对比了以下多标记学习算法：

(1) BR [43]：为每个标记独立训练一个二分类线性 SVM(实现使用 LIbBLINEAR 工具包[67])；

(2) MLLOC [49]：通过将标记关系编码到样本的特征表示来利用局部标记关系；

(3) LEML [19]：学习一个线性的样本到标记的有低秩结构的映射矩阵，隐式地利用了全局标记关系；

(4) ML-LRC [37]：显式地学习并利用了低秩的全局标记关系。

从利用标记关系的角度来看，BR 没有考虑任何标记关系；MLLOC 只考虑局部标记关系；LEML 隐式地利用了全局标记关系；ML-LRC 直接对全局标记关系进行建模；而 GLOCAL 同时利用了全局和局部的标记关系。

从处理标记缺失任务的能力来看，BR、MLLOC，均只能处理训练集标记完整的情况；而 LEML，ML-LRC，和 GLOCAL 则可处理有标记缺失的多标记学习任务。为了能够在训练集部分标记缺失数据集上对比 BR 和 MLLOC 的性能，首先使用 MAXIDE 算法 [18] 恢复所有的缺失标记。用 MBR 和 MMLLOC 分别代表 MAXIDE+BR 和 MAXIDE+MLLOC 的组合。

### 2.3.2 实验结果

训练集中观察标记以不同比率缺失时的恢复结果如表 2.2 所示。其中因为 BR 和 MLLOC 不能直接处理缺失标记，所以用 MAXIDE 做缺失标记恢复。测试集标记预测结果则如表 2.3 所示。

由表2.2和2.3可见，观察到的标记越多，学习结果越好，这与直觉相吻合：越多标记意味着提供了更多的监督信息，从而可以得到更好的模型。总体上，GLOCAL 在不同的观察标记的比率下，在缺失标记恢复和测试集标记预测任务上都取得了最好的结果。它的成功是因为同时优化了标记矩阵低秩分解、学习了特征空间到隐标记的映射、以及包含了全局和局部标记关系的各标记拉普拉斯矩阵。由标记矩阵低秩分解，可以得到更紧致、信息量更大的隐标记。隐标记空间是稠密的、实值的、低维的，更容易学得特征空间到隐标记的映射。从总体上，全局标记关系流形提供了整体上标记是如何相关联的，使得学习器能够预测出现次数较少的标记：如果这些少量标记和某些标记正相关，全局标记流形会鼓励它们的预测尽可能接近，反之亦然。而局部标记流形则进一步调整分类器。学到的标记拉普拉斯矩阵将最好地匹配全局和局部数据，可以避免人工指定标记关系的种种困难和产生的偏差。

**表 2.2** 缺失标记恢复结果。Rkl 和 Cvg 越小越好，Auc 和 Ap 越大越好。斜体表示 GLOCAL 显著优于对比方法（t 检验，$\alpha = 0.05$）。粗体显示的是对应数据集上的最佳结果。

| | | $\rho$ | GLOCAL | MAXIDE | LEML | ML-LRC | | | $\rho$ | GLOCAL | MAXIDE | LEML | ML-LRC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Art | Rkl | 30 | **0.103** | *0.131* | *0.133* | *0.137* | Bus | Rkl | 30 | **0.029** | *0.044* | *0.046* | *0.046* |
| | | 70 | **0.074** | *0.083* | *0.090* | *0.083* | | | 70 | **0.021** | *0.026* | *0.027* | *0.024* |
| | Auc | 30 | **0.897** | *0.871* | *0.848* | *0.879* | | Auc | 30 | **0.971** | *0.956* | *0.954* | *0.954* |
| | | 70 | **0.928** | *0.918* | *0.912* | *0.910* | | | 70 | **0.979** | *0.974* | *0.973* | *0.974* |
| | Cvg | 30 | **4.189** | *5.195* | *5.231* | *5.161* | | Cvg | 30 | **1.830** | *2.550* | *2.622* | *2.622* |
| | | 70 | **3.234** | *3.616* | *3.733* | *3.778* | | | 70 | **1.477** | *1.742* | *1.783* | *1.746* |
| | Ap | 30 | **0.652** | *0.645* | *0.634* | *0.640* | | Ap | 30 | **0.893** | *0.876* | *0.878* | *0.876* |
| | | 70 | **0.720** | **0.720** | **0.720** | *0.709* | | | 70 | **0.908** | *0.905* | *0.901* | *0.903* |
| Com | Rkl | 30 | **0.073** | *0.101* | *0.098* | *0.097* | Edu | Rkl | 30 | **0.069** | *0.097* | *0.093* | *0.089* |
| | | 70 | **0.052** | *0.059* | *0.063* | *0.061* | | | 70 | **0.058** | *0.061* | *0.061* | *0.061* |
| | Auc | 30 | **0.933** | *0.905* | *0.908* | *0.909* | | Auc | 30 | **0.932** | *0.902* | *0.907* | *0.911* |
| | | 70 | **0.955** | *0.947* | *0.943* | *0.945* | | | 70 | **0.942** | *0.938* | *0.938* | 0.940 |
| | Cvg | 30 | **3.511** | *4.627* | *4.586* | *4.565* | | Cvg | 30 | **3.171** | *4.672* | *4.372* | *3.914* |
| | | 70 | **2.586** | *2.912* | *3.100* | *3.095* | | | 70 | **2.815** | *3.113* | *3.106* | 3.000 |
| | Ap | 30 | **0.726** | *0.709* | *0.700* | *0.705* | | Ap | 30 | **0.655** | 0.653 | *0.648* | 0.653 |
| | | 70 | **0.787** | **0.787** | **0.787** | **0.787** | | | 70 | **0.711** | **0.711** | *0.702* | 0.710 |
| Ent | Rkl | 30 | **0.085** | *0.104* | *0.103* | *0.106* | Hea | Rkl | 30 | **0.041** | *0.060* | *0.057* | *0.054* |
| | | 70 | **0.062** | *0.063* | *0.063* | *0.063* | | | 70 | **0.030** | *0.037* | *0.036* | 0.032 |
| | Auc | 30 | **0.916** | *0.898* | *0.899* | *0.899* | | Auc | 30 | **0.960** | *0.941* | *0.943* | *0.947* |
| | | 70 | **0.940** | **0.940** | 0.938 | **0.940** | | | 70 | **0.971** | *0.964* | *0.964* | 0.968 |
| | Cvg | 30 | **2.512** | *3.058* | *2.994* | *3.022* | | Cvg | 30 | **2.567** | *3.577* | *3.462* | *3.465* |
| | | 70 | **1.957** | 1.987 | 2.051 | 2.080 | | | 70 | **2.152** | *2.524* | *2.465* | *2.450* |
| | Ap | 30 | **0.704** | 0.704 | *0.698* | *0.698* | | Ap | 30 | **0.801** | *0.796* | *0.794* | 0.798 |
| | | 70 | **0.768** | *0.763* | 0.765 | 0.765 | | | 70 | **0.848** | **0.848** | *0.842* | **0.848** |

续表 2.2

| | | $\rho$ | GLOCAL | MAXIDE | LEML | ML-LRC | | | $\rho$ | GLOCAL | MAXIDE | LEML | ML-LRC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rec | Rkl | 30 | **0.110** | *0.130* | *0.133* | *0.135* | Ref | Rkl | 30 | **0.063** | *0.083* | *0.083* | *0.083* |
| | | 70 | **0.068** | *0.078* | *0.080* | *0.080* | | | 70 | **0.048** | **0.048** | 0.049 | 0.049 |
| | Auc | 30 | **0.895** | *0.873* | *0.870* | *0.869* | | Auc | 30 | **0.939** | *0.919* | *0.919* | *0.918* |
| | | 70 | **0.934** | *0.925* | *0.923* | *0.920* | | | 70 | **0.955** | **0.955** | 0.953 | 0.953 |
| | Cvg | 30 | **3.291** | *3.899* | *3.919* | *4.048* | | Cvg | 30 | **2.520** | *3.436* | *3.392* | *3.372* |
| | | 70 | **2.262** | *2.560* | *2.607* | *2.620* | | | 70 | **1.972** | *2.039* | *2.103* | *2.195* |
| | Ap | 30 | **0.681** | *0.680* | *0.663* | *0.660* | | Ap | 30 | 0.679 | **0.681** | *0.664* | 0.674 |
| | | 70 | **0.770** | *0.767* | *0.763* | *0.760* | | | 70 | **0.746** | 0.745 | **0.746** | **0.746** |
| Soc | Rkl | 30 | **0.102** | *0.129* | *0.128* | *0.123* | Enr | Rkl | 30 | **0.075** | *0.091* | *0.115* | *0.085* |
| | | 70 | **0.073** | 0.074 | *0.081* | **0.073** | | | 70 | **0.040** | 0.042 | *0.060* | **0.040** |
| | Auc | 30 | **0.898** | *0.871* | *0.872* | *0.877* | | Auc | 30 | **0.926** | *0.910* | *0.887* | *0.918* |
| | | 70 | **0.929** | 0.926 | *0.919* | 0.928 | | | 70 | **0.962** | 0.960(3 | *0.942* | **0.962** |
| | Cvg | 30 | **4.496** | *5.557* | *5.459* | *5.167* | | Cvg | 30 | **12.05** | *14.24* | *16.65* | *13.45* |
| | | 70 | **3.442** | *3.641* | *3.824* | *3.608* | | | 70 | 7.510 | *7.961* | *10.33* | **7.480** |
| | Ap | 30 | **0.652** | 0.646 | *0.629* | 0.650 | | Ap | 30 | **0.739** | **0.739** | *0.711* | **0.739** |
| | | 70 | **0.719** | **0.719** | 0.717 | **0.719** | | | 70 | **0.855** | 0.854 | *0.842* | **0.855** |
| Cor | Rkl | 30 | **0.185** | *0.226* | *0.214* | *0.206* | Ima | Rkl | 30 | **0.173** | *0.302* | *0.184* | 0.175 |
| | | 70 | 0.125 | *0.138* | *0.131* | **0.123** | | | 70 | **0.148** | *0.251* | **0.148** | **0.148** |
| | Auc | 30 | **0.814** | *0.773* | *0.786* | *0.794* | | Auc | 30 | **0.828** | *0.820* | **0.828** | 0.826 |
| | | 70 | 0.874 | 0.874 | 0.874 | 0.874 | | | 70 | 0.855 | *0.834* | **0.857** | 0.855 |
| | Cvg | 30 | **153.82** | *204.90* | *182.76* | *178.60* | | Cvg | 30 | **0.950** | *1.493* | *1.104* | 0.967 |
| | | 70 | **102.30** | 103.63 | 102.42 | **102.30** | | | 70 | **0.760** | *0.790* | **0.760** | 0.770 |
| | Ap | 30 | **0.275** | **0.275** | *0.259* | **0.275** | | Ap | 30 | **0.785** | *0.739* | *0.776* | *0.775* |
| | | 70 | 0.279 | 0.279 | 0.279 | 0.279 | | | 70 | **0.841** | *0.768* | **0.841** | 0.834 |

**表 2.3** 测试集预测结果。Rkl 和 Cvg 越小越好，Auc 和 Ap 越大越好。斜体表示 GLOCAL 显著优于对比方法（t 检验，$\alpha = 0.05$）。粗体显示的是对应数据集上的最佳结果。

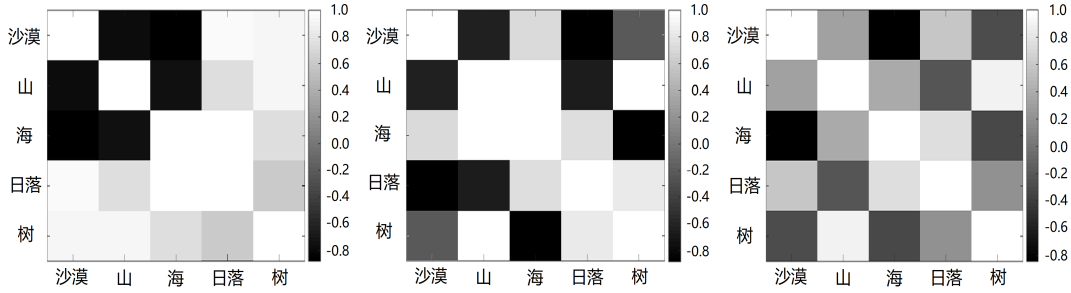
| | | $\rho$ | GLOCAL | MBR | MMLLOC | LEML | ML-LRC | | | $\rho$ | GLOCAL | MBR | MMLLOC | LEML | ML-LRC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Art | Rkl | 30 | **0.144** | *0.287* | *0.225* | *0.204* | *0.184* | Bus | Rkl | 30 | **0.054** | *0.105* | *0.083* | *0.063* | *0.061* |
| | | 70 | **0.139** | *0.244* | *0.193* | *0.181* | *0.159* | | | 70 | **0.046** | *0.084* | *0.064* | *0.058* | **0.046** |
| | Auc | 30 | **0.831** | *0.769* | *0.781* | *0.801* | 0.828 | | Auc | 30 | **0.937** | *0.898* | *0.917* | *0.928* | **0.937** |
| | | 70 | **0.840** | *0.790* | *0.819* | *0.825* | 0.838 | | | 70 | **0.952** | *0.920* | *0.935* | *0.942* | 0.950 |
| | Cvg | 30 | **5.867** | *9.296* | *9.033* | *7.369* | *6.281* | | Cvg | 30 | **2.863** | *5:195* | *4.643* | *3.954* | *3.279* |
| | | 70 | **5.352** | *8.234* | *7.262* | *6.431* | 5.432 | | | 70 | **2.579** | *4.313* | *3.670* | *3.303* | 2.580 |
| | Ap | 30 | **0.572** | *0.497* | *0.529* | *0.503* | *0.517* | | Ap | 30 | **0.879** | *0.823* | *0.843* | *0.866* | *0.858* |
| | | 70 | **0.607** | *0.537* | *0.583* | *0.589* | *0.588* | | | 70 | **0.881** | *0.845* | *0.861* | *0.870* | *0.870* |
| Com | Rkl | 30 | 0.154 | *0.227* | *0.201* | *0.179* | **0.152** | Edu | Rkl | 30 | **0.137** | *0.289* | *0.187* | *0.176* | *0.144* |
| | | 70 | **0.113** | *0.173* | *0.150* | *0.141* | 0.115 | | | 70 | **0.111** | *0.234* | *0.165* | *0.151* | 0.113 |
| | Auc | 30 | **0.883** | *0.834* | *0.849* | 0.880 | *0.873* | | Auc | 30 | **0.846** | *0.777* | *0.815* | *0.817* | 0.845 |
| | | 70 | **0.896** | *0.850* | *0.868* | 0.894 | 0.895 | | | 70 | **0.860** | *0.790* | *0.844* | *0.842* | **0.860** |
| | Cvg | 30 | **5.798** | *9.123* | *8.808* | *7.392* | *6.052* | | Cvg | 30 | **6.338** | *11.839* | *11.089* | *9.672* | 6.350 |
| | | 70 | **4.976** | *7.263* | *6.871* | *6.306* | 5.000 | | | 70 | **5.070** | *10.005* | *8.096* | *7.595* | 5.075 |
| | Ap | 30 | **0.669** | *0.586* | *0.631* | *0.646* | *0.636* | | Ap | 30 | **0.592** | *0.503* | *0.538* | *0.537* | *0.543* |
| | | 70 | **0.691** | *0.623* | *0.674* | *0.665* | *0.667* | | | 70 | **0.622** | *0.538* | *0.586* | *0.591* | *0.600* |
| Ent | Rkl | 30 | **0.122** | *0.260* | *0.229* | *0.175* | *0.152* | Hea | Rkl | 30 | **0.085** | *0.163* | *0.137* | *0.095* | **0.085** |
| | | 70 | **0.109** | *0.206* | *0.164* | *0.159* | *0.129* | | | 70 | **0.065** | *0.129* | *0.109* | *0.074* | *0.071* |
| | Auc | 30 | **0.859** | *0.784* | *0.832* | *0.826* | *0.849* | | Auc | 30 | 0.906 | *0.856* | *0.894* | *0.896* | **0.907** |
| | | 70 | **0.871** | *0.800* | *0.842* | *0.850* | 0.870 | | | 70 | **0.920** | *0.875* | *0.901* | **0.920** | **0.920** |
| | Cvg | 30 | **4.153** | *6.588* | *6.029* | *5.755* | 4.170 | | Cvg | 30 | **4.814** | *6.415* | *7.104* | *6.248* | 4.924 |
| | | 70 | **3.117** | *5.433* | *4.857* | *4.643* | *3.483* | | | 70 | 3.963 | *5.246* | *5.866* | *5.167* | **3.960** |
| | Ap | 30 | **0.645** | *0.565* | *0.601* | *0.601* | *0.601* | | Ap | 30 | **0.752** | *0.678* | *0.727* | *0.715* | *0.720* |
| | | 70 | **0.670** | *0.610* | *0.635* | *0.645* | *0.643* | | | 70 | **0.775** | *0.713* | *0.762* | *0.770* | *0.766* |

续表 2.3

| | | $\rho$ | GLOCAL | MBR | MMLLOC | LEML | ML-LRC | | | $\rho$ | GLOCAL | MBR | MMLLOC | LEML | ML-LRC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rec | Rkl | 30 | **0.165** | *0.279* | *0.266* | *0.245* | *0.202* | Ref | Rkl | 30 | **0.098** | *0.251* | *0.199* | *0.187* | *0.137* |
| | | 70 | **0.156** | *0.220* | *0.204* | *0.196* | *0.167* | | | 70 | **0.086** | *0.181* | *0.155* | *0.145* | *0.098* |
| | Auc | 30 | **0.839** | *0.782* | *0.785* | *0.828* | *0.802* | | Auc | 30 | **0.886** | *0.821* | *0.851* | *0.847* | *0.868* |
| | | 70 | **0.845** | *0.790* | *0.800* | *0.837* | *0.836* | | | 70 | **0.898** | *0.837* | *0.861* | *0.869* | 0.895 |
| | Cvg | 30 | **4.545** | *7.396* | *7.084* | *6.842* | *5.397* | | Cvg | 30 | **3.367** | *8.657* | *7.549* | *6.463* | *5.052* |
| | | 70 | **4.430** | *6.091* | *5.952* | *5.685* | 4.490 | | | 70 | **3.348** | *6.522* | *6.419* | *6.130* | *3.694* |
| | Ap | 30 | **0.573** | *0.511* | *0.547* | *0.540* | *0.540* | | Ap | 30 | **0.638** | *0.578* | 0.631 | *0.609* | *0.611* |
| | | 70 | **0.614** | *0.556* | *0.597* | *0.567* | *0.600* | | | 70 | 0.672 | *0.622* | **0.675** | *0.653* | *0.653* |
| Soc | Rkl | 30 | **0.139** | *0.264* | *0.252* | *0.202* | *0.175* | Enr | Rkl | 30 | **0.149** | *0.224* | *0.179* | *0.172* | *0.173* |
| | | 70 | **0.136** | *0.222* | *0.208* | *0.194* | *0.141* | | | 70 | **0.129** | *0.206* | *0.170* | *0.162* | *0.152* |
| | Auc | 30 | **0.826** | *0.782* | *0.804* | *0.808* | **0.826** | | Auc | 30 | **0.853** | *0.775* | *0.820* | *0.830* | *0.843* |
| | | 70 | **0.840** | *0.790* | *0.816* | *0.816* | **0.840** | | | 70 | **0.872** | *0.794* | *0.829* | *0.839* | *0.849* |
| | Cvg | 30 | **5.816** | *9.415* | *9.550* | *8.637* | *6.944* | | Cvg | 30 | **19.01** | *26.61* | *22.72* | *21.41* | *20.42* |
| | | 70 | **5.750** | *8.229* | *8.227* | *7.638* | **5.750** | | | 70 | **17.16** | *24.37* | *21.90* | *19.53* | *18.17* |
| | Ap | 30 | **0.601** | *0.533* | *0.569* | *0.563* | *0.565* | | Ap | 30 | **0.589** | *0.537* | *0.580* | *0.582* | *0.580* |
| | | 70 | **0.625** | *0.576* | *0.606* | *0.589* | *0.590* | | | 70 | **0.635** | *0.569* | *0.585* | *0.601* | *0.607* |
| Cor | Rkl | 30 | **0.285** | *0.328* | *0.332* | *0.308* | *0.331* | Ima | Rkl | 30 | **0.200** | *0.247* | *0.224* | *0.204* | *0.220* |
| | | 70 | **0.194** | *0.283* | *0.248* | *0.250* | 0.199 | | | 70 | **0.187** | *0.207* | *0.195* | 0.188 | *0.197* |
| | Auc | 30 | **0.714** | *0.677* | *0.673* | *0.693* | *0.670* | | Auc | 30 | **0.801** | *0.796* | *0.796* | *0.795* | 0.800 |
| | | 70 | **0.805** | *0.688* | *0.747* | *0.749* | 0.801 | | | 70 | **0.813** | 0.812 | 0.812 | 0.811 | 0.810 |
| | Cvg | 30 | **211.84** | *281.69* | *275.41* | *233.83* | *240.17* | | Cvg | 30 | **1.070** | *1.176* | *1.160* | *1.103* | *1.131* |
| | | 70 | **151.23** | *214.27* | *212.84* | *190.83* | *160.59* | | | 70 | **1.025** | *1.105* | *1.066* | 1.030 | *1.040* |
| | Ap | 30 | **0.174** | *0.147* | *0.158* | *0.166* | *0.165* | | Ap | 30 | **0.760** | *0.745* | *0.745* | *0.752* | *0.744* |
| | | 70 | **0.192** | *0.150* | *0.176* | *0.185* | 0.188 | | | 70 | **0.777** | *0.766* | *0.768* | 0.772 | *0.770* |

(a) 第 1 组样例图.

(b) 第 2 组样例图.




(c) 第 1 组局部标记关系 (d) 第 2 组局部标记关系 (e) 全局标记关系

**图 2.1** Image 数据集中两个局部组的示例图以及对应的 5 × 5 标记关系矩阵。标记依次为沙漠，山，海，日落，和树。

相较于 GLOCAL，其它对比方法所关注的只是部分方面：MBR 没有考虑任何标记相关性，且是一个单一学习器，因此总是取得最差或者次差的结果。因为 MBR 为每个标记单独训练分类器，而在不考虑标记相关性的情况下，对于那些正样本很少的标记（例如 Coral5k 数据集中，超过 100 个标记的正样本数少于 10），单一学习器很难建立很好的模型。LEML 利用了低秩结构，但是并未显式利用标记关系；MMLLOC 仅仅学了局部标记关系；而 ML-LRC 仅考虑全局标记关系。因而整体上同时学习、利用了全局和局部标记关系的 GLOCAL 更优。除此之外，MBR 和 MMLLOC 这类两阶段学习方法对预测测试样本标记并不是很有效（见表 2.3）。因为在恢复缺失标记阶段，MAXIDE 会引入额外的恢复误差（见表 2.2），这些误差将会传播到训练学习器的过程中。

图 2.1展示了 GLOCAL 学到的标记关系。可以看到，局部标记关系在不同的局部数据可以大不相同。对于第一组来说，标记 “日落” 与 “沙漠”、“海” 非常正相关，如图 2.1(c) 所示。这符合第一组（如图 2.1(a) 所示）的图片关系。

**表 2.4** 标记完全情况下小聚类上的结果 (样本数小于总样本数的 5%)。粗体表示最好的结果，斜体表示 GLOCAL 显著优于对比方法 (t 检验，$\alpha = 0.05$)。

| | | GLOCAL | GLObal | loCAL | | | GLOCAL | GLObal | loCAL |
|---|---|---|---|---|---|---|---|---|---|
| Art | Rkl | **0.130±0.005** | *0.137±0.003* | *0.137±0.002* | Bus | Rkl | **0.040±0.003** | 0.040±0.002 | 0.040±0.002 |
| | Auc | **0.870±0.005** | *0.863±0.003* | *0.863±0.002* | | Auc | **0.958±0.003** | 0.958±0.003 | 0.958±0.003 |
| | Cvg | **5.197±0.065** | *5.286±0.046* | *5.286±0.046* | | Cvg | **2.528±0.040** | 2.529±0.035 | 2.528±0.040 |
| | Ap | **0.631±0.011** | *0.602±0.013* | *0.602±0.010* | | Ap | **0.886±0.003** | *0.882±0.002* | *0.882±0.002* |
| Com | Rkl | **0.092±0.002** | *0.095±0.002* | *0.095±0.002* | Edu | Rkl | **0.097±0.002** | *0.101±0.002* | *0.101±0.002* |
| | Auc | **0.908±0.001** | *0.905±0.002* | *0.905±0.002* | | Auc | **0.903±0.002** | *0.899±0.002* | *0.899±0.002* |
| | Cvg | **4.364±0.055** | *4.482±0.032* | *4.486±0.040* | | Cvg | **4.672±0.051** | *4.803±0.033* | *4.805±0.036* |
| | Ap | **0.678±0.005** | 0.677±0.003 | 0.676±0.003 | | Ap | **0.624±0.005** | *0.605±0.003* | *0.605±0.003* |
| Ent | Rkl | **0.086±0.003** | *0.091±0.002* | *0.091±0.002* | Hea | Rkl | **0.053±0.004** | 0.054±0.002 | 0.054±0.003 |
| | Auc | **0.914±0.002** | *0.909±0.002* | *0.909±0.002* | | Auc | **0.947±0.003** | 0.945±0.003 | 0.946±0.003 |
| | Cvg | **2.709±0.059** | *2.817±0.027* | *2.797±0.035* | | Cvg | **3.504±0.041** | 3.508±0.036 | 3.506±0.049 |
| | Ap | **0.759±0.006** | *0.748±0.003* | *0.749±0.004* | | Ap | **0.812±0.006** | 0.810±0.004 | 0.810±0.004 |
| Rec | Rkl | **0.118±0.002** | *0.124±0.002* | *0.124±0.002* | Ref | Rkl | **0.054±0.004** | *0.060±0.002* | *0.061±0.003* |
| | Auc | **0.872±0.004** | 0.871±0.003 | 0.870±0.003 | | Auc | **0.946±0.004** | *0.940±0.003* | *0.939±0.004* |
| | Cvg | **3.700±0.042** | 3.704±0.033 | 3.700±0.037 | | Cvg | **2.325±0.060** | *2.552±0.043* | *2.559±0.057* |
| | Ap | **0.672±0.005** | 0.670±0.004 | 0.670±0.004 | | Ap | **0.783±0.005** | *0.739±0.004* | *0.739±0.004* |
| Soc | Rkl | **0.113±0.005** | *0.126±0.003* | *0.126±0.005* | Enr | Rkl | **0.105±0.005** | *0.117±0.002* | *0.119±0.003* |
| | Auc | **0.887±0.005** | *0.874±0.003* | *0.874±0.004* | | Auc | **0.895±0.004** | *0.883±0.004* | *0.881±0.004* |
| | Cvg | **5.208±0.059** | *5.554±0.047* | *5.553±0.053* | | Cvg | **17.511±1.231** | *19.440±0.833* | *19.372±0.915* |
| | Ap | **0.711±0.005** | *0.670±0.004* | *0.670±0.005* | | Ap | **0.706±0.007** | *0.685±0.005* | *0.673±0.005* |
| Cor | Rkl | **0.160±0.002** | *0.163±0.002* | *0.163±0.002* | Ima | Rkl | **0.190±0.004** | *0.197±0.003* | *0.199±0.004* |
| | Auc | **0.840±0.002** | *0.837±0.002* | *0.837±0.002* | | Auc | **0.810±0.003** | *0.803±0.003* | *0.801±0.003* |
| | Cvg | **128.40±1.30** | *130.84±1.01* | *131.13±1.21* | | Cvg | **1.027±0.027** | *1.064±0.015* | *1.066±0.021* |
| | Ap | **0.214±0.005** | 0.212±0.003 | 0.212±0.003 | | Ap | **0.771±0.005** | *0.764±0.003* | *0.763±0.004* |

而且在第一组中，有的图片中 “树” 会和 “沙漠” 同时出现（如图 2.1(a) 中的第一张和最后一张图），所以学到的标记关系有一定相关性。而在第二组标记关系矩阵图 2.1(b) 中，“山” 和 “海” 经常同时出现，“树” 和 “沙漠” 则很少同时出现（如图 2.1(b) 所示）。图 2.1(e) 展示了学到的全局标记关系：“海” 和 “日落”，“山” 和 “树” 正相关，而 “沙漠” 和 “海”，“树” 和 “沙漠” 负相关，这和常识一致。

为了进一步验证利用全局和局部标记关系的有效性，研究了 GLOCAL 的两个退化版本：(1) GLObal，仅仅利用全局标记关系（只有全局标记流形正则化项）；(2) loCAL，仅利用局部标记关系（只有局部标记流形正则化项）。由于局部数据组是通过聚类算法得到的，因而每组并非包含相同数量的样本。对于有的数据集，最大的聚类包括超过 40% 的样本，而一些比较小的聚类仅仅不到 5% 的样本。这种情况下，全局标记关系可能会由大组的局部标记关系主导 (命题 2.2)，导致与退化版本的差别不是很显著。因此这里着重关注它们在小聚类上的性能，如表 2.4所示。即使是 GLOCAL 的退化版本，在一些数据集上（如

Health 数据集）也可以取得较好的结果。而同时考虑全局和局部标记关系则可在大部分数据集上取得更好的结果。

### 2.3.3 算法收敛性

本节通过实验验证了 GLOCAL 算法的收敛性。图 2.2展示了在 Arts、Business、Enron 和 Image 数据集上，目标函数值随着迭代轮数的变化曲线。可见目标函数值在数十轮后就收敛了。在其它数据集上也可得到类似结果。

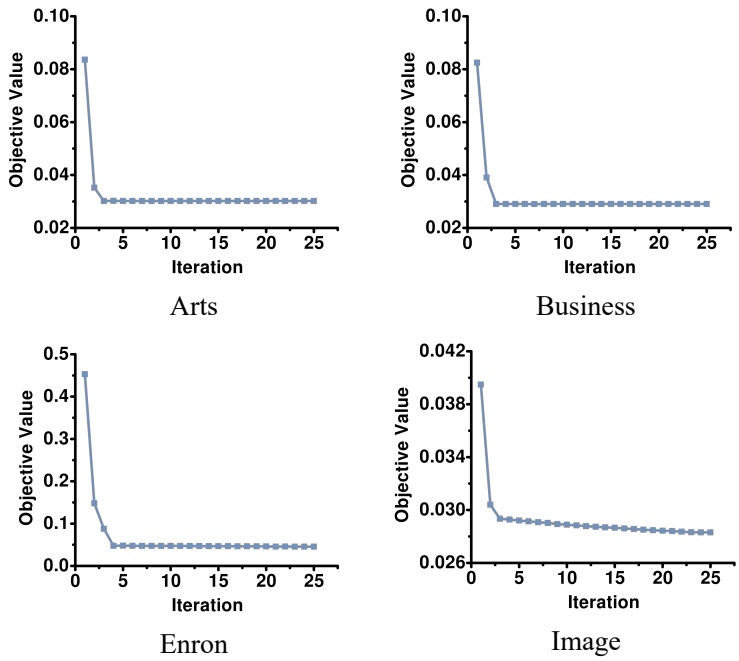



图 2.2 GLOCAL 在 Arts、Business、Enron 和 Image 数据集上的收敛性。

## 2.4 本章小结

在本工作中，提出了一种多标记关系学习算法 GLOCAL 以解决多标记训练集标记增广问题。它能够通过学习隐标记表示及优化标记流形同时恢复缺失标记，训练分类器，探索与利用全局和局部标记关系。与之前的工作相较，它同时利用全局和局部标记关系，且直接通过数据学习标记拉普拉斯矩阵，而不需要其它关于标记关系的先验知识。除此之外，GLOCAL 为标记完整情况和标记缺失情况的多标记学习提供了一个统一的解决方案。实验结果验证了 GLOCAL 方法的有效性。

本工作已总结成文：

- Y. Zhu and J. T. Kwok and Z.-H. Zhou. Multi-Label Learning with Global and Local Correlation. **IEEE Transactions on Knowledge and Data Engineering**, 2018, 39(6):1081–1094.（CCF-A 类期刊）

# 第三章　测试集标记增广学习

## 3.1　引言

测试集标记增广源于训练集上未观察到的新标记，要求学习系统不仅能够预测已知的目标标记，还能够预测出测试样本中是否包含这些新标记。新标记对应的语义可能在训练集中已经存在，但从未被标记出；也可能由数据流上动态增广的样本引入。前者是一种静态的测试集标记增广，对应真实标记集合固定但真实标记矩阵整列缺失的情况，此时无法利用低秩结构恢复出标记矩阵中的缺失标记。后者是一种动态的测试集标记增广，真实标记集合随着数据流中出现的新标记不断增加。一般不知道新标记何时会出现，且在多标记设置下，新标记经常会和一些已知标记同时出现在同一个样本上，因此很难将这些新标记样本与那些没有新标记但有相同已知标记的样本区分开来。本章主要考虑多标记的情况，分别对静态和动态测试集标记增广问题展开研究。

静态测试集标记增广问题可以看成是一种特殊的标记缺失问题。例如，图片标注任务要求标注图片是否有建筑和车辆，那么标注者则只关心建筑和车辆的语义，而对图片出现的天空、树木等语义可能不会进行标注，导致这些标记在训练集中完全没有出现过，即新标记。与第二章训练集标记增广相比，第二章考虑的是部分标记缺失情况，其中某些样本上缺失的标记在另一些样本上可以被观察到。假设训练集的标记包含了所有训练集中出现的语义，即真实标记集合等于所有训练样本标记集合的并集。此时，那些缺失标记可以利用低秩结构进行恢复。但是在本章测试集标记增广设置下，真实标记集合是训练集中所有观察到的标记集合与与训练集中隐藏的新标记集合的并集，相当于真实标记矩阵整列缺失，无法利用矩阵的低秩结构进行补全。本章的第一个工作旨在从训练数据中检测多个新标记并在测试集上进行预测，为简化工作，假设训练集中的已知标记在每个训练样本上没有缺失。

直接在多标记学习框架下预测多个新标记是非常困难的：每个样本仅仅由一个特征向量表示，却与多个标记关联，那些新标记样本可能同时有已知标记和新标记，很难将它们与仅有相同已知标记的样本区分开。即使能够区分，在

静态测试集标记增广学习设置下，新标记对应的语义在训练集中可能多次出现，因此无法通过判断特征是否在训练集上非常少见，来判断样本是否与新标记关联。即使新标记确实出现频次少、通过异常检测能够检测出，由于样本只由单一特征向量表示，表示能力有限，难以区分不同的新标记。

为了解决多新标记预测问题，考虑利用表达能力更强的多示例多标记学习框架[68]。其中每个对象由一个多示例包表示，包中的每个示例包含某一个语义概念，包的标记即为包中示例标记的并集。在训练过程中，只有包的标记可以被观察到，而包中具体每个示例对应的标记是未知的。以图像分类任务为例，一个包对应的是一幅图像，而包中的示例是图像分割成的各个小块所对应的特征向量。这样一来在多示例多标记学习框架下，只要能够预测每个示例的标记，并找到不属于已知标记的示例，那么这个示例以及包含这个示例的包存在新标记。进一步地，相同标记或者语义对应的示例更加相似，反之亦然。根据这一性质，能够区分开不同的新标记。

已有的多示例多标记学习 [65, 70, 71] 主要假设所有的包属于固定的目标标记集合，而不考虑新标记的存在。而能够适应新标记对建立鲁棒的学习系统来说是非常重要的 [72]，不仅仅是从训练集中辨识出新标记，而且要能够预测出测试样本中的那些新标记。Pham 等人 [73] 最近提出了一种多示例多标记概率模型能够辨识出新标记样本。他们假设所有新标记样本都有相同的新类标记，即只检测一个新标记。通过最大化对数似然学习一个概率模型，并利用动态规划方法将每个包 $i$ 估计似然的复杂度从 $O((c_i+1)^{z_i})$ 降到 $O((c_i+1)2^{(c_i+1)}z_i)$。其中，$c_i$ 是观察到的标记数目，$z_i$ 是包中示例数目。在很多应用中，新标记样本可能属于不同的标记，且学习器不仅仅要检测出样本是否有新标记，而且还要能够区分它是否属于不同的新类别。例如，在鸟类声音识别任务中，人们希望能够区分出不同的新叫声，这可能表示不同种类的鸟类新迁徙到声音采集的地区。但是将 Pham 等人的方法 [73] 直接扩展到多个新标记建模是非常困难的。在对每个包进行似然估计时，由于只有一个新标记的设定，一共只有 2 种情况需要考虑：(1) 这个包只有已知标记；(2) 这个包同时有已知标记和新标记。但是潜在的新标记数 $k>1$ 时，要考虑的情况将复杂得多：所有 $k$ 个新标记的子集在计算似然的时候都要被考虑进去，即一共有 $2^k$ 种情况。且当标记数目比较多的时候，算法复杂度非常高。

在静态测试集标记增广学习中，训练集和真实标记集合都是固定的。而还有很多应用会不断产生新的样本，数据流中的新样本可能会带来新的语义产生

新的标记，即真实标记集合会随着新标记的出现不断增大。具体地，假设样本呈数据流形式到来，除了一开始的训练集样本是有完整标记的数据，数据流中的样本在训练中是无标记的。在这样的问题中，学习系统必须能够检测新标记、并对出现的新标记建立模型。这种动态测试集标记增广学习最困难的部分是检测出新标记样本。

当把动态测试集标记增广学习中的检测新标记问题当做是异常检测问题来处理时，即把异常样本当做是新标记样本，可以利用很多现有的异常检测方法。例如 OC-SVM [21] 为每个已知类训练一个边界，把落到边界外的样本当做是异常样本；iForest [22] 则把那些落到密度稀疏区域的样本当做是异常样本。但是在多标记学习问题中，新标记和已知标记可能同时出现在一个新样本上。这个新标记样本由于同时存在的已知标记，可能会被已有的分类器以很高置信度预测为某个已知标记，也可能会落在某个已知标记相关样本密度稠密的区域。这使得在多标记设置下，OC-SVM 和 iForest 等异常检测方法很难仅通过区分特征上的不同，来将那些同时有新标记和已知标记的样本与那些没有新标记但是有相同已知标记的样本区分开来。

当把动态测试集标记增广学习当做特殊的增量学习 (E-IL 和特殊 C-IL 的结合) 看待时，由于现有 C-IL 只能处理多类问题，而不能处理一个样本同时与多个标记相关联的情况，需要改变 E-IL+C-IL 使得它能够处理动态测试集标记增广学习：将每一种可能的标记组合当成是一个独立的类别，从而把多标记问题转化为多类学习问题 [20]。但是这种方法有两个非常严重的问题：(1) 检测到的新类可能并不意味着一个新的标记，而可能是一些在训练集上没有出现过的已知标记的组合；(2) 当标记集合比较大时将导致可能的标记组合数即转化为多类学习后的类别数非常大，有可能一些类只有非常少量的相关样本，使得训练任务变得非常困难。也因此，这种简单的转换在实际问题中并不可行。

为了迎接以上挑战，针对静态测试集标记增广学习问题，提出了 DMNL (Discovering Multiple New Labels) 方法，将多示例多标记学习下的多个新标记学习问题形式化为一个非负正交约束下的优化问题，通过最小化多示例包级损失函数和聚类正则化项进行高效求解；针对动态测试集标记增广学习问题，提出了 MuENL (Multi-label learning with Emerging New Labels) 方法，设计了基于特征与已知标记上的预测值的新标记检测器以及鲁棒的更新模型以用于整合检测到的新标记，且能够容忍新标记检测的误差。最后通过大量实验，验证了 DMNL 方法和 MuENL 方法分别在动态和静态测试集标记增广学习上的有效性。

## 3.2 本文方法

### 3.2.1 DMNL

将静态测试集标记增广学习形式化成一种特殊的多示例多标记学习，在示例级的标注及包级的预测任务考虑多个新标记，设计了一种新的判别式方法DMNL(Discovering Multiple New Labels)，最小化了包级的损失函数（在已知标记上的误差）并通过聚类正则化项利用了所有包中示例之间的示例级聚类结构关系。

#### 形式化

令 $\mathcal{X}$ 表示样本特征空间，$\boldsymbol{v}=\{1,\cdots,c\}$ 是大小为 $c$ 的已知标记集合[①]。定义 $D=\{(\boldsymbol{B}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{B}_m,\boldsymbol{y}_m)\}$ 是大小为 $m$ 的训练集合，其中 $\boldsymbol{B}_i=\{\boldsymbol{x}_{i,1},\cdots,\boldsymbol{x}_{i,z_i}\}$ 是包含 $z_i$ 个样本的包，每个样本 $\boldsymbol{x}_{i,j}\in\mathcal{X}$；$\boldsymbol{y}_i=[y_{i,1};\cdots;y_{i,c}]\in\{0,1\}^c$ 是这个包上观察到的包标记向量。如果第 $i$ 个包属于标记 $j$，则 $y_{i,j}=1$，否则 $y_{i,j}=0$。

假定有 $k$ 个新标记，则真实标记集合为 $\hat{\boldsymbol{v}}=\boldsymbol{v}\cup\overline{\boldsymbol{v}}$，其中 $\overline{\boldsymbol{v}}=\{c+1,\cdots,c+k\}$ 代表 $k$ 个新标记。令 $\hat{\boldsymbol{y}}_{i,j}=[\hat{y}_{i,j,1};\cdots;\hat{y}_{i,j,c+k}]\in\{0,1\}^{c+k}$ 表示第 $i$ 个包中第 $j$ 样本的标记向量。我们沿袭了多示例多标记学习中的一个常用的设定，即每个样本仅属于一个标记[70, 73, 74]。因此对每个样本 $\boldsymbol{x}_{i,j}$ 有 $\sum_{l=1}^{c+k}\hat{y}_{i,j,l}=1$。需要注意的是，这些新标记以及样本级的标记在训练的时候是不能被观察到的，无法使用。

令 $\boldsymbol{A}=[\boldsymbol{B}_1,\cdots,\boldsymbol{B}_m]$ 是由所有包中的所有示例拼成的示例矩阵，$\boldsymbol{a}_i$ ($i\in\{1,\cdots,n\}$) 表示 $\boldsymbol{A}$ 的第 $i$ 列，对应一个示例。其中 $n=\sum_{i=1}^{m}z_i$ 是所有示例的总数。定义 $\hat{\boldsymbol{Y}}_i=[\hat{\boldsymbol{y}}_{i,1},\cdots,\hat{\boldsymbol{y}}_{i,z_i}]$ 为第 $i$ 个包对应的示例标记矩阵，$\tilde{\boldsymbol{Y}}=[\hat{\boldsymbol{Y}}_1,\cdots,\hat{\boldsymbol{Y}}_m]$ 表示所有示例标记矩阵。$\boldsymbol{A}$ 和 $\tilde{\boldsymbol{Y}}$ 的每一列分别代表一个示例和它所对应的标记向量。

另定义常用符号如下：$\boldsymbol{I}$ 代表单位矩阵；$\mathbf{1}$ 代表全 1 向量；$\circ$ 代表矩阵对应元素相乘，$\div$ 代表对应元素相除；tr(.) 代表矩阵的迹；$\vee$ 代表对应元素或运算；$\bigvee$ 代表并运算 $\bigvee_{i=1}^{n}\boldsymbol{a}_i=\boldsymbol{a}_1\vee\boldsymbol{a}_2\vee\cdots\vee\boldsymbol{a}_n$；$\Omega_c(.)$ 返回输入矩阵的前 $c$ 行；$\mathbb{I}(.)$ 代表指示函数，如果输入为真则返回 1，否则返回 0；Diag(.) 输出一个对角矩阵，其中，对角线元素对应输入向量中的元素。

[①] $\boldsymbol{v}$ 中数字表示标记的序号。

定义多示例多标记框架下的多个新标记学习问题如下：给定训练集 $D$（包括多示例包及相应已知标记），在多示例多标记学习框架下的新标记学习的任务是检测出训练集中每个包中所包含的多个新标记，并且能够为未知包预测已知标记和新标记。进一步，这个问题包括两个子任务：(1) 示例级的标注任务是学一个由示例到标记（包括已知标记和新标记）的映射 $f : \mathcal{X} \to \hat{\mathcal{Y}}$；(2) 包级预测任务是学一个由包到标记集合（同样包括已知标记和新标记）的映射 $\Psi : 2^{\mathcal{X}} \to 2^{\hat{\mathcal{Y}}}$。假设包的真实标记包括所有包中示例的真实标记。所以当任务 (1) 解决后，任务 (2) 可由 $\bigvee_{j=1}^{z_i} \hat{\boldsymbol{y}}_{i,j}$ 预测包的标记。

首先通过两个命题来介绍关于示例标记的两个性质，为简化符号，只用已知标记来阐述，而这两个性质对多个新标记的设定是成立的，只需要简单更改一下符号。由于多示例多标记学习中只有包级的标记能够被观察到，损失项的设计基于包级已知标记上的经验误差，即利用一个包中的所有示例去计算包上的损失。这和以往的多示例多标记学习方法不同：以往方法将每个标记上包中示例预测的最大值作为包上标记的预测值[70, 71, 75]，相当于为每个标记选最大预测值的示例作为整个包在相应标记的代表。但是包上的标记经常是由包中样本的某些结构决定的，而不是由某个单一示例决定[76]。这种关键的结构一般是不知道的，因此考虑了包中每个示例对包标记的贡献。此外，每个包上的标记是同等重要的，尽管它们在包中和不同数量的样本相关联，故而采用了再缩放策略 (Rescaling) [77]。这种策略被广泛应用于类别不平衡学习问题中去平衡每个标记的贡献。

**命题 3.1** $\boldsymbol{y}_i = \hat{\boldsymbol{Y}}_i \boldsymbol{\beta}_i$，其中 $\boldsymbol{\beta}_i = [\beta_{i,1}; \cdots ; \beta_{i,z_i}]$，$\beta_{i,j} = 1 / \sum_{l=1}^{c} (\mathbb{I}(\hat{y}_{i,j,l} = 1) \sum_{q=1}^{z_i} \hat{y}_{i,q,l})$。

在命题3.1中，每个包中的示例标记对包标记均有贡献，其中 $\beta_i$ 代表贡献的权重。为了平衡每个包标记的重要程度（因为它们可能在包中对应不同数量的示例），引入了再缩放策略 [77]。假定在一个包中有 $n_l$ 个示例有第 $l$ 个标记，那么它们每个的权重即为 $1/n_l$ (即假定每个相同标记示例权重相同) 。$\beta_{i,j}$ 表示的就是第 $i$ 个包中第 $j$ 个示例的权重。这个命题满足：$\boldsymbol{y}_i = \bigvee_{j=1}^{z_i} \hat{\boldsymbol{y}}_{i,j}$。例如给定一个包 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$ [①]，如果 $\boldsymbol{x}_1$ 属于标记 1，$\boldsymbol{x}_2$ 和 $\boldsymbol{x}_4$ 属于标记 2，$\boldsymbol{x}_3$ 属于标记 3，那么由命题3.1可得 $\boldsymbol{\beta}_i = [1, 0.5, 1, 0.5]^{\top}$。其中，$\boldsymbol{x}_2$ 和 $\boldsymbol{x}_4$ 共享一个标记，并对标记 2 有相同贡献。$\boldsymbol{\beta}_i$ ($i \in \{1, 2, \cdots, m\}$) 的集合记作 $\tilde{\beta} = \{\beta_1, \cdots, \beta_m\}$。

正则化项的设计利用了示例之间的聚类关系去正则化假设空间。其基本

[①]为了简化标记，这里省略了包的下标 $i$。

思想是：在示例级上，属于相同的标记的示例在特征空间应该更为相似，应属于同一个聚类。相对的，不属于同一标记的示例，则应分属不同的聚类。直观上来看，利用这种聚类关系可以大大降低搜索复杂度。假定有两个包 $\{\boldsymbol{x}_{1,1},\cdots,\boldsymbol{x}_{1,n_1}\}$ 和 $\{\boldsymbol{x}_{2,1},\cdots,\boldsymbol{x}_{2,n_2}\}$，包上的标记均为 $\{1,2\}$。如果不考虑任何结构关系，对于样本标注任务，一共有 $(2^{n_1}-2)(2^{n_2}-2)$ 种可能的样本标记组合。相反，如果知道样本的聚类结构，如 $\{\boldsymbol{x}_{1,1},\boldsymbol{x}_{1,2},\boldsymbol{x}_{2,1},\boldsymbol{x}_{2,2}\}$ 属于同一个聚类，剩余的样本属于另一个聚类，那么可能的样本标记组合将只有 2 种：$\boldsymbol{x}_{1,1},\boldsymbol{x}_{1,2},\boldsymbol{x}_{2,1},\boldsymbol{x}_{2,2}$ 同属于标记 1(标记 2)，其它样本同属于标记 2(标记 1)。

为了利用聚类结构，假设每个标记存在一个示例原型，属于相同标记的示例将离该标记对应的示例原型很近，而与其它标记对应的原型距离很远。这即是聚类假设。

**命题 3.2** 给定标记 $l$ 的示例原型 $\boldsymbol{p}_l$，示例标记矩阵 $\tilde{\boldsymbol{Y}}$ 是式 (3.1) 的解[78]：

$$\min_{\boldsymbol{G}} \sum_{l=1}^{c}\sum_{i=1}^{n} G_{l,i}\|\boldsymbol{a}_i-\boldsymbol{p}_l\|^2 : \boldsymbol{G}\in\{0,1\}^{c\times n}, \boldsymbol{G}\boldsymbol{G}^\top=\boldsymbol{S}, \tag{3.1}$$

其中 $\boldsymbol{S}=\mathrm{Diag}(\boldsymbol{G}\boldsymbol{1})$。

由于每个示例属于一个标记，所以在式 (3.1) 中，加入了正交约束 $\boldsymbol{G}\boldsymbol{G}^\top=\boldsymbol{S}$，且 $\boldsymbol{S}$ 中的非 0 元素代表每个标记正示例的个数。

基于命题 3.1，设计了包级的损失项，考虑了包中所有示例的贡献。具体地，采用已知包标记上的平方误分损失：$\sum_{i=1}^{m}\|\boldsymbol{y}_i-\Omega_c(\hat{\boldsymbol{Y}}_i)\boldsymbol{\beta}_i\|^2$.

基于命题 3.2，设计了聚类正则化项。由于原型 $\boldsymbol{p}$ 与每个标记的真实分布有关，是未知的，优化过程必须避免直接计算 $\boldsymbol{p}$。通过定义 $\boldsymbol{H}=\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}}$，对式 (3.1) 的优化可以改写为式 (3.2) [79]。

$$\max_{\boldsymbol{H}} \mathrm{tr}(\boldsymbol{H}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{H}^\top) : \boldsymbol{H}\boldsymbol{H}^\top=I, \boldsymbol{H}\geq 0. \tag{3.2}$$

进而聚类正则化项定义为 $-\mathrm{tr}((\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}})\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}})^\top$。

将损失项和正则化项整合到一起，并加入非负正交约束，即可得优化任务如式 (3.3)：

$$\begin{aligned}\min_{\tilde{\boldsymbol{Y}}} \quad & \sum_{i=1}^{n}\|\boldsymbol{y}_i-\Omega_c(\hat{\boldsymbol{Y}}_i)\boldsymbol{\beta}_i\|^2-\lambda\,\mathrm{tr}((\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}})\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}})^\top, \\ \text{s.t.} \quad & (\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}})(\boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}})^\top=\boldsymbol{I},\ \ \tilde{\boldsymbol{Y}}\in\{0,1\}^{(c+k)\times n},\end{aligned} \tag{3.3}$$

其中 $\lambda$ 是一个权衡参数。

考虑归纳学习设定，且需要把式 (3.3) 从复杂困难的整数优化放松到一个相对简单的 $[0,1]^{(c+k)\times n}$ 上的连续优化问题。因此，定义 $W=[\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{c+k}]$ 为需要学习的变量，并定义 $g_l(\boldsymbol{x},\boldsymbol{w}_l)=\exp(\boldsymbol{w}_l^\top\boldsymbol{x})/\sum_{l'=1}^{c+k}\exp(\boldsymbol{w}_{l'}^\top\boldsymbol{x})$ 为对样本 $\boldsymbol{x}$ 在标记 $l$ 上的预测函数，且它的值域为 $[0,1]$。令 $\boldsymbol{g}=[g_1;\cdots;g_{c+k}]$ 可得 $\boldsymbol{g}(\boldsymbol{x},\boldsymbol{W})=[g_1(\boldsymbol{x},\boldsymbol{w}_1);\cdots;g_{c+k}(\boldsymbol{x},\boldsymbol{w}_{c+k})]$ 且 $\boldsymbol{g}(\boldsymbol{B},\boldsymbol{W})=[\boldsymbol{g}(\boldsymbol{x}_1,\boldsymbol{W}),\cdots,\boldsymbol{g}(\boldsymbol{x}_z,\boldsymbol{W})],\boldsymbol{x}_i\in\boldsymbol{B}$，则 $\tilde{\boldsymbol{Y}}$ 可由 $\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W})$ 建模。将 $\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W})$ 代替式 (3.3) 的 $\tilde{\boldsymbol{Y}}$，即可得 DMNL 最终的优化式 (3.4)：

$$\begin{aligned}\min_{\boldsymbol{W}}\quad & \sum_{i=1}^{m}\|\boldsymbol{y}_i-\Omega_c(\boldsymbol{g}(\boldsymbol{B}_i,\boldsymbol{W}))\boldsymbol{\beta}_i\|^2-\lambda\,\mathrm{tr}((\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W}))\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W}))^\top),\\ \text{s.t.}\quad & (\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W}))(\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W}))^\top=\boldsymbol{I}.\end{aligned}\tag{3.4}$$

**评注 3.1** 当考虑流形结构[54]，可以用 $+\lambda\,\mathrm{tr}((\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W})\boldsymbol{L}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W})^\top)$ 替换式 (3.4) 中的聚类正则化项 $-\lambda\,\mathrm{tr}((\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W}))\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W}))^\top)$。其中，$L$ 是样本拉普拉斯矩阵，使得在相似示例上的预测值相似。流形假设下的优化与聚类假设下的优化方法相似。

**样本级标注：**给定学到的 $\boldsymbol{W}$，用最大的预测值对应的标记作为 $\boldsymbol{x}_{i,j}$ 的预测标记

$$\hat{y}_{i,j,l}=\begin{cases}1, & l=\arg\max_{l'} g_{l'}(\boldsymbol{x}_{i,j},W),l'\in\{1,\cdots,c+k\};\\ 0, & \text{otherwise.}\end{cases}\tag{3.5}$$

**包级标记预测：**第 $i$ 个包的预测标记预测为 $\bigvee_{j=1}^{z_i}\hat{\boldsymbol{y}}_{i,j}$。

## 优化求解

对式 (3.4) 的优化并不直接，原因主要有二：(1)$\tilde{\boldsymbol{\beta}}$ 和 $\boldsymbol{S}$ 是基于 $\tilde{\boldsymbol{Y}}$ 的值（通过式 (3.5) 得到的离散的 0-1 矩阵）；(2) 在非负正交约束中有一个关于 $\boldsymbol{W}$ 的复杂方程。为了解决 (1)，采用一种交替优化策略：固定 $\tilde{\boldsymbol{\beta}}$ 和 $\boldsymbol{S}$ 更新 $\boldsymbol{W}$；然后根据更新后的 $\boldsymbol{W}$ 得到 $\tilde{\boldsymbol{\beta}}$ 和 $\boldsymbol{S}$：给定 $\boldsymbol{W}$ 根据式 (3.5) 得到 $\tilde{\boldsymbol{Y}}$，然后根据命题3.1 计算 $\tilde{\boldsymbol{\beta}}$，$\boldsymbol{S}$ 由 $S=\mathrm{diag}(\tilde{\boldsymbol{Y}}\boldsymbol{1})$ 得到。为了解决 (2)，定义一个新的变量 $\hat{\boldsymbol{H}}=\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A},\boldsymbol{W})$。这样，约束可以简化表示为 $\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top=\boldsymbol{I},\hat{\boldsymbol{H}}\geq 0$。令

$\hat{\boldsymbol{H}} = [\boldsymbol{H}_1, \cdots, \boldsymbol{H}_m]$，其中 $\boldsymbol{H}_i = \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{B}_i, W)$，则式 (3.4) 可以重写为式 (3.6)：

$$\min_{\boldsymbol{W},\hat{\boldsymbol{H}}} \phi(\boldsymbol{W}) + \psi(\hat{\boldsymbol{H}}) : \hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top = \boldsymbol{I}, \hat{\boldsymbol{H}} = \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{W}), \tag{3.6}$$

其中 $\phi(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{y}_i - \Omega_c(\boldsymbol{g}(\boldsymbol{B}_i, \boldsymbol{W}))\boldsymbol{\beta}_i\|^2$，$\psi(\hat{\boldsymbol{H}}) = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{y}_i - \Omega_c(\boldsymbol{H}_i\boldsymbol{S}^{\frac{1}{2}})\boldsymbol{\beta}_i\|^2 - \lambda\,\mathrm{tr}(\hat{\boldsymbol{H}}\boldsymbol{A}^\top\boldsymbol{A}\hat{\boldsymbol{H}}^\top)$。这里，在 $\phi(\boldsymbol{W})$ 和 $\psi(\hat{\boldsymbol{H}})$ 中均考虑误分类误差，从而得到更好的结果和更快的收敛速度。

为了求解式 (3.6)，采用增广拉格朗日优化框架 [80]，其中式 (3.6) 的增广拉格朗日为：

$$\mathcal{L}(\boldsymbol{W}, \hat{\boldsymbol{H}}, \boldsymbol{\Lambda}) = \phi(\boldsymbol{W}) + \psi(\hat{\boldsymbol{H}}) + \frac{\rho}{2}\|\hat{\boldsymbol{H}} - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{W}) + \boldsymbol{\Lambda}\|_F^2 + \zeta,$$

其中 $\|.\|_F$ 是矩阵的 F 范数，$\boldsymbol{\Lambda}$ 是对偶变量，$\rho$ 是惩罚参数，$\zeta$ 是一个常数，在优化过程中可以被省去。解式 (3.6) 等价于解如下优化问题：

$$\min_{\boldsymbol{W},\hat{\boldsymbol{H}},\boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{W}, \hat{\boldsymbol{H}}, \boldsymbol{\Lambda}) : \hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top = \boldsymbol{I}, \hat{\boldsymbol{H}} \geq 0. \tag{3.7}$$

分别关于 $\boldsymbol{W}$、$\hat{\boldsymbol{H}}$ 和 $\Lambda$，迭代地优化式 (3.7)。令 $\boldsymbol{W}^{(t)}$、$\hat{\boldsymbol{H}}^{(t)}$ 和 $\boldsymbol{\Lambda}^{(t)}$ 表示第 $t$ 轮的解，更新规则如下：

$$\boldsymbol{W}^{(t+1)} = \arg\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}, \hat{\boldsymbol{H}}^{(t)}, \boldsymbol{\Lambda}^{(t)}); \tag{3.8}$$

$$\hat{\boldsymbol{H}}^{(t+1)} = \arg\min_{\hat{\boldsymbol{H}}\geq 0} \mathcal{L}(W^{(t+1)}, \hat{\boldsymbol{H}}, \boldsymbol{\Lambda}^{(t)}) : \hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top = I; \tag{3.9}$$

$$\boldsymbol{\Lambda}^{(t+1)} = \boldsymbol{\Lambda}^{(t)} + \hat{\boldsymbol{H}}^{(t+1)} - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{W}^{(t+1)}). \tag{3.10}$$

算法 3.1总结了 DMNL 算法。

### 更新 $\boldsymbol{W}$

为了能够高效求解式 (3.8)，采用随机梯度下降进行优化。解式 (3.8) 等价于最小化

$$\mathcal{L}_{\boldsymbol{W}} = \phi(\boldsymbol{W}) + \frac{\rho}{2}\|\hat{\boldsymbol{H}} - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{W}) + \boldsymbol{\Lambda}\|_F^2.$$

将 $\mathcal{L}_{\boldsymbol{W}} = \sum_{i=1}^{m}\mathcal{L}_{\boldsymbol{W}}^{(i)}$ 根据包进行拆解：$\phi(\boldsymbol{W})$ 可以自然写作 $\phi(\boldsymbol{W}) = \sum_{i=1}^{m}\phi_i(\boldsymbol{W})$，其中每个 $\phi_i(\boldsymbol{W}) = \|\boldsymbol{y}_i - \Omega_c(\boldsymbol{g}(\boldsymbol{B}_i, \boldsymbol{W}))\boldsymbol{\beta}_i\|^2$。回忆 $\boldsymbol{A} = [\boldsymbol{B}_1, \cdots, \boldsymbol{B}_m]$，$\hat{\boldsymbol{H}} = [\boldsymbol{H}_1, \cdots, \boldsymbol{H}_m]$，继而分解 $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_m]$，其中每个 $\Lambda_i$ 与包 $i$ 对应。然后

**算法 3.1** DMNL

**输入**: $D, \lambda, \rho$
**输出**: $W$

初始化 $\boldsymbol{W}$, $\hat{\boldsymbol{H}}$, $\boldsymbol{\Lambda}$, $\boldsymbol{S}$ and $\boldsymbol{\beta}_i, i = 1, \cdots, m$ ;
**repeat**:
  通过随机梯度下降算法更新 $\boldsymbol{W}$ ; \\ 解式 (3.8)
  通过式 (3.11) 更新 $\hat{H}$; \\ 解式 (3.9)
  通过式 (3.10) 更新 $\boldsymbol{\Lambda}$;
  根据式 (3.5) 预测 $\tilde{\boldsymbol{Y}}$;
  根据命题 3.1计算 $\boldsymbol{\beta}_i, i = 1, \cdots, m$;
  计算 $\boldsymbol{S} = \text{diag}(\tilde{\boldsymbol{Y}}\mathbf{1})$;
**until** 算法收敛或达到最大迭代轮数

$\frac{\rho}{2}\|\hat{\boldsymbol{H}} - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{W}) + \boldsymbol{\Lambda}\|_F^2$ 可以分解为 $\sum_{i=1}^{m} \frac{\rho}{2}\|\boldsymbol{H}_i - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{B}_i, \boldsymbol{W}) + \boldsymbol{\Lambda}_i\|_F^2$。基于以上分解可得：

$$\mathcal{L}_{\boldsymbol{W}}^{(i)} = \phi_i(\boldsymbol{W}) + \frac{\rho}{2}\|\boldsymbol{H}_i - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{g}(\boldsymbol{B}_i, \boldsymbol{W}) + \boldsymbol{\Lambda}_i\|_F^2.$$

令 $\boldsymbol{F}_i = \boldsymbol{g}(\boldsymbol{B}_i, \boldsymbol{W})$，给定 $(\boldsymbol{B}_i, \boldsymbol{y}_i, \boldsymbol{H}_i, \boldsymbol{\Lambda}_i)$，$\mathcal{L}_{\boldsymbol{W}}^{(i)}$ 关于 $\boldsymbol{W}$ 的梯度为：

$$\nabla_{\boldsymbol{W}} = \boldsymbol{B}_i\Big((\Delta_{\boldsymbol{W}1}\boldsymbol{\beta}_i^\top) \circ \boldsymbol{G}_i \circ \boldsymbol{J}_i\Big)^\top + \rho\boldsymbol{B}_i((\boldsymbol{S}^{-\frac{1}{2}}\Delta_{\boldsymbol{W}2}) \circ \boldsymbol{G}_i)^\top,$$

其中，$\Delta_{\boldsymbol{W}1} = [\boldsymbol{y}_i; 0^{k\times 1}] - (\boldsymbol{F}_i\boldsymbol{\beta}_i) \circ \boldsymbol{J}'$, $\Delta_{\boldsymbol{W}2} = \boldsymbol{H}_i + \boldsymbol{\Lambda}_i - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{F}_i$，$\boldsymbol{G}_i = \boldsymbol{F}_i \circ (\boldsymbol{F}_i - 1)$，$\boldsymbol{J}_i = [1^{c\times n_i}; 0^{k\times n_i}]$, $\boldsymbol{J}_i' = [1^{c\times 1}; 0^{k\times 1}]$。

用 $W^{(t)}$ 初始化 $W$，然后通过 $\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_i\nabla_{\boldsymbol{W}}$，$(i \in \{1, \cdots, m\})$ 对 $\boldsymbol{W}$ 进行更新。其中 $\eta_i$ 是步长。最后输出所有随机梯度下降算法中所有轮得到的 $\boldsymbol{W}$ 平均作为输出 $\boldsymbol{W}^{(t+1)}$。

**更新 $\hat{\boldsymbol{H}}$.**

式 (3.9) 的优化任务中有一个非负正交约束 $\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top = \boldsymbol{I}, \hat{\boldsymbol{H}} \geq 0$。$\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top = \boldsymbol{I}$ 是一种 Stiefel 流形[61]，因此尝试在该流形上解式 (3.9)。为了简化描述，这里省略了 $\hat{\boldsymbol{H}}^{(t)}$ 的角标。基于 Choi 的工作[81]，可以推出 $\hat{\boldsymbol{H}}$ 的更新规则如式 (3.11) 所示。

$$\hat{\boldsymbol{H}}^{(t+1)} = \hat{\boldsymbol{H}} \circ \frac{[\nabla_{\hat{\boldsymbol{H}}}]^+ + \hat{\boldsymbol{H}}([\nabla_{\hat{\boldsymbol{H}}}]^-)^\top\hat{\boldsymbol{H}}}{[\nabla_{\hat{\boldsymbol{H}}}]^- + \hat{\boldsymbol{H}}([\nabla_{\hat{\boldsymbol{H}}}]^+)^\top\hat{\boldsymbol{H}}}, \tag{3.11}$$

其中 $\nabla_{\hat{\boldsymbol{H}}}$ 是 $\mathcal{L}(\boldsymbol{W}^{(k+1)}, \hat{\boldsymbol{H}}, \Lambda^{(k)})$ 关于 $\hat{\boldsymbol{H}}$ 的梯度；$[\nabla_{\hat{\boldsymbol{H}}}]^+$ 和 $[\nabla_{\hat{\boldsymbol{H}}}]^-$ 满足 $[\nabla_{\hat{\boldsymbol{H}}}]^+ > 0$，$[\nabla_{\hat{\boldsymbol{H}}}]^- > 0$，且 $[\nabla_{\hat{\boldsymbol{H}}}] = [\nabla_{\hat{\boldsymbol{H}}}]^+ - [\nabla_{\hat{\boldsymbol{H}}}]^-$。

由于式 (3.11) 中的所有运算都是基于矩阵对应元素之间的操作，即对应元素乘法、除法，它可以根据包分解成多个小块。目标函数关于每个小块对应的梯度 $\nabla_{\hat{\boldsymbol{H}}_i}$ 为：

$$\nabla_{\hat{\boldsymbol{H}}_i} = \boldsymbol{S}^{\frac{1}{2}}\left(\Delta_{\boldsymbol{H}i1}\boldsymbol{\beta}_i^\top\right) \circ \boldsymbol{J}_i - \lambda \sum_{j=1}^{m} \boldsymbol{H}_j\boldsymbol{B}_j^\top\boldsymbol{B}_i + \rho\, \Delta_{\boldsymbol{H}i2},$$

其中 $\Delta_{\boldsymbol{H}i1} = (\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{H}_i\boldsymbol{\beta}_i) \circ \boldsymbol{J}_i' - [\boldsymbol{y}_i; 0^{k\times 1}]$，$\Delta_{\boldsymbol{H}i2} = \boldsymbol{H}_i - \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{F}_i + \boldsymbol{\Lambda}_i$。

为简化描述，令 $\boldsymbol{M}_i = \sum_{j=1}^{m} \boldsymbol{H}_j\boldsymbol{B}_j^\top\boldsymbol{B}_i$。因为 $\boldsymbol{M}_i$ 和 $\boldsymbol{\Lambda}_i$ 中有些元素可能是负数，而无法满足应用式 (3.11) 的非负条件。为解决这个问题，定义 $\Phi^+(\boldsymbol{M}) = (\text{abs}(\boldsymbol{M}) + \text{abs}(\boldsymbol{M}))/2$，$\Phi^-(\boldsymbol{M}) = (\text{abs}(\boldsymbol{M}) - \text{abs}(\boldsymbol{M}))/2$。其中 abs(.) 返回输入的绝对值。从而可得 $\Phi^+(\boldsymbol{M}) \geq 0$ 和 $\Phi^-(\boldsymbol{M}) \geq 0$ 满足 $\boldsymbol{M} = \Phi^+(\boldsymbol{M}) - \Phi^-(\boldsymbol{M})$。由此可得 $[\nabla_{\hat{\boldsymbol{H}}_i}]^+$ 和 $[\nabla_{\hat{\boldsymbol{H}}_i}]^-$，如下所示。

$$[\nabla_{\hat{\boldsymbol{H}}_i}]^+ = \left(\boldsymbol{S}^{\frac{1}{2}}\left((\boldsymbol{S}^{\frac{1}{2}}\hat{\boldsymbol{H}}\boldsymbol{\beta}_i) \circ \boldsymbol{J}_i'\right)\boldsymbol{\beta}_i^\top\right) \circ \boldsymbol{J}_i + \lambda\, \Phi^-(\boldsymbol{M}_i) + \rho(\boldsymbol{H}_i + \Phi^+(\boldsymbol{\Lambda}_i)),$$

$$[\nabla_{\hat{\boldsymbol{H}}_i}]^- = \left(\boldsymbol{S}^{\frac{1}{2}}[\boldsymbol{y}_i; 0^{k\times 1}]\boldsymbol{\beta}_i^\top\right) \circ \boldsymbol{J}_i + \lambda\, \Phi^+(\boldsymbol{M}_i) + \rho(\boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{F}_i + \Phi^-(\boldsymbol{\Lambda}_i)).$$

**初始化**

$\boldsymbol{W}$ 中的每个元素由 $[0, 1]$ 之间的随机数初始化。然后通过对所有样本进行 k-means 聚类得到 $(c + k)$ 个簇，用得到的簇指示矩阵（第 $i$ 行第 $j$ 列表示第 $i$ 个样本属于第 $j$ 个簇）对 $\tilde{\boldsymbol{Y}}$ 进行初始化。然后计算 $S = \text{Diag}(\tilde{\boldsymbol{Y}}\boldsymbol{1})$，并根据命题 3.1计算 $\tilde{\boldsymbol{\beta}}$。最后 $\hat{\boldsymbol{H}}$ 由 $\hat{\boldsymbol{H}} \leftarrow \boldsymbol{S}^{-\frac{1}{2}}\tilde{\boldsymbol{Y}}$ 初始化。

### 3.2.2 MuENL

针对动态测试集标记增广学习，提出了 MuENL (Multi-label learning with Emerging New Labels) 方法，设计了基于特征与已知标记上的预测值的新标记检测器以及鲁棒的更新模型以用于整合检测到的新标记，且能够容忍新标记检测的误差。

**形式化**

首先形式化多标记设置下的动态测试集标记增广学习。该设置下，初始有一个标记完整的训练集，然后未标记样本以数据流的形式不断出现。令 $\mathcal{X}$ 表示特征空间并定义 $\boldsymbol{X}_0 = [\boldsymbol{x}_{-n+1}, \cdots, \boldsymbol{x}_{-1}, \boldsymbol{x}_0]^\top \subseteq \mathcal{X}$ 为初始观察到的 $n$ 个

完整标记的样本。令 $\boldsymbol{x}_t$ 表示在第 $t$ 时刻观察到的数据流上的未标记样本。令 $\boldsymbol{X}_t, t \in \{1,2,\cdots,T\}$ 表示在第 $t$ 时刻能够访问的数据。令 $\boldsymbol{v}_0 = \{1,2,\cdots,\ell\}$ 表示 $t=0$ 时刻的已知标记集合[①]，即一开始的训练集上观察到的所有不同标记的并集。令 $\boldsymbol{v}_t$ 表示 $t$ 时刻的标记集合，且它的最大标记序号为 $\ell = |\boldsymbol{v}_t|$。在 $t$ 时刻，当一个新标记转化为已知标记时，标记集合将加入该新标记 $\ell' = \ell + 1$: $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} \cup \{\ell'\}$ [②]；否则标记集合将不会发生变化：$\boldsymbol{v}_t = \boldsymbol{v}_{t-1}$ 。令 $\boldsymbol{Y}_0 = [\boldsymbol{y}_{-n+1};\cdots;\boldsymbol{y}_{-1};\boldsymbol{y}_0] \in \{-1,1\}^{\ell \times n}$ 表示初始训练集上 $\boldsymbol{X}_0$ 对应观察到的标记矩阵；$\boldsymbol{y}_t = [y_{t,1},\cdots,y_{t,\ell}] \in \{-1,1\}^{\ell}$ 表示新出现的样本的标记向量，其中，$y_{t,j}=1$ 表示 $\boldsymbol{x}_t$ 上有标记 $j$，否则没有该标记。注意到 $\boldsymbol{y}_t$ 中的所有元素都是未被观察到的，因为当 $t>0$ 时，$\boldsymbol{x}_t$ 是无标记样本。给定 $\boldsymbol{X}_t$ 和 $\boldsymbol{Y}_0$，动态测试集标记增广学习的目标是学习一个函数集合 $\mathcal{H}_t = [h_{t,1};h_{t,2};\cdots;h_{t,\ell}]$，其中，$h_{t,j}: \mathcal{X} \to \{-1,1\}^{\ell}$ 表示在 $t(t \in \{1,2,\cdots,T\})$ 时刻对标记 $j$ $(j \in \{1,2,\cdots,\ell\})$ 的分类器。对每个 $\boldsymbol{x}_t$，$\hat{\boldsymbol{y}}_t = \mathcal{H}_t(\boldsymbol{x}_t)$ 是它的预测标记向量，其中 $\boldsymbol{x}_t$ 同时与已知标记和新标记关联。

直接估计 $\mathcal{H}_t$ 将会十分困难，因为 $\boldsymbol{v}_t$ 是一个未知的变量，即不知道一个新来的样本 $\boldsymbol{x}_t$ 是否有一个新标记。此外，假设在整个数据流上真实标记都是无法获得的。因此，准确预测出新样本中是否有新标记对新标记建模，在数据集上维持较高的性能都是非常重要的。

首先为未知标记建立一个检测函数 $\mathcal{D}_t(\boldsymbol{x}_t)$，当 $\boldsymbol{x}_t$ 与一新标记关联则输出1，否则输出-1。对数据流上的每一个样本 $\boldsymbol{x}_t$，当前的分类器用来做预测，可得 $\hat{\boldsymbol{y}}_t = \mathcal{H}_t(\boldsymbol{x}_t)$，其中 $\mathcal{H}_t = [h_{t,1};h_{t,2};\cdots;h_{t,\ell};\mathcal{D}_t]$。

算法3.2 总结了 MuENL 方法。它包括 3 个主要组件：(1) 为已知标记建立多标记分类器 $(\mathcal{H}_t = [h_{t,1};\cdots;h_{t,\ell}])$；(2) 为检测新标记建立检测器 $(\mathcal{D}_t)$；(3) 更新 $\mathcal{H}_t$ 和 $\mathcal{D}_t$ 为 $\mathcal{H}_{t+1}$ 和 $\mathcal{D}_{t+1}$ 。

对 $\boldsymbol{x}_t$ 做出预测 $\hat{\boldsymbol{y}}_t$ 后，当下列条件满足时，对分类器和检测器进行更新：

(1) 当 $\mathcal{D}_t(\boldsymbol{x}_t)=1$，新标记样本被加入缓冲区 $B$；

(2) $|B|$ 缓冲达到设定的最大容量。

用 $\mathcal{T}_t = (\boldsymbol{X}_t, \mathcal{H}_t(\boldsymbol{X}_t))$ 更新分类器和检测器，如算法3.2所示。此时，已知的标记集合将被扩展，以包括新标记 $\ell' = \ell + 1$: $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} \cup \{\ell'\}$。由于当 $0 < |B| <$ MAX_BUFFER_SIZE 时，尽管检测到新标记，但是由于数据并不足以训练/更新一个好的分类器。在这种情况下，用检测器的输出作为新标记的预测。

[①]其中 $\boldsymbol{v}$ 中的数字代表标记的序号。

[②]此后，$\ell$ 会自动更新为 $\ell'$。

**算法 3.2** MuENL

**输入**: $\boldsymbol{X}_0, \boldsymbol{Y}_0, \{\boldsymbol{x}_t, t \in \{1, 2, \cdots, T\}\}$
**输出**: 每个 $\boldsymbol{x}_t$ 的预测 $\hat{\boldsymbol{y}}_t$

1: 基于 $\boldsymbol{X}_0$ 和 $\boldsymbol{Y}_0$ 训练 $\mathcal{H}_0$；在 $\boldsymbol{X}_0$ 上建立新标记检测器 $\mathcal{D}_0$；令 $t = 1$；
2: 初始化采样权重向量 $\boldsymbol{s}_0 = \mathbf{1}_{|\boldsymbol{X}_0|}$；
3: $\mathcal{H}_1 = [\mathcal{H}_0; \mathcal{D}_\prime]$; $\mathcal{D}_1 = \mathcal{D}_0$；
4: **repeat**
5: 接受一个新样本 $\boldsymbol{x}_t$, $\boldsymbol{X}_t = [\boldsymbol{X}_{t-1}, \boldsymbol{x}_t^\top]$；
6: 更新采样权重向量 $\boldsymbol{s}_t = [\boldsymbol{s}_{t-1}; 1]$；
7: $\hat{\boldsymbol{y}}_t = \mathcal{H}_t(\boldsymbol{x}_t)$，其中 $\mathcal{H}_t = [h_{t,1}; h_{t,2}; \cdots; h_{t,\ell}; \mathcal{D}_t]$；
8: **if** $\mathcal{D}_t(\boldsymbol{x}_t) = 1$
9: 将 $\boldsymbol{x}_t$ 加入 $B$；
10: **if** $|B|$ = MAX_BUFFER_SIZE
11: 在 $\mathcal{T}_t = (\boldsymbol{X}_t, \mathcal{H}_t(X_t))$ 建立 $\mathcal{D}_{t+1}$ 和 $\mathcal{H}_{t+1}$；
12: 清空 $B$；
13: 当前新标记转为已知标记:
$\ell \leftarrow \ell + 1$; $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} \cup \{\ell\}$；
14: 更新采样权重 $\boldsymbol{s}_t \leftarrow 0.8\boldsymbol{s}_t$；
15: $X_t \leftarrow$ 根据 $\boldsymbol{s}_t$ 采样 $X_t$；
16: 基于 $X_t$ 更新 $\boldsymbol{s}_t$；
17: **end if**
18: **end if**
19: $t \leftarrow t + 1$；
20: $t = t - 1$； $\mathcal{D}_t = \mathcal{D}_{t-1}$； $\mathcal{H}_t = \mathcal{H}_{t-1}$；
21: **until** $t = T$。

为了减小存储和计算开销，并非所有历史数据均保存下来，即当分类器和检测器更新后，只有 $\boldsymbol{X}_t$ 的一个子集被保留用于以后的处理。模型更偏好更近出现的样本，因此，为每一个样本维护了一个采样权重 $\boldsymbol{s}_t$。当 $\boldsymbol{x}_t$ 第一次出现的时候 (算法 3.2第 6 行)，相应权重被初始化为 1.0。然后在每一次分类器和检测器更新后，它将乘以一个衰退系数 0.8(算法 3.2第 14 行) [①]。最后，为 $\boldsymbol{X}_t$ 中的每个样本产生一个 $[0, 1]$ 上的随机数，当产生的随机数小于样本对应的采样权重时，则被选中保留 (算法 3.2第 15 行)。

## 多标记分类器 (PLR)

为每个标记 $i$ 建立一个线性分类器 $(\boldsymbol{w}_i, b_i)$，即 $h_i(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)$。除了最小化错分损失之外，为了利用标记之间的二阶关系[42]，同时最小化正负标记对排序损失。因此这个分类器被命名为 PLR(Pairwise Label Ranking)。整个

[①]这个系数时根据经验设定，与其它设定相比，在不同的数据集上更为鲁棒。

优化过程是一个迭代优化的过程，对于每个标记 $i$，固定其它 $\boldsymbol{w}_j$ $(j \neq i)$ 优化 $\boldsymbol{w}_i$，直到收敛。对每个 $\boldsymbol{w}_i$ 的优化问题形式化如式 (3.12) 所示。

$$\begin{aligned} \min_{\boldsymbol{w}_i, b_i, \xi, \zeta} \quad & \frac{1}{2}\|\boldsymbol{w}_i\|^2 + C_1 \sum_{k=1}^{n} \xi_k + C_2 \sum_{j=1}^{\ell} \sum_{k=1}^{n} \zeta_{j,k}, \\ \text{s.t.} \quad & y_{i,k} f_{i,k} \geq 1 - \xi_k, \\ & \Delta_{j,k}(f_{i,k} - f_{j,k}) \geq 1 - \zeta_{j,k}, \\ & \xi_k \geq 0,\ \zeta_{j,k} \geq 0, \\ & j \in \{1, 2, \cdots, \ell\},\ k \in \{1, 2, \cdots, n\}, \end{aligned} \tag{3.12}$$

其中 $\Delta_{j,k} = y_{i,k} - y_{j,k}$，$f_{i,k} = \boldsymbol{w}_i^{\top} \boldsymbol{x}_k + b_i$，$C_1, C_2$ 是权衡参数。

为了简化优化过程，将 $(\boldsymbol{w}_i, b_i)$ 转化为 $\boldsymbol{w}_i$：在 $x_k$ 向量的后面增加一个元素 1，即 $f_{i,k} = \boldsymbol{w}_i^{\top}[\boldsymbol{x}_k; 1]$。那么，式 (3.12) 可以被重写为式 (3.13)

$$\min_{\boldsymbol{w}_i} \sum_{j=1}^{\ell} \sum_{k=1}^{n} [1 - (y_{i,k} - y_{j,k})(f_{i,k} - f_{j,k})]_+ + \lambda_1 \sum_{k=1}^{n} [1 - y_{i,k} f_{i,k}]_+ + \frac{\lambda_2}{2}\|\boldsymbol{w}_i\|^2, \tag{3.13}$$

其中 $\lambda_1$ 和 $\lambda_2$ 是两个权衡参数。解式 (3.13) 时，首先计算目标函数的次梯度，然后利用梯度下降进行求解。

## 新标记检测

新标记的出现可能是由于出现了之前从未见过的特征属性值，或是新的标记关系，或者两者都有。因此同时考虑特征以及标记空间：如果一个新样本，它在特征空间与以前观测得到的样本特性不同，那么它可能会与一个新标记相关联；如果在标记空间中，它在已知标记上得到的预测与众不同，那么它也可能会与一个新标记相关联。基于这个想法，提出了 `MuENLForest` 作为新标记检测器 $\mathcal{D}$。类似于以前的工作[45, 49]，用标记预测信息编码扩展特征空间。当数据被重新编码后，新标记检测器的构建根据一种有效的异常检测技术 `iForest`[22] 改进而成。`MuENLForest` 与 `iForest` 的不同如下所示：

(1) `iForest` 仅仅考虑特征空间，而 `MuENLForest` 则将标记预测信息编码到特征中去，同时考虑特征与标记信息。由于除初始训练集以外的样本都是无标记的，因此，在训练 `MuENLForest` 时，使用预测值代替真实标记。

(2) 在 `iForest` 中，每棵树的节点会随机选择一个维度随机分裂。而在

MuENLForest 中，每棵树的每个节点会随机选择固定数量的维度，然后根据这些选中的维度上的聚类结果进行分裂。

(3) 当对一个测试样本进行预测时，iForest 把该样本在各个树上的平均路径长度作为异常得分。当这个分数比较小的时候表示该样本落在一个样本稀疏的区域，更有可能是一个异常点。但是这样的策略在多标记设置下的新标记检测中不会起作用，因为那些有新标记的样本往往同时也与一些已知标记相关联，它们可能同时也落到那些有着相同已知标记样本的稠密区域。而 MuENLForest 同时刻画了特征和标记模式的特性。且在构造每一棵树的过程中，会根据落在叶子节点的样本构造一个能够覆盖所有该节点上样本的最小球。当一个测试样本落在这个球外的话，则被预测为与新标记关联；否则认为该样本在特征或标记组合上与训练样本无异。

$t>0$ 时刻的训练集为 $\mathcal{T}_t=(\boldsymbol{X}_t,\mathcal{H}_t(\boldsymbol{X}_t))$，且 $\mathcal{T}_0=(\boldsymbol{X}_0,\boldsymbol{Y}_0)$. MuENLForest 由 $g$ 棵 MuENLTree 组成，每棵 MuENLTree 建在 $\mathcal{T}_t$ 的一个大小为 $\psi$ 的子集上。其中，每个子集根据采样权重 $\boldsymbol{s}_t$ 随机采样得到。MuENLTree 的定义如下：

**定义 3.1** MuENLTree 是一棵二叉树，由中间节点和叶子节点组成。令 $\boldsymbol{a}=[\boldsymbol{x};\mathcal{H}_t(\boldsymbol{x})]$ 表示训练集样本以及对它的预测值。每个中间节点都由 $\|\boldsymbol{a}^{\boldsymbol{q}}-\boldsymbol{p}_1\|\leq\|\boldsymbol{a}^{\boldsymbol{q}}-\boldsymbol{p}_2\|$ 分裂成两个子节点，其中 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是选中维度 $\boldsymbol{q}$ 上的两个聚类中心，$\boldsymbol{a}^{\boldsymbol{q}}$ 是 $\boldsymbol{a}$ 在选中维度 $\boldsymbol{q}$ 上的投影。在每个叶子节点上定义最小的球覆盖 $\boldsymbol{S}$($\boldsymbol{S}$ 为落在该叶子节点的所有训练样本)，它的半径为 $r=\max_{\boldsymbol{x}\in\boldsymbol{S}}\|\boldsymbol{a}-\boldsymbol{m}\|$，其中 $\boldsymbol{m}=\text{mean}(\boldsymbol{S})$。

训练过程中通过递归生成一棵 MuENLTree，样本被逐渐划分，直到以下任一条件 (条件集合用 $\boldsymbol{C}$ 表示) 满足：(a) 树达到了最大的高度限制 $e_m$；(b)$|\boldsymbol{S}|=1$；(c)$\boldsymbol{S}$ 中所有的样本在 $\boldsymbol{q}$ 上的投影都相同，即都为 $\boldsymbol{x}^{\boldsymbol{q}}$。过程 3.1 总结了 MuENLTree 建树的过程。 在生成 MuENLForest 时，用每个样本的预测值增广它们的特征表示，这样一来，特征空间和标记信息都能被考虑到。每个中间节点的分裂都是基于当前属性的一个随机选中的子集上的聚类结果，更相近的样本将被划分到同一个子节点。因此，相同叶子节点上的样本，必然会存在某些属性 (或是特征，或是某些标记上的预测值，或是两者) 相似。

MuENLForest，即 $\mathcal{D}_t(\cdot)$ 建好后，将用它进行新标记检测。在用每棵 MuENLTree 评估测试样本 $\boldsymbol{x}_t$ 时，如果 $\boldsymbol{x}_t$ 落在相应叶子节点的球外，则

[①] 当 $|\boldsymbol{S}|=1$ 的情况下，聚类中心和球将计算它的父节点的聚类中心和球代替。

**过程 3.1** MuENLTree

**输入**: 采样的训练集 $\boldsymbol{S}$，当前树的高度 $e$，最大树高 $e_m$，维度数 $k$
**输出**: MuENLTree

**if** 满足 $\boldsymbol{C}$ 中任一条件：
    建立一个半径为 $r = \max_{\boldsymbol{a}\in\boldsymbol{S}}(\|\boldsymbol{a}-\boldsymbol{m}\|)$ 的球[①]，其中 $\boldsymbol{m} = \text{mean}(\boldsymbol{S})$；
    **return** $\boldsymbol{N}\{\boldsymbol{N}.\boldsymbol{S} \leftarrow \boldsymbol{S}, \boldsymbol{N}. \leftarrow , \boldsymbol{N}.r \leftarrow r\}$;
**else**
    令 $\boldsymbol{Q}_1$ 表示 $\boldsymbol{S}$ 的特征矩阵；
    令 $\boldsymbol{Q}_2$ 表示 $\boldsymbol{Q}_1$ 相应的预测值矩阵；
    从 $\boldsymbol{Q}_1$ 随机挑选 $k$ 列：$\boldsymbol{q}_1 \subset \boldsymbol{Q}_1$；
    从 $\boldsymbol{Q}_2$ 随机挑选 $k$ 列：$\boldsymbol{q}_2 \subset \boldsymbol{Q}_2$；
    $\boldsymbol{q} = [\boldsymbol{q}_1, \boldsymbol{q}_2]$；
    聚类中心 $\{\boldsymbol{p}_1, \boldsymbol{p}_2\} \leftarrow \text{Clustering}(\boldsymbol{q}, S)$；
    $\boldsymbol{S}_l = \{\boldsymbol{a} \in \boldsymbol{S} \mid \|\boldsymbol{a}^{\boldsymbol{q}} - \boldsymbol{p}_1\| \le \|\boldsymbol{a}^{\boldsymbol{q}} - \boldsymbol{p}_2\|\)$;
    $\boldsymbol{S}_r = \{\boldsymbol{a} \in \boldsymbol{S} \mid \|\boldsymbol{a}^{\boldsymbol{q}} - \boldsymbol{p}_1\| > \|\boldsymbol{a}^{\boldsymbol{q}} - \boldsymbol{p}_2\|\)$;
    **return** $\boldsymbol{N}\{\boldsymbol{N}.\boldsymbol{S} \leftarrow \boldsymbol{S}, \boldsymbol{N}. \leftarrow , \boldsymbol{N}.r \leftarrow r,$
        $\boldsymbol{N}.\boldsymbol{q} \leftarrow \boldsymbol{q}, \boldsymbol{N}.\{\boldsymbol{p}_1, \boldsymbol{p}_2\} \leftarrow \{\boldsymbol{p}_1, \boldsymbol{p}_2\},$
        $\boldsymbol{N}.\boldsymbol{N}_{left} \leftarrow \text{MuENLTree}(\boldsymbol{S}_l, e+1, e_m, k)$
        $\boldsymbol{N}.\boldsymbol{N}_{right} \leftarrow \text{MuENLTree}(\boldsymbol{S}_r, e+1, e_m, k)\}$;
**end if**

$\mathcal{D}_t(\boldsymbol{x}_t) = 1$，即有新标记，否则 $\mathcal{D}_t(\boldsymbol{x}_t) = -1$，即没有新标记。最终检测输出是根据 $g$ 棵 `MuENLTree` 的结果用多数投票得到。

它的基本思想如下：相同叶子节点上的样本，必然会存在某些属性 (或是特征，或是某些标记上的预测值，或是两者) 相似。所以当一个测试样本落在某个叶子节点，但却在该节点上定义的球外，说明它必然与其所经路径对应的特征和标记预测值以外的其它特征和标记预测值非常不同，则更有可能与一个新标记相关联。

## 模型更新

当缓冲区 $B$ 满了，需要对 $\mathcal{H}_t$ 进行更新。这个更新主要包括为新的标记创建新的分类器。因为新标记检测的结果并非完美，可能会错过一些有新标记的样本或者把没有任何新标记的样本当成了新标记样本。所以必须设计一种鲁棒的分类器模型，使其能够容忍检测带来的错误。基本思想是并非直接使用新标记检测器的输出作为监督信息，而是额外引入了一个隐变量，用于估计 $\boldsymbol{X}_t$ 中样本的真实标记分配（即是否有新标记）。然后通过优化同时学习这个隐变量和新标记上的分类模型，使它们能够最好地匹配训练数据。这样一来，学到的

模型将对新标记检测器检测误差更加鲁棒。

在 $\boldsymbol{X}_t$ 中，令 $\boldsymbol{X}_B$ 表示缓冲区中收集的样本 (即检测器预测有新标记的样本)，$\boldsymbol{X}_U$ 表示检测器预测没有新标记的样本，其中 $X_U = X_t \setminus X_B$。令 $\boldsymbol{d} = [d_1, d_2, \cdots, d_m]^\top$ 表示 $X_t = [X_B, X_U]$ 对应的潜在新标记的指示向量，其中 $m$ 是 $[X_B, X_U]$ 中的样本数，如果 $\boldsymbol{x}_k \in [X_B, X_U]$ 真正与新标记关联，则 $d_k = 1$ 否则为 0。$\boldsymbol{d}$ 是我们要估计的变量，并由检测器的输出值进行初始化。

为了得到鲁棒的分类器，提出了多标记新标记学习更新方法 (MNL)，同时优化 $\boldsymbol{d}$ 并为新标记 $\ell \leftarrow \ell + 1$ 建立模型 $\boldsymbol{w}_a$。该方法基于式 (3.13) 实现：将 $y_{i,k}$ 替换为 $2d_k - 1$，同时优化了标记对排序损失 (第一项，鼓励正标记排在负标记之前)、错分损失 (第二项，鼓励每个标记都被正确预测出) 以及正则化项 (最后两项)。优化问题如式 (3.14) 所示：

$$\begin{aligned} \min_{\boldsymbol{w}_a, \boldsymbol{d}} \quad & \sum_{j=1}^{\ell} \sum_{k=1}^{m} [1 - (2d_k - 1 - y_{j,k})(f_{\ell,k} - f_{j,k})]_+ \\ & + \lambda_1 \sum_{k=1}^{m} [1 - (2d_k - 1) f_{\ell,k}]_+ + \frac{\lambda_2}{2} \|\boldsymbol{w}_a\|^2 + \frac{\lambda_3}{2} \|\boldsymbol{d}\|^2, \\ \text{s.t.} \quad & d_k \in \{0, 1\}, k \in \{1, 2, \cdots, m\}, \end{aligned} \tag{3.14}$$

其中，$y_{j,k} = h_{t,j}(\boldsymbol{x}_k)$，$\lambda_1, \lambda_2, \lambda_3$ 是权衡参数。

上面的优化问题是一个 NP 难问题。为了简化问题，将约束从 $d_k \in \{0, 1\}$ 放松到 $d_k \in [0, 1]$，并交替优化 $\boldsymbol{d}$ 和 $\boldsymbol{w}_a$。当给定 $\boldsymbol{w}_a$ 更新 $\boldsymbol{d}$ 时，求解子问题如式 (3.15) 所示：

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & \sum_{j=1}^{\ell} \sum_{k=1}^{m} [1 - (2d_k - 1 - y_{j,k})(f_{\ell,k} - f_{j,k})]_+ + \lambda_1 \sum_{k=1}^{m} [1 - (2d_k - 1) f_{\ell,k}]_+ + \frac{\lambda_3}{2} \|\boldsymbol{d}\|^2, \\ \text{s.t.} \quad & d_k \in [0, 1], k \in \{1, 2, \cdots, m\}. \end{aligned} \tag{3.15}$$

当给定 $\boldsymbol{d}$，更新 $\boldsymbol{w}_a$ 时，求解子问题如式 (3.16) 所示：

$$\min_{\boldsymbol{w}_a} \sum_{j=1}^{\ell} \sum_{k=1}^{m} [1 - (2d_k - 1 - y_{j,k})(f_{\ell,k} - f_{j,k})]_+ + \lambda_1 \sum_{k=1}^{m} [1 - (2d_k - 1) f_{\ell,k}]_+ + \frac{\lambda_2}{2} \|\boldsymbol{w}_a\|^2, \tag{3.16}$$

为了求解式 (3.15) 和 (3.16)，分别计算目标函数的次梯度，然后利用梯度下降算法进行更新。在每次变量更新后，将 $\boldsymbol{d}$ 投影到 $[0, 1]^m$：$\boldsymbol{d} \leftarrow \min(\boldsymbol{1}, [\boldsymbol{d}]_+)$，以满足 $[0, 1]$ 盒约束。

过程 3.2 MNL

输入: $\boldsymbol{X}_B, \boldsymbol{X}_U, \boldsymbol{w}_i, i \in \{1, 2, \cdots, \ell\}, \boldsymbol{d}_{init}, \boldsymbol{w}_{a,init}$
输出: $\boldsymbol{w}_a$

初始化 $0 \leftarrow init$； $\boldsymbol{w}_{a,0} \leftarrow \boldsymbol{w}_{a,init}$；
$t = 1$；
**repeat**:
    $\boldsymbol{d}_t \leftarrow$ 解式 (3.15)，使用 $\boldsymbol{d}_{t-1}$ 初始化 $\boldsymbol{d}$；
    $\boldsymbol{w}_{a,t} \leftarrow$ 解式 (3.16) ，使用 $\boldsymbol{w}_{a,t-1}$ 初始化 $\boldsymbol{w}_a$；
    $t \leftarrow t + 1$；
**until** 算法收敛或者达到最大迭代轮数；
$\boldsymbol{w}_a = \boldsymbol{w}_{a,t}$。

为了能够更快收敛，且最终得到一个好结果，采用了热启动策略。在第一轮，将 $\boldsymbol{X}_B$ 当作正样本集合，把 $[\boldsymbol{X}_0, \boldsymbol{X}_U]$ 当作负样本集合，训练一个线性分类器，作为 $\boldsymbol{w}_a$ 的初始值。在后续的每一轮迭代优化中，用上一轮的输出作为下一轮优化变量的初始值。过程3.2总结了 MNL 学习分类器的过程。

为了更新检测器，在 $\boldsymbol{X}_t$ 上重新构建MuENLForest，其中，每棵 MuENLTree 都是根据 $\boldsymbol{X}_t$ 上的采样权重随机采样 $\psi$ 个样本，以产生 $\mathcal{D}_{t+1}$ 替代 $\mathcal{D}_t$。

## 3.3 实验测试

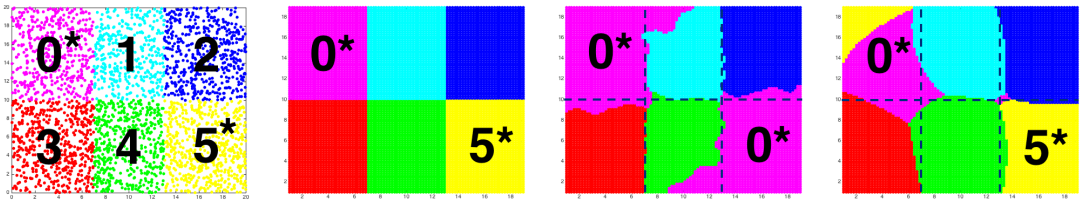
### 3.3.1 静态测试集标记增广实验

#### 人工数据实验设置

随机产生 300 个训练包作为训练样本，包括 6 个不同的标记（0-5）分别对应图3.1(a) 中 6 个不同颜色的矩形区域。其中平均每个包包括 10 个示例以及 2.47 个标记。在训练阶段，只观察到包级标记 1-4 作为已知标记，而标记 0 和 5 为未被标注的新标记（图 3.1中用星号标记），在训练集中是未知的。另外从各个标记的区域一共采样 10000 个示例作为测试集如图3.1(b) 所示，用以评价包括新标记的示例标注任务上的性能。最后与 MIML-NC 算法 [73] 对比，该算法将所有新标记示例当做同一个新标记。

## 人工数据实验结果

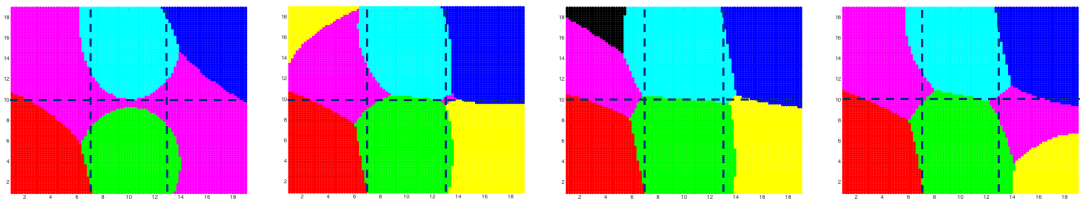
图 3.1(c) 和 (d) 分别展示了 MIML-NC 和 DMNL 在测试集上的预测结果，其中虚线表示不同真实标记的分界线。如图所示，DMNL 和 MIML-NC 和在已知标记上的性能是可比的。但是 MIML-NC 把所有的新标记示例都预测成同样的标记 0，而 DMNL 能够预测多个不同的新标记 0 和 5。




(a) Training data (b) Test data (c) MIML-NC (d) DMNL

**图 3.1** 人工数据实验

在实际应用时，新标记的个数是未知的，所以它的数目 $k$ 需要作为参数由用户手动设置。图 3.2 展示了 DMNL 的性能随不同 $k$ 值变化的结果。



(a) $k$=1 (b) $k$=2 (c) $k$=4 (d) $k$=8

**图 3.2** 不同 $k$ 取值的影响

当 $k$ 与真实新标记个数相同（即 $k = 2$）时，在已知标记和新标记预测上 DMNL 都能取得最好的性能。由此可以通过交叉验证 (根据在验证集已知标记上的性能) 来挑选合适的 $k$ 值。

当 $k$ 比真实新标记数大时，一些属于相同新标记的示例可能会被细分到不同的子类：见图 3.2(c)。但是算法不一定总是产生用户指定的 $k$ 那么多个新标记。因为优化过程中同时考虑了包上的错分损失和所有示例的聚类结构。那些检测到的只有非常少量正示例的新标记 (由正交约束产生) 会被认为是噪声，而不是新标记。图 3.2中，当 $k = 4$ 时只产生了 3 个新标记，而当 $k = 8$ 时只产生了 2 个新标记。

当 $k = 1$ 时，多个标记检测的性能以及在已知标记上预测的结果都不甚理

想，因为分属两个不同标记的示例被强制当做属于同一个标记，而这与本工作的基本假设不符合：相同标记的示例应当属于同一个簇。

## 真实数据集实验设置

使用常用的 $\text{MSRCv}_2$ 图像数据集 [82]，两个字母数据集 [70]（Letter Carroll 和 Letter Frost），以及 MNIST 手写数字数据集 [83] ①。由于 MNIST 是一个单示例单标记数据集，通过随机从 10 个数字中采样 200 个包得到多示例多标记数据集。其中每个包平均包括 27.6 个示例，3.09 个标记。

沿用了 Pham 等人[73] 工作中设定：把每个数据集标记集合分成已知标记和未知标记（新标记）两个部分。对每个数据集考虑 3 种不同数量的新标记：对前三个数据集，第 1-4，第 1-8，第 1-16 标记作为新标记，剩余标记作为已知标记；对 MNIST，由于一共只有 10 个标记，把第 1-2，第 1-4，第 1-6 标记作为新标记，剩余标记作为已知标记。作为训练集时，这些新标记被移除。

性能评价考虑在以下几个方面的预测表现：(A) 示例级的标注包括 (A1) 不同新标记检测，(A2) 新标记示例检测，(A3) 示例标注在已知标记上的预测；(B) 包级的预测表现包括：(B1) 对每个包上多个新标记的预测，(B2) 对包上已知标记的预测。由于对于多个新标记检测问题，目前尚无评价指标，必须为示例标注和包标记预测中多个新标记预测设计新的指标。

为了评价 A1 任务上的新标记标注表现，定义了一个新指标 $\text{F}_{\text{INL}}$。令 $h_i$ 和 $t_i$ 分别表示样本 $\boldsymbol{x}_i$ 的预测标记和真实标记的下标，且令已知标记排在新标记之前，$\text{F}_{\text{INL}}$ 定义如下：

$$
\begin{aligned}
\text{F}_{\text{INL}} &= 2\text{Prec}_{\text{INL}}\text{Rec}_{\text{INL}}/(\text{Prec}_{\text{INL}} + \text{Rec}_{\text{INL}});\\
\text{Prec}_{\text{INL}} &= \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(h_i = h_j)\mathbb{I}(t_i = t_j)\mathbb{I}(t_i > c)\mathbb{I}(h_i > c)}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(h_i = h_j)\mathbb{I}(h_i > c)};\\
\text{Rec}_{\text{INL}} &= \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(h_i = h_j)\mathbb{I}(t_i = t_j)\mathbb{I}(t_i > c)\mathbb{I}(h_i > c)}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(t_i = t_j)\mathbb{I}(t_i > c)},
\end{aligned}
$$

其中，$c$ 是已知标记数。$\text{Prec}_{\text{INL}}$ 度量了从预测为同一新标记的所有示例对中，真正属于同一新标记的示例对的比例；$\text{Rec}_{\text{INL}}$ 度量了所有属于同一新标记的示例对中被预测为相同新标记示例对的比例；$\text{F}_{\text{INL}}$ 是 $\text{Prec}_{\text{INL}}$ 和 $\text{Rec}_{\text{INL}}$ 的调和平均。

① 使用 MS，LC，LF 和 MN 分别作为 $\text{MSRCv}_2$，Letter Carroll，Letter Frost 和 MNIST 的缩写

对 B1 任务无法使用 $F_{INL}$ 指标，因为在一个包上可能有多个新标记，而一个示例只有一个标记（每个示例最多只可能出现一个新标记）。因此，为包级的新标记预测定义了评价指标 $\mathrm{F_{BNL}}$。令 $\boldsymbol{G} \in \{0,1\}^{(c+k)\times n}$ 表示包级的预测标记矩阵，$\boldsymbol{Y} \in \{0,1\}^{(c+k')\times n}$ 为包级的真实标记矩阵，$\boldsymbol{G}_{l,:}$ 表示 $\boldsymbol{G}$ 的第 $l$ 行。定义 $\mathcal{F}(\boldsymbol{y},\boldsymbol{g})$ 为 F1-度量函数，其中 $\boldsymbol{y}$ 为预测标记向量，$\boldsymbol{g}$ 为真实标记向量。同样令已知标记排在未知标记之前，$\mathrm{F_{BNL}}$ 定义如下：

$$\mathrm{F_{BNL}} = \frac{1}{k}\sum_{i=1}^{k} \max(\left\{\mathcal{F}(\boldsymbol{G}_{c+i,:}, \boldsymbol{Y}_{c+j,:}), j \in \{1,\cdots,k\}\right\}).$$

它度量了与真实标记最佳匹配情况下预测标记的平均表现。$\mathrm{F_{INL}}$ 和 $\mathrm{F_{BNL}}$ 都是越高越好。

对其它任务的评价，可以采用现有评价指标：A2 与文献 [73] 的任务相同，因此采用相同的指标 AUC 来评价；A3 用 Accuracy 来评价在已知标记上示例标注的表现；对于 B2，用 Hamming Loss 评价对已知标记的包级预测的好坏。

用 3 个现有的多示例多标记学习方法作为对比方法：`ORLR` [74]、`MIMLfast` [71] 和 `MIML-NC` [73]。前两种方法没法直接处理新标记，因此根据文献 [84]，指定对示例的预测值均小于某个阈值则认为该示例属于一个新标记。由于所有的对比方法都无法检测多个新标记，因此，用 `kmeans` 将检测出来的新标记示例划分成 $k$ 个簇作为后处理。

所有方法的参数都通过 5 折交叉验证选出，除了对比方法的聚类数 $k$。由于对比方法是通过后处理得到多个新标记的预测，因而无法通过训练集选择 $k$。为了公平比较，将对比方法的 $k$ 参数设为与 `DMNL` 选出的 $k$ 一样的值。

## 真实数据集实验结果

对于 A1 多个新标记检测任务，$\mathrm{F_{INL}}$ 结果如图3.3所示。无一例外，`DMNL` 在所有的数据集不同新标记数的情况下都取得了最好的效果。`DMNL` 比其它对比方法好的原因是，其它方法必须采用额外的步骤去处理多个新标记（使用聚类区分不同的新标记，对于 `ORLR` 和 `MIMLfast` 必须通过设置阈值才能检测新标记），这样做忽略了包本身的结构，也会引入其他误差。相对的，`DMNL` 为新标记和已知标记建立了一个整体优化框架，考虑了示例间的聚类结构以及每个包中示例对包标记的贡献。由于 $\mathrm{F_{INL}}$ 是一种比较保守的度量，即只有当一

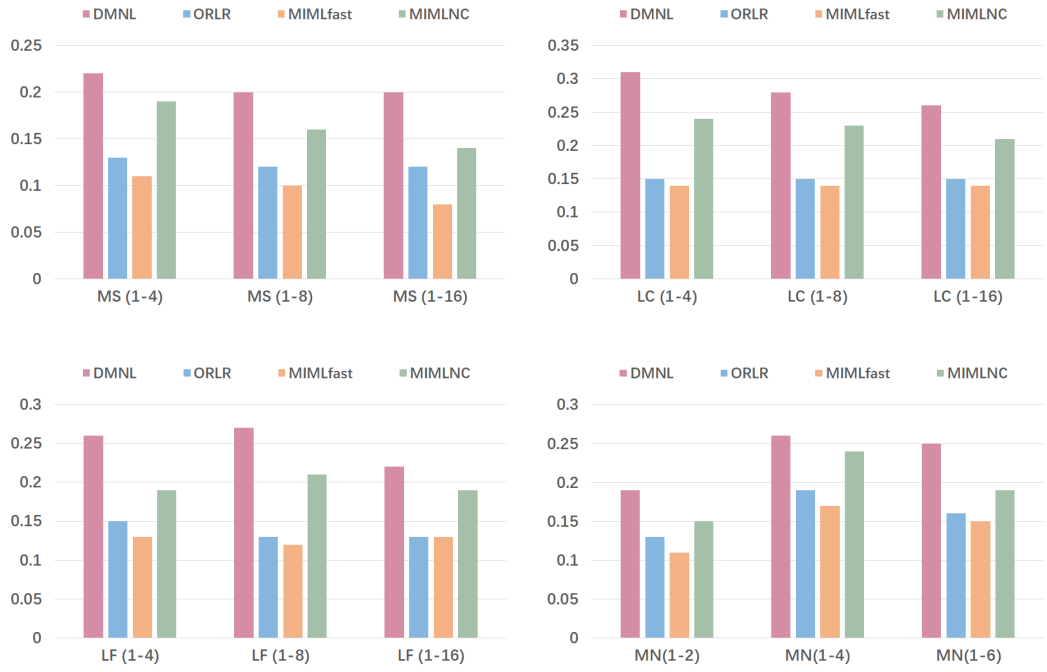



**图 3.3** 示例级多个新标记检测 (A1) $F_{INL}$ 比较

对新标记示例对属于相同的新标记才会对这个值作出贡献，而如果新标记示例对中任何一个示例没有新标记或是不属于同一个新标记都不会作出贡献。这也是 $F_{INL}$ 比较小的原因，但在 $F_{INL}$ 上很小的区别可能都意味着巨大的实际效果上的差异。

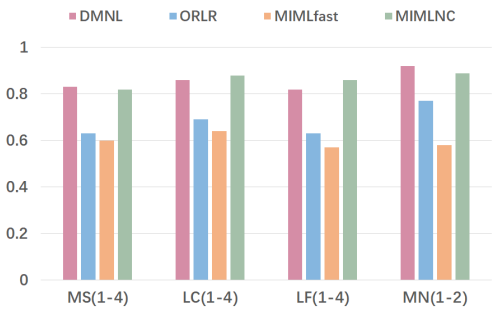



**图 3.4** 新标记示例检测 (A2) AUC 比较

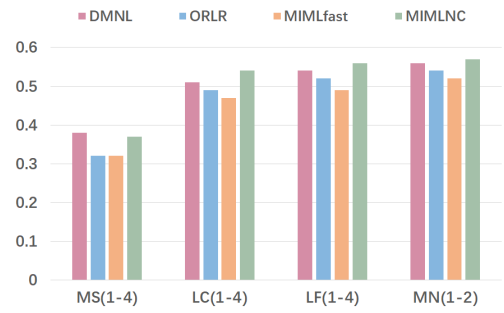


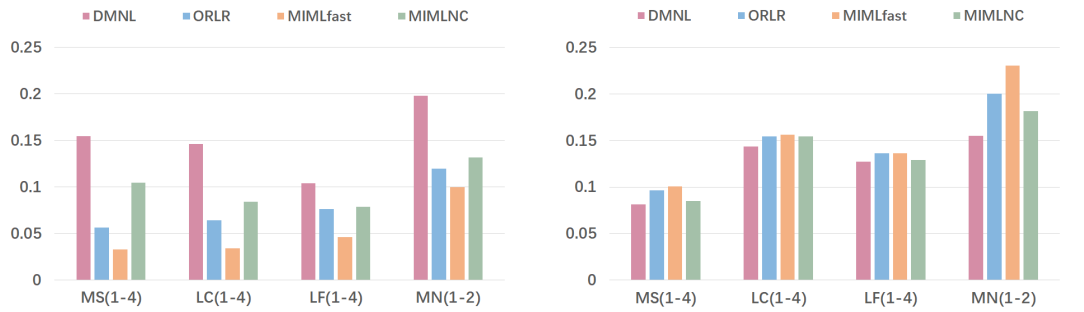

**图 3.5** 已知标记示例标注 (A3) Accuracy 比较

图3.4展示了新标记示例检测的 AUC 结果。如果一个示例包含任一新标记则为新标记示例。DMNL 显著优于 ORLR 和 MIMLfast，而与 MIML-NC（现有的多示例多标记下新标记示例检测的最好方法）相匹。

图3.5展示了示例已知标记的预测准确率。如图所示，DMNL 与 MIML-NC

不分伯仲，比 ORLR 和 MIMLfast 稍好。这表明，即使同时考虑了多个新标记，也不会降低 DMNL 在已知标记标注任务上的性能。

图 3.6总结了包级预测 B1 和 B2 任务的比较结果。正如预期的那样，DMNL 在包级多个新标记检测 (B1) 任务上的性能超过了其它对比方法。在已知标记的包级预测 (B2) 任务上，DMNL 的表现超过了 ORLR 和 MIMLfast，符合 Pham 等人在文献 [73] 中的论述：在多示例多标记设置下同时为新标记建模能够提升对已知标记预测的结果。由于 DMNL 同时考虑了多个新标记并为其建立了统一的模型，因此在 (B2) 任务上 DMNL 也超过了 MIML-NC。




(a) 新标记预测 (B1) $F_{BNL}$ 比较　　(b) 已知标记预测 (B2) Hamming Loss 比较

**图 3.6** 包级预测结果 (B1 和 B2)

## 3.3.2 动态测试集标记增广实验

### 实验设置

用 5 个常用的标准多标记数据集评估 MuENL 方法的性能。5 个数据集为：Birds [85], CAL500 [86], Emotions [87], Enron [63] 和 Yeast [88]，数据集详细信息如表 3.1所示。首先产生初始标记数据集，即 $(X_0, Y_0)$，然后模拟数据流：令未标记样本 $\boldsymbol{x}_t$ 在每个时间点 $t \in \{1, 2, \ldots, T\}$ 陆续出现，根据算法记录结果。

**表 3.1** 数据集

| 数据集 | 样本数 | 样本维度 | 标记数 | 平均标记数 (每样本) |
|---|---|---|---|---|
| Birds | 645 | 260 | 19 | 1.014 |
| CAL500 | 502 | 68 | 174 | 26.044 |
| Emotions | 593 | 72 | 6 | 3.378 |
| Enron | 1702 | 1001 | 53 | 3.378 |
| Yeast | 2417 | 103 | 14 | 4.237 |

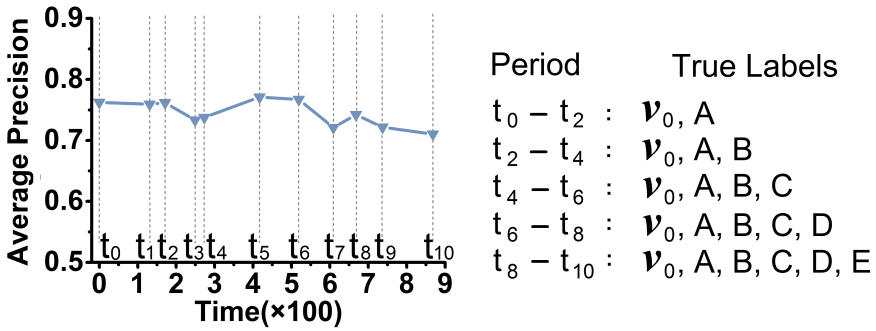



**图 3.7** Yeast 数据集上的结果，包括 5 个新标记

图 3.7 展示了在 Yeast 数据集模拟数据流上的结果，其中 5 个新标记 (A 到 E) 在不同的时间段首次出现。在 $t_0$，初始训练集中只观察到已知标记集合 ($\boldsymbol{v}_0$) 中的标记，然后在此训练集上训练一个多标记学习器。紧接着在 $t_0 - t_1$ 时间段，出现的新样本可能有新标记 A。在 $t_1$ 时刻，缓冲区 $B$(被检测为新标记样本) 满了，MuENL 增广已知标记集合为 $\{\boldsymbol{v}_0, A\}$，对多标记学习器 ($[h_{t,1}; \cdots; h_{t,\ell}]$) 以及新标记检测器 ($\mathcal{D}_t$) 进行更新。更新后的学习器和检测器将用于从 $t_1 + 1$ 时刻起的预测和检测任务。注意，在 $t_1$ 时刻后，标记 A 仍有可能出现在新样本中，只是此时不再作为新标记，而是属于已知标记集合。在 $t_2$ 时刻，新标记 B 开始出现，$t_2 - t_4$ 时的处理与 $t_0 - t_2$ 相同；重复这些过程，直到数据流停止 ($t_{10}$)。在每个时间段所用的学习器和检测器如表 3.2所示。值得注意的是，始终使用 MuENLForest 作检测，而对于预测任务，在新标记模型未建立之前 (缓冲区 $B$ 未满)，使用检测器 MuENLForest 的输出作为新标记的预测结果，而当 MNL 更新后，则改为使用 MNL 得到的模型对新标记预测。 如图 3.7所示，

**表 3.2** 新标记预测模型

| 时间段 | 新标记预测模型 | 真实标记集合 |
|---|---|---|
| $t_0 - t_1$ | MuENLForest | $\boldsymbol{v}_0$, A |
| $t_1 - t_2$ | MNL | $\boldsymbol{v}_0$, A |
| $t_2 - t_3$ | MuENLForest | $\boldsymbol{v}_0$, A, B |
| $t_3 - t_4$ | MNL | $\boldsymbol{v}_0$, A, B |
| $t_4 - t_5$ | MuENLForest | $\boldsymbol{v}_0$, A, B, C |
| $t_5 - t_6$ | MNL | $\boldsymbol{v}_0$, A, B, C |
| ... | ... | ... |

即使经过了 900 个时间点，陆续出现了 5 个新标记，MuENL 总体上的预测性能并没有随着新标记不断出现而下降很多。

对于给定数据集，生成上例数据流的具体过程如下：首先将该数据集上的

标记集合 $\boldsymbol{v}$ 分成两个互补子集，即 $\boldsymbol{v}_N = \{\boldsymbol{v}_{a1}, \ldots, \boldsymbol{v}_{a5}\}$，代表 5 个新标记集合；以及 $\boldsymbol{v}_K$ 表示已知标记集合。令 $\boldsymbol{P}_K$ 表示没有新标记 $\boldsymbol{v}_N$ 的数据集子集。需要注意的是必须对数据集做一些预处理：(1) 如果两个标记总是同时出现，则将它们结合成一个标记处理，即取它们的并集。因为在没有先验知识的情况下，几乎不可能将总是同时出现的标记区分开。(2) 如果一个标记与其它标记均相互独立，即很少与其它标记一同出现。这种情况下将不会选它作为新标记候选，因为这将会使检测任务退化为类增量学习中的新类检测问题。(3) 那些与其它标记相关程度适中（在相关性得分排序在中间位置）的标记，将是潜在的新标记候选。这些标记可能会与不同的已知标记同时出现[①]。

采样 $\boldsymbol{P}_K$ 的 90% 作为初始 $t_0$ 时刻的训练集 $\mathcal{T}_0 = (\boldsymbol{X}_0, \boldsymbol{Y}_0)$。对于 $t > 0$，$\boldsymbol{x}_t$ 从 $\boldsymbol{P}_K \setminus \mathcal{T}_0$ 中无放回地随机抽取，且属于 $\boldsymbol{P}_N$ 中与 $\boldsymbol{v}_{ai}$ 关联的样本集合的一个子集，作为时间段 $t_{2i-2} - t_{2i}$ 的数据流数据，其中 $i = 1, \ldots, 5$。

为了评价算法性能，使用多标记学习中常用的度量，例如 Average Precision，它计算了排在某个正标记之前的正标记的平均个数 (第**??**中有详细介绍)。Average Precision 的值越大越好。

实验基于模拟数据流，测试了在 $t_1, \ldots, t_6$ 点的预测性能，主要考虑 2 方面的评估：(1)$\boldsymbol{v}_t$-评价：评价了算法在 $t$ 时刻整个标记集合上的性能，包括出现的新标记；(2)$\boldsymbol{v}_0$-评价：评价了算法在初始标记集合 $\boldsymbol{v}_0$ 上的表现，即不考虑在新标记上的表现。

对比方法主要考虑与三类工作：现有的多标记学习算法、改造后的类别增量学习算法以适应多标记设置以及 MuENL 的变体方法。对于现有的多标记学习算法，将 MuENL 与 BR [43]、CLR [44]、ECC [45]、PLR、LIMO [89]、以及 GenEML [90] 进行比较试验。这些多标记学习方法均只考虑了初始训练集上的已知标记，而不能处理出现的新标记。BR、CLR 和 ECC 分别考虑了标记的一阶、二阶和高阶关系；PLR 是本工作提出的多标记学习方法，同时最小化标记对排序损失和错分损失；LIMO 和 GenEML 是最近提出的多标记学习方法：LIMO 同时考虑了样本对和标记对间隔，GenEML 是一个生成式模型且可以处理标记缺失问题。

对于增量学习方法，SENCForest [8] 最近被提出用于在数据流上处理多类别类增量学习问题，并取得了成功。它提供了一种统一的基于树的新类检测和预测框架。SENCForest 针对的是多类别学习设置，即每个样本只属于多个类

[①]两个标记之间的相关程度用余弦距离度量。通过对标记两两余弦距离求和得到最终的相关性得分。例如有 A, B, C 三个标记向量，A 的相关性得分为 $\cos(A, B) + \cos(A, C)$。

别中的其中一类，而本工作讨论的是每个样本可能同时与多个标记相关联。为了将 SENCForest 应用于多标记设置，首先采用标记幂集 (LP)[20] 方法将不同的标记组合编码为不同的类，从而将多标记学习设置转化为多类别学习设置。完成转换后，再使用 SENCForest 方法，为了与转换前区分，将其命名为 LP-SENCForest。将 MuENL 与 LP-SENCForest 进行对比。

为了进一步验证 MuENL 各组件的有效性，将 MuENL 与其变体对比：

(1) MuENL-IF：用 iForest [22] (代替 MuENLForest) 作为新标记检测器；
(2) MuENL-OC：用 OC-SVM [21] (代替 MuENLForest) 作为新标记检测器；
(3) MuENL-SVM：用 MuENLForest 作为新标记检测器，但是用线性 SVM 为新标记训练模型 (训练数据与 MNL 所用相同)；
(4) MuENL-OR：用 MuENLForest 作为新标记检测器，但是假设能够获得样本的真实标记用于模型更新。它的性能将是上界，因为实际中真实标记无法获得。

这些变体的各个组件如表 3.3所示。

**表 3.3** MuENL 变体

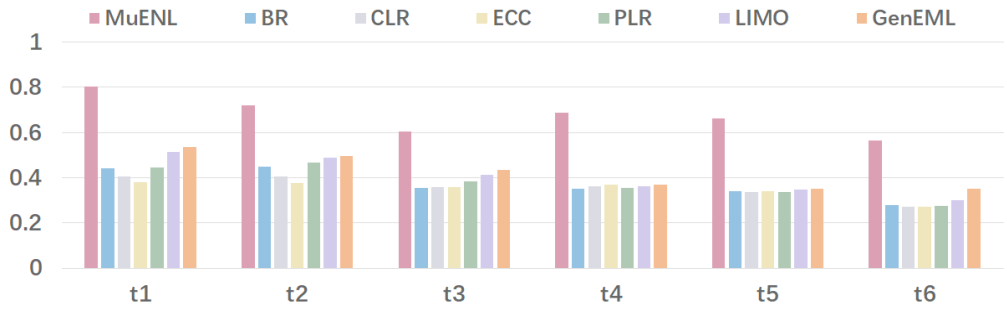
| 算法 | 多标记学习器 | 新标记检测器 | 新标记模型 |
|---|---|---|---|
| MuENL-SVM | PLR | MuENLForest | SVM |
| MuENL-IF | PLR | iForest | MNL |
| MuENL-OC | PLR | OC-SVM | MNL |
| MuENL-OR | PLR | MuENLForest | Oracle+PLR |
| MuENL | PLR | MuENLForest | MNL |

## 实验结果

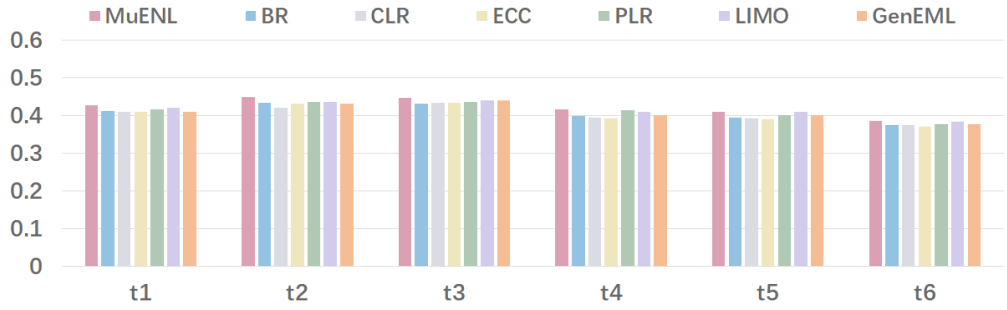
图3.7展示了在 Yeast 数据集上在 $\boldsymbol{v}_t$ 上评价的一个样例[①]。从整体上看，MuENL 从 $t_0$ 到 $t_{10}$ 保持了不错的预测性能。直接原因是 MuENL 有一个好的检测器和鲁棒的学习器。图3.8总结了 MuENL 与 BR、ECC、CLR、PLR、LIMO 和 GenEML 在 $\boldsymbol{v}_t$ 上的对比结果。MuENL 总是比对比方法好，因为这些对比方法均无法检测新标记，更无法为新标记建立模型。

5 个数据集上的 $\boldsymbol{v}_0$-评价结果如图3.9所示。MuENL 在初始已知标记上的性能比其它对比方法更好或是可比。这说明 MuENL 检测并为新标记建立模型并不会影响到在原来多标记任务（初始已知标记）上的性能。

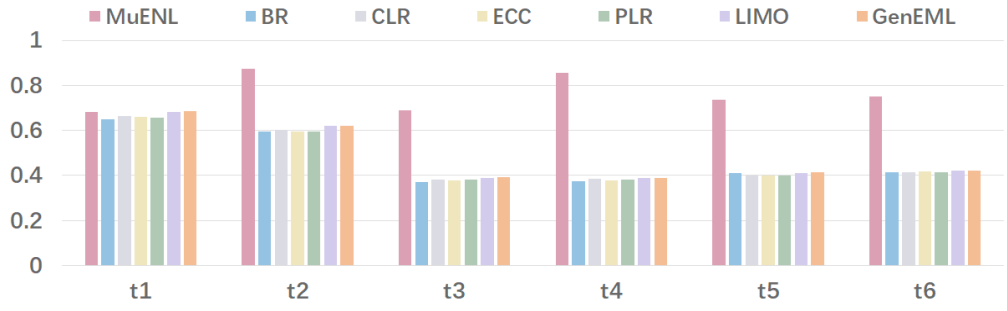
[①]注意，在不同的时间段缓冲区填满的时间点可能不同，即使在同一个时间段，检测器不同缓冲区填满的时间也可能不同。因此，时间段的长短可能因实验而异。



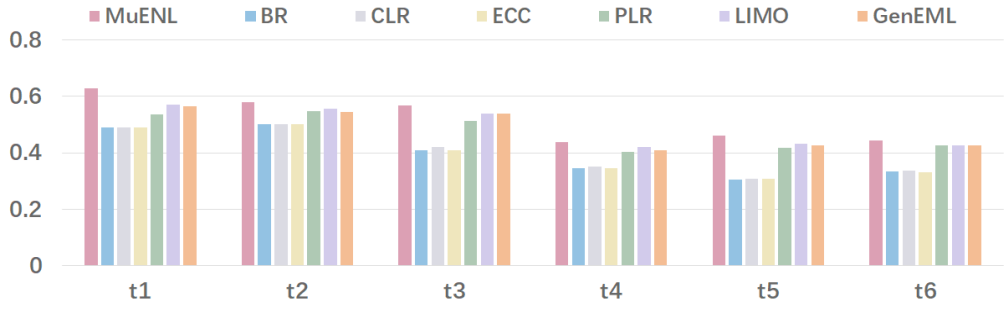

Birds 数据集



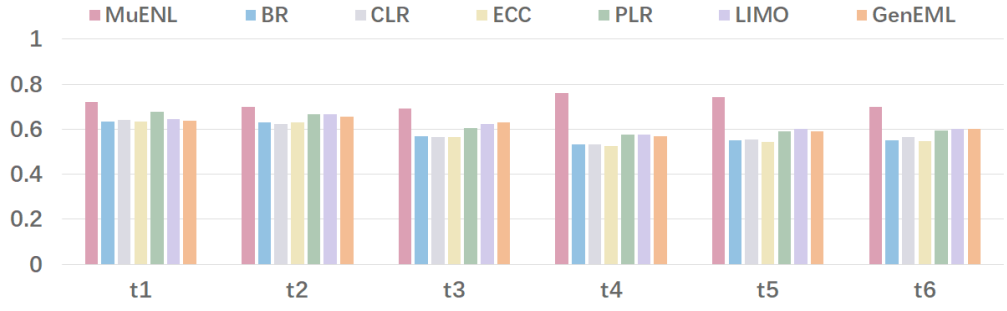

CAL500 数据集




Emotions 数据集




Enron 数据集




Yeast 数据集

**图 3.8** $v_t$-评价结果 (Average Precision)。横坐标 $t_1 - t_6$ 为 6 个时间点。

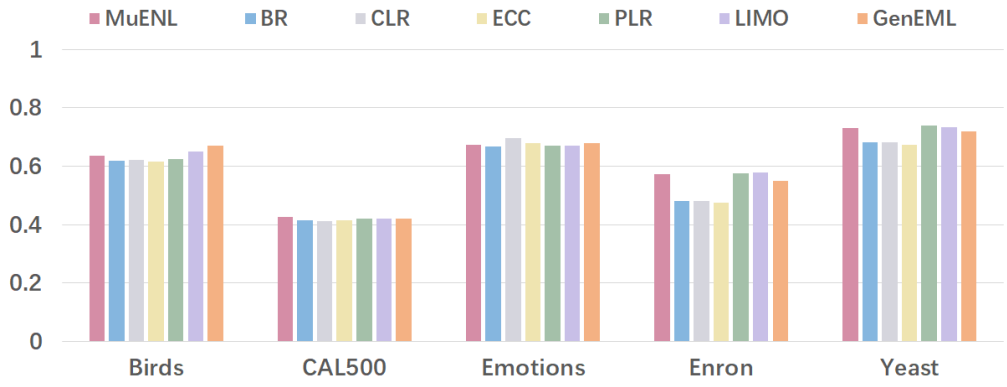



图 3.9 $v_0$-评价结果

对 MuENL 与 LP-SENCForest 在 $t_1$ 至 $t_6$ 的 $v_t$-评价求平均后，得到的结果如图 3.10所示：可以看到，MuENL 的结果比 LP-SENCForest 好很多。这个结果在情理之中，原因有二：(1) LP 转换可能产生非常多的类别，有些类别可能只与非常少量的样本相关。在这种情况下，SENCForest 效果可能不能令人满意，因为它需要每个类有足够的样本才能取得比较好的效果。(2) 经过 LP 转换后，SENCForest 检测出的新类可能只是比较少见的已知标记的组合，可能并不包括任何新标记。

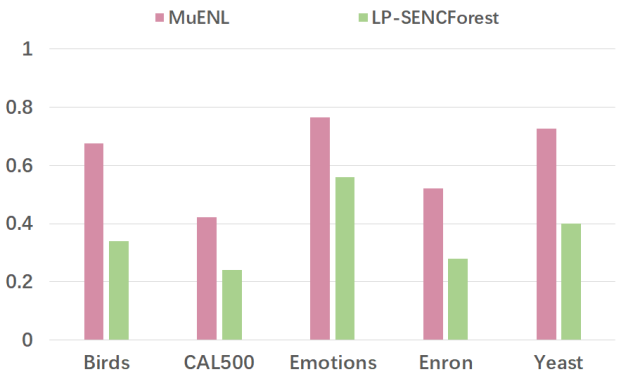



图 3.10 MuENL 与 LP-SENCForest 对比结果 (Average Precision)

对 MuENL 与 MuENL 的变体方法在 $t_1$ 至 $t_6$ 的 $v_t$-评价求平均后，得到的结果如图 3.11 所示。在对比的 4 个 MuENL 变体算法中，MuENL-OR 的效果最好，因为它假设真实标记可能获得，而其它方法均不能得到真实标记。注意这些真实标记在实际中并不能拿到，MuENL-OR 仅用来展示性能的上界。MuENL 并不假设可以获得真实标记，但能够与 MuENL-OR 在大部分情况下可比，在性能上并没有显著差别，这说明了 MuENL 方法的有效性。

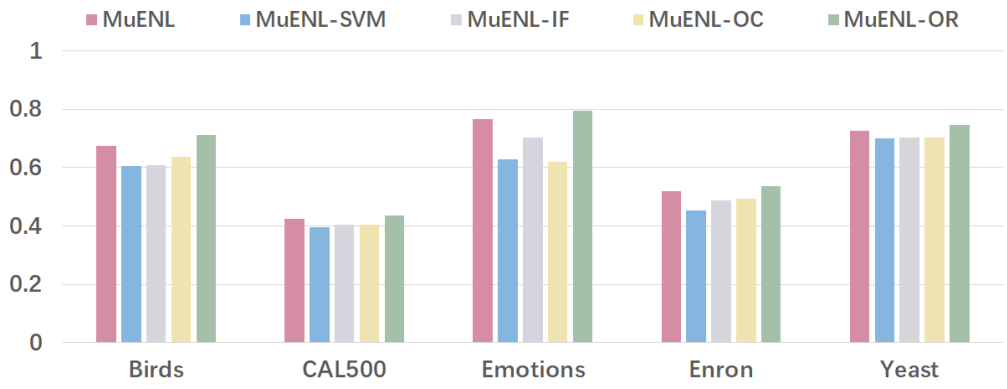



**图 3.11** MuENL 与 MuENL 变体方法对比结果 (Average Precision)

和 $\text{MuENL}_{\text{SVM}}$ 相比，MuENL 在所有的数据集上都取得了更好的成绩。这也说明了 MuENL 中鲁棒的模型更新要比直接使用SVM为新标记建立模型更好，这也是 MuENL 在动态多标记环境中保持良好性能的因素之一。

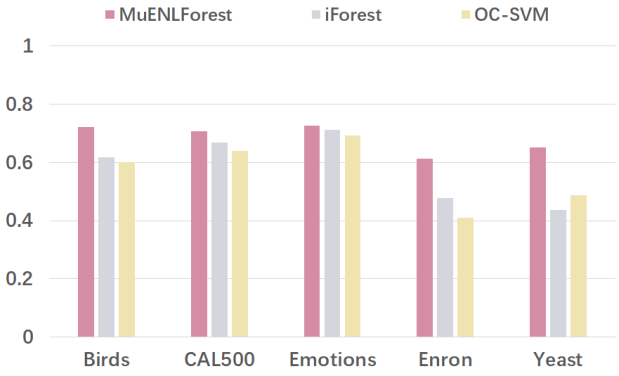



**图 3.12** $t_1$ 时刻 MuENLForest、OC-SVM 和 iForest 的检测性能的比较 (F1-score)

$\text{MuENL}_{\text{IF}}$ 和 $\text{MuENL}_{\text{OC}}$ 分别将 MuENL 中的新标记检测器替换为 iForest 和 OC-SVM。尽管 $\text{MuENL}_{\text{IF}}$ 和 $\text{MuENL}_{\text{OC}}$ 都使用了鲁棒的模型更新策略，MuENL 还是比它们取得了更好的性能 (在 5 个数据集上的 4 个数据集，MuENL 都显著优于其它对比方法)。结果验证了 MuENL 中检测器的有效性。其主要原因是在多标记设置下，新标记往往与已知标记同时出现，这种情况将使得现有的异常检测方法 OC-SVM 和 iForest 难以辨识出新标记 (它仅仅考虑了特征空间)。而 MuENLForest 同时考虑了特征和标记模式，因此能够更好的检测出新标记。为了进一步印证这一结果，比较了不同检测器的性能。具体地，MuENLForest、OC-SVM 和 iForest 在动态多标记学习问题中检测新标记的性能 ($t_1$ 时刻的 F1 值) 比较如图 3.12所示。正如所预计，MuENLForest 与另外

两个对比方法相比能够更好地检测出新标记。

除此之外，还比较了 MNL 、SVM 、PLR 对新标记建模的性能 (在 $t_2$ 时刻的 F1 值)，如图 3.13所示。其中 SVM 和 PLR 都是直接把缓冲区中的样本当成新标记的正样本，而其它样本则当成是负样本，训练分类器。为了公平比较，在实验中使用了同样的检测器 MuENLForest 。可以观察到，MNL 超过了 SVM 和 PLR，因为它采用了鲁棒的更新策略，在检测结果有误差的情况下自然做得更好。

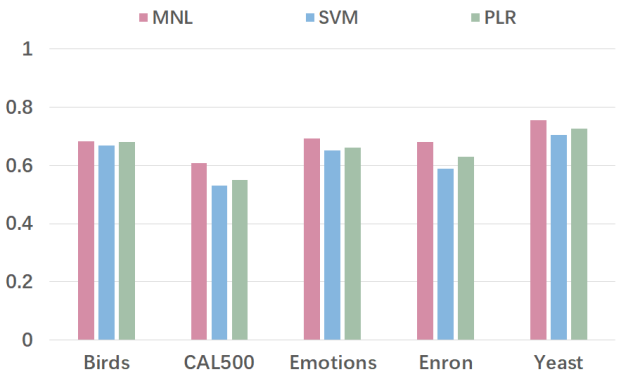



**图 3.13** MNL、SVM 和 PLR 在 $t_2$ 新标记上的分类性能比较 (F1-score)

## 3.4 本章小结

本章主要阐述了测试集标记增广情况下的研究成果。针对静态测试集标记增广问题，真实标记矩阵整列缺失，提出了 DMNL 方法，在多示例多标记框架下高效地最小化包损失项和聚类正则化项，使得已知标记和新标记可以被同时建模。针对动态测试集标记增广问题，真实标记集合随着新标记样本的出现动态增广，提出了 MuENL 方法，在构造检测器时同时考虑特征与标记模式，并设计了鲁棒的更新模型。大量实验分别验证了 DMNL 方法和 MuENL 在解决静态和动态测试集标记增广问题上的有效性。

本工作已成文发表：

- Y. Zhu and K. M. Ting and Z.-H. Zhou. Multi-Label Learning with Emerging New Labels. **IEEE Transactions on Knowledge and Data Engineering**, in press. （CCF-A 类期刊）

- Y. Zhu and K. M. Ting and Z.-H. Zhou. Discover multiple novel labels in multi-instance multi-label learning. In: **Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)**, San Francisco, CA, 2017, pp.2977-2983.（CCF-A 类会议）
- Y. Zhu and K. M. Ting and Z.-H. Zhou. Multi-label learning with emerging new labels. In: **Proceedings of the 16th IEEE International Conference on Data Mining (ICDM'16)**, Barcelona, Spain, 2016, pp.1371-1376.（CCF-B 类会议）

# 第四章　特征增广学习

## 4.1　引言

在分类任务中，训练数据的多少与优劣对分类器的性能有重要的影响。特别是当训练数据信息量不足时，就很难训练出令人满意的分类器。在这种情况下，需要找一些容易获取的且对训练有用的额外的信息提取特征，即增广特征参与分类器的训练，以提高分类器的性能。例如：在图像标注应用中，图像四周的文本描述就是一种增广特征，可以用于改善学习效果。在推荐系统中，用户和用户之间的关系及物品和物品之间的关系是一种增广特征，可以用于提高推荐准确度。

在基于扫描摘要信息的科学文献分类任务中，科学文献常因版权限制只有摘要部分可以公开免费下载。由于摘要的字数限制以及作者强调重心往往在工作的创新点，所以与分类主题相关的高频词不一定会出现在摘要中。于是仅仅利用摘要信息做预测，结果可能并不令人满意。幸运的是，大部分情况下科学文献参考文献的题目和摘要能够公开免费获得，从而可以作为增广特征信息弥补原始信息不足的问题。与之类似，在微博情感预测任务中，微博文本很短，因为信息不足，导致直接进行预测非常困难，但可以把微博的回复或是讨论当作相应的增广特征信息以降低分类的难度；在网页主题分类任务中，则可以把链接到的网页当作原网页的增广特征信息帮助分类。

针对上述例子，如何更好地利用这类增广特征信息提升学习性能是本章的研究目标。多视图学习用于处理由多个视图描述的数据，即多个特征集合。这里，每个视图是一个特征集合，每个样本在每个视图上都有相应的示例表示。多视图学习的目标是通过利用多个视图之间的关系提升学习性能或是降低样本复杂度。观察以上例子，可以把原始文献的摘要 (表示为一个 TF-IDF 特征向量) 当作一个单示例视图，而把参考文献的摘要作为另一个视图。但是对于参考文献来说，并非每篇参考文献都与目标主题相关，如果直接把所有的参考文献摘要拼在一起提取特征将会引入噪声，甚至会造成学习性能下降。为了解决这个问题，对增广特征考虑多示例学习 (Multi-Instance Learning，简称 MIL) 框

架 [76, 91–95]，即将每篇参考文献的摘要表示为一个 TF-IDF 特征向量，所有参考文献的集合则构成一个多示例包作为增广多示例视图；如果这个包为正，则包中至少存在一个示例为正，对应目标主题相关文章的参考文献中至少存在一篇与目标主题相关，否则包中示例均为负。对这类学习任务，将原数据当作单示例视图，把增广特征信息当作多示例视图，提出了增广多示例视图学习 AMIV (Learning with Augmented Multi-Instance View) 框架。

目前，有几篇文献将多示例学习和多视图学习结合在一起。2011，Mayo [96] 通过实验说明将多示例学习和多视图学习结合起来可以在监督图像分类任务上取得很好的效果。他在每个视图上训练一个 MIL 分类器，然后将所有训练好的分类器集成起来得到最终的分类器。2012，Li [97] 提出了一个迭代式的通用方法，可以应用于多示例多视图学习：在每一轮迭代中，首先用多核学习方法在候选标记集上训练，然后根据其它视图上的分类器的输出结果更新当前视图的候选标记集。2013，Zhang [98] 提出了 MI2LS 方法解决多源多示例学习问题。MI2LS 最小化了多示例分类误差、结构化风险以及对相同样本在不同视图上的分类器差异的惩罚。以上工作均为多视图学习，且在每个视图上，样本表示为一个多示例包，每个示例都能在其它视图上找到一一对应的示例。而在 AMIV 框架中，一个视图是单示例视图，另一个视图为多示例视图，示例之间找不到一一对应的关系，因而上述方法均无法应用于 AMIV 框架。

为了在 AMIV 框架下求解问题，提出了 AMIV-lss 方法：首先为这两个视图建立一个共同隐藏语意子空间，使得同一样本的单示例视图和多示例视图在该子空间上的表示相近；然后再在该子空间中处理分类任务。最后在 TechPaper 和 WebPage 两种不同类型的文本数据集上进行了实验，结果表明本文方法的分类性能有明显提升。

## 4.2 本文方法

在本节中，提出了 AMIV 框架并设计了 AMIV-lss 方法来解决利用增广多示例视图学习问题，通过建立一个隐藏语义子空间 LSS (Latent Semantic Subspace) 使得同一样本在该子空间中的单示例视图的表示与多示例视图的表示相近。然后在此空间里处理传统监督学习任务。

令 $\mathcal{X}$ 表示原始单示例视图的特征空间；$\mathcal{X}^*$ 为增广多示例视图的特征空间；$\mathcal{Y}$ 表示标记空间。AMIV 的任务是给定数据集 $\boldsymbol{D} = \{(\boldsymbol{x}_i, B_i, y_i)\}^n$ 学习一个函

数 $f : \mathcal{X} \times 2^{\mathcal{X}^*} \to \mathcal{Y}$，将单示例经过增广多示例视图映射到标记空间，去预测新的未知示例 $\boldsymbol{x}_{new}$ 的类别标记。其中，$\boldsymbol{x}_i \in \mathcal{X}$ 表示原始单视图的一个示例；$\boldsymbol{B}_i = \{\boldsymbol{b}_{i,1}, \boldsymbol{b}_{i,2}, \cdots, \boldsymbol{b}_{i,n_i}\} \subseteq \mathcal{X}^*$ 表示增广多示例视图上相应大小为 $n_i$ 的多示例包；$y_i \in \mathcal{Y}$ 是相应类别标记。

### 4.2.1 AMIV-lss 方法形式化

在传统多视图学习中，建立公共子空间的方法能够很好刻画多个视图之间数据的固有结构信息，可以大大降低学习任务的难度 [30]。为了解决 AMIV 这样的异构多视图问题，受这个思想的启发，提出了 AMIV-lss 方法。AMIV-lss 采用了两阶段优化策略。在第一阶段，学习一个最优隐藏语义子空间（LSS），由 $\mathcal{J}$ 表示，使得同一个样本的单示例视图 $\boldsymbol{x}$ 和对应多示例视图 $\boldsymbol{B}$ 在该子空间上的表示相互接近。但是直接指定 $\boldsymbol{B}$ 在 $\mathcal{J}$ 上的表示比较困难，因为 $\boldsymbol{B}$ 中多示例的类别标记未知。因此，考虑为 $\boldsymbol{B}$ 寻找一个原型 $\boldsymbol{s}$，这样包在 $\mathcal{J}$ 上的表示即可以转变为原型 $\boldsymbol{s}$ 在 $\mathcal{J}$ 上的表示。接着第二阶段在第一阶段学到的最优子空间上训练一个最大间隔分类器。

令 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ 表示单示例视图上的 $n$ 个示例；$\boldsymbol{P}_X = [\boldsymbol{p}_{\boldsymbol{x}_1}, \boldsymbol{p}_{\boldsymbol{x}_2}, \cdots, \boldsymbol{p}_{\boldsymbol{x}_n}] \in \mathbb{R}^{d_s \times n}$ 表示 $\boldsymbol{X}$ 在 $\mathcal{J}$ 上的表示，其中 $\boldsymbol{p}_{\boldsymbol{x}_i}$ 是 $\boldsymbol{x}_i$ 在 $\mathcal{J}$ 的对应表示。$\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_t]^\top \in \mathbb{R}^{d_s \times d}$ 表示 $\boldsymbol{J}$ 的基，则有 $\boldsymbol{X}^\top = \boldsymbol{P}_X^\top \boldsymbol{Q}$。这种表示有很多实际应用，尤其是在文本数据上 [99]。其中 $\boldsymbol{X}$ 是一个对象-单词矩阵，表示对象和单词之间的语义相关性；$\boldsymbol{p}_{\boldsymbol{x}}$ 代表 $\boldsymbol{x}$ 的 $d_s$ 个语义主题；$\boldsymbol{q} \in \boldsymbol{Q}$ 代表了一个语义主题对应的单词的出现模式。通过 $\boldsymbol{X}^\top = \boldsymbol{P}_X^\top \boldsymbol{Q}$ 可对对象的语义主题模式与语义主题的单词模式之间的相互作用进行建模。因为 $\boldsymbol{P}_X$ 和 $\boldsymbol{Q}$ 分别代表语义主题模式和单词的出现模式，所以它们都应该是非负的。注意到仅仅有少量的语义主题和对象的真正主题相关，且相对于整个字典而言，与每个语义主题相关的单词非常少，因此，$\boldsymbol{P}_X$ 和 $\boldsymbol{Q}$ 都应该有稀疏的结构。

为了建立这样的隐藏语义子空间（LSS）并获取示例在该空间里的表示，采用非负矩阵分解（NMF）框架 [100] 如式 (4.1) 所示，

$$\min_{\boldsymbol{P}_X \geq 0, \boldsymbol{Q} \geq 0} \left\| \boldsymbol{X}^\top - \boldsymbol{P}_X^\top \boldsymbol{Q} \right\|_F^2 + \mathcal{R}(\boldsymbol{P}_X, \boldsymbol{Q}), \tag{4.1}$$

其中 $\mathcal{R}(\boldsymbol{P}_X, \boldsymbol{Q}) = \|\boldsymbol{P}_X\|_1 + \|\boldsymbol{Q}\|_1$。

令 $\boldsymbol{Z} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \cdots, \boldsymbol{B}_n] \in \mathbb{R}^{r \times d}$ 表示增广多示例视图上的示例，其中 $r = \sum_{i=1}^{n} n_i$；并用 $\boldsymbol{P}_Z = [\boldsymbol{P}_{\boldsymbol{B}_1}, \boldsymbol{P}_{\boldsymbol{B}_2}, \cdots, \boldsymbol{P}_{\boldsymbol{B}_n}] \in \mathbb{R}^{d_s \times r}$ 表示 $\boldsymbol{Z}$ 在 $\mathcal{J}$ 上的表示，其中

$\boldsymbol{P}_{\boldsymbol{B}_i} = \left[\boldsymbol{p}_{\boldsymbol{b}_{i1}}, \boldsymbol{p}_{\boldsymbol{b}_{i2}}, \cdots, \boldsymbol{p}_{\boldsymbol{b}_{i,n_i}}\right]$，$\boldsymbol{p}_{\boldsymbol{b}_{ij}}$ 是 $\boldsymbol{b}_{ij}$ 在 $\mathcal{J}$ 上的表示。类似式 (4.1)，有式 (4.2)：

$$\min_{\boldsymbol{P}_Z \geq 0, \boldsymbol{Q} \geq 0} \left\| \boldsymbol{Z}^\top - \boldsymbol{P}_Z^\top \boldsymbol{Q} \right\|_F^2 + \mathcal{R}(\boldsymbol{P}_Z, \boldsymbol{Q}). \tag{4.2}$$

尽管上文已经给出了示例在 $\mathcal{J}$ 上的表示，但仍然不能直接指定在该空间里多示例包的表示，因为包里面的所有示例的标记都是未知的。为了给出包在 $\mathcal{J}$ 上的表示，首先为包 $\boldsymbol{B}$ 定义了关键原型 $\boldsymbol{s}$，且包 $\boldsymbol{B}$ 的类别标记就是由 $\boldsymbol{s}$ 决定的。令 $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n] \in \mathbb{R}^{d \times n}$ 表示关键原型空间，其中 $\boldsymbol{s}_i$ 表示第 $i$ 个包 $\boldsymbol{B}_i$ 的关键原型；令 $\boldsymbol{P}_S = [\boldsymbol{p}_{\boldsymbol{s}_1}, \boldsymbol{p}_{\boldsymbol{s}_2}, \cdots, \boldsymbol{p}_{\boldsymbol{s}_n}] \in \mathbb{R}^{d_s \times n}$ 代表 $\boldsymbol{S}$ 在 $\mathcal{J}$ 上的表示，则 $\boldsymbol{p}_{\boldsymbol{s}_i}$ 就是 $\boldsymbol{B}_i$ 在隐藏语义子空间 $\mathcal{J}$ 上的表示。类似式 (4.2)，可得到式 (4.3)：

$$\min_{\boldsymbol{P}_S \geq 0, \boldsymbol{Q} \geq 0} \left\| \boldsymbol{S}^\top - \boldsymbol{P}_S^\top \boldsymbol{Q} \right\|_F^2 + \mathcal{R}(\boldsymbol{P}_S, \boldsymbol{Q}). \tag{4.3}$$

一种指定多示例包关键原型的方法是把多示例包的中心点当作这个包的关键原型。这种方法虽然直接简单，但是需要承担很大的风险：如果一个正包中负标记示例的数量远多于正标记示例，那么这个包的中心点的标记应该为负，恰恰与包的标记相反。这种情况下，指定中心点作为关键原型将会误导学习，并很可能造成学习性能下降。为了降低这样的风险，考虑了不同示例的权重。给定单示例视图的上的示例 $\boldsymbol{x}$，对应多示例包 $\boldsymbol{B}$ 中的示例离它越近就有更高的概率与 $\boldsymbol{x}$ 拥有相同的类别标记，那么该示例就有更高的权重。新的问题是如何度量远近关系。当示例维度很高且非常稀疏的情况下（文本数据往往存在这样的情况），用欧氏距离衡量示例的远近关系非常不可靠。

为了在高维稀疏数据上求解包的关键原型，引入了局部线性假设[101]。假设包 $\boldsymbol{B}_i$ 的原型 $\boldsymbol{s}_i$ 为示例 $\boldsymbol{x}_i$ 在包 $\boldsymbol{B}_i$ 中的近邻示例的线性组合。其中，近邻在低维子空间 $\mathcal{J}$ 上由欧氏距离定义。定义 $\boldsymbol{\delta}_i$ 为包 $\boldsymbol{B}_i$ 中的近邻指示向量，$\mathrm{N}_k$ 表示 $k$ 个最近邻。因此，如果 $\boldsymbol{p}_{\boldsymbol{b}_{i,j}} \in \mathrm{N}_k(\boldsymbol{p}_{\boldsymbol{x}_i})$，则有 $\delta_{i,j} = 1$；否则 $\delta_{i,j} = 0$。令 $\boldsymbol{\alpha}_i$ 为 $\boldsymbol{B}_i$ 的线性组合系数向量，则有 $\boldsymbol{s}_i = B_i \boldsymbol{\alpha}_i$. $\boldsymbol{\alpha}_i$ 中的各个系数由距离反比加权决定：$\alpha_{i,j} = \exp(-dist_{i,j}^2)\delta_{i,j} / (\sum_{j=1}^{n_i} \exp(-dist_{i,j}^2)\delta_{i,j})$，其中 exp 为指数函数，$dist_{i,j}$ 为 $\boldsymbol{p}_{\boldsymbol{x}_i}$ 到 $\boldsymbol{p}_{\boldsymbol{b}_{i,j}}$ 的欧氏距离：$dist_{i,j} = \|\boldsymbol{p}_{\boldsymbol{x}_i} - \boldsymbol{p}_{\boldsymbol{b}_{i,j}}\|_2$。则 $\boldsymbol{S}$ 可重写为重写为式 (4.4)：

$$\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n],\ \boldsymbol{s}_i = \sum_{j=1}^{n_i} \frac{\exp(-\|\boldsymbol{p}_{x_i} - \boldsymbol{p}_{\boldsymbol{b}_{ij}}\|_2^2)\delta_{i,j}\boldsymbol{b}_{i,j}}{\sum_{j=1}^{n_i} \exp(-\|\boldsymbol{p}_{x_i} - \boldsymbol{p}_{\boldsymbol{b}_{i,j}}\|_2^2)\delta_{i,j}}, \forall \boldsymbol{s}_i \in \boldsymbol{S}. \tag{4.4}$$

`AMIV-lss` 包括两阶段优化：第一阶段，学习一个最佳隐藏语义子空间，

**算法 4.1** AMIV-lss

**输入**: 训练集 $(\boldsymbol{x}_i, \boldsymbol{B}_i, y_i), i = 1, 2, \cdots, n$
**输出**: $\boldsymbol{Q}, \boldsymbol{w}$

初始化 $\boldsymbol{P}_X, \boldsymbol{P}_Z, Q \leftarrow$ 式 (4.10)；
**repeat**:
  更新原型 $\boldsymbol{S} \leftarrow$ 式 (4.4)；
  固定 $\boldsymbol{S}$ 和 $\boldsymbol{P}_Z$，更新 $\boldsymbol{P}_X, \boldsymbol{P}_S, \boldsymbol{Q} \leftarrow$ 式 (4.8)；
  固定 $\boldsymbol{S}$、$\boldsymbol{P}_X$、$\boldsymbol{P}_S$ 和 $\boldsymbol{Q}$，更新 $\boldsymbol{P}_Z \leftarrow$ 式 (4.9)；
**until** 算法收敛或者达到最大迭代轮数
给定 $\boldsymbol{P}_X$ 和 $\boldsymbol{P}_S$, 解 $\boldsymbol{w} \leftarrow$ 式 (4.7)。

使得单示例视图样本 $\boldsymbol{x}$ 与其对应多示例视图包 $\boldsymbol{B}$ 在该空间中的表示相似；第二阶段在学得的最优子空间上训练分类器。其形式化如下。

定义 $\boldsymbol{P} = [\boldsymbol{P}_X, \boldsymbol{P}_Z, \boldsymbol{P}_S]$，综合式 (4.1) - (4.3)，可得到：

$$\min_{\boldsymbol{P}\geq 0, \boldsymbol{Q}\geq 0} \left\| [\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}]^\top - [\boldsymbol{P}_X, \boldsymbol{P}_Z, \boldsymbol{P}_S]^\top \boldsymbol{Q} \right\|_F^2 + \mathcal{R}(\boldsymbol{P}, \boldsymbol{Q}). \tag{4.5}$$

如前文所述, 在隐藏子空间 $\mathcal{J}$ 中，单示例视图样本和其相应的增广多示例视图包在该空间中的表示应该彼此相近。换言之 $\boldsymbol{P}_X$ 与 $\boldsymbol{P}_S$ 应该相似，因此除稀疏项 $\|\boldsymbol{P}\|_1$ 和 $\|\boldsymbol{Q}\|_1$ 外，增加一项正则项 $\|\boldsymbol{P}_X - \boldsymbol{P}_S\|_F^2$。于是，第一阶段优化可表示为式 (4.6)：

$$\min_{\boldsymbol{P}\geq 0, \boldsymbol{Q}\geq 0} \left\| [\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}]^\top - \boldsymbol{P}^\top \boldsymbol{Q} \right\|_F^2 + \lambda_1 \|\boldsymbol{P}_X - \boldsymbol{P}_S\|_F^2 + \lambda_2 \left( \|\boldsymbol{P}\|_1 + \|\boldsymbol{Q}\|_1 \right). \tag{4.6}$$

令 $(\boldsymbol{P}^*, \boldsymbol{Q}^*)$ 表示式 (4.6) 的解，则 $\boldsymbol{Q}^*$ 为最优隐藏语义子空间 $\mathcal{J}^*$ 的基，$\boldsymbol{P}^* = [\boldsymbol{P}_X^*, \boldsymbol{P}_Z^*, \boldsymbol{P}_S^*]$ 为示例在 $\mathcal{J}^*$ 上的表示，$\lambda_1$ 和 $\lambda_2$ 为权衡参数。

在第二阶段优化中，在 $\mathcal{J}^*$ 空间上训练一个最大化间隔分类器模型 $\boldsymbol{w}^*$ 来处理分类任务。该分类器模型需要满足一个额外约束：同一样本的两个相应视图在 $\mathcal{J}^*$ 上的表示有相同的标记，即 $\boldsymbol{x}_i$ 与 $\boldsymbol{s}_i$ 均有类别标记 $y_i$。综上所述，第二阶段优化可写为式 (4.7)：

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \boldsymbol{p}_{\boldsymbol{x}_i}^* \geq 1 - \xi_i \\ & y_i \boldsymbol{w}^\top \boldsymbol{p}_{\boldsymbol{s}_i}^* \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, n. \end{aligned} \tag{4.7}$$

### 4.2.2 优化求解

在第一阶段优化中，式 (4.6) 的目标函数中涉及到变量 $\boldsymbol{S}$、$\boldsymbol{P}_X$、$\boldsymbol{P}_Z$ 和 $\boldsymbol{P}_S$，联合优化所有这些变量非常困难。因此采用了交替优化方法进行优化。具体来说，在每轮迭代中，首先通过式 (4.4) 更新关键原型 $\boldsymbol{S}$ 以学习包的表示。然后固定 $\boldsymbol{S}$ 和 $\boldsymbol{P}_Z$，优化式 (4.6) 等价于优化式 (4.8)，即通过求解式 (4.8) 更新 $\boldsymbol{P}_X$、$\boldsymbol{P}_S$ 和 $\boldsymbol{Q}$。

$$\min_{\boldsymbol{P}_X\geq 0,\boldsymbol{P}_S\geq 0,\boldsymbol{Q}\geq 0} \left\|[\boldsymbol{X},\boldsymbol{S}]^\top-[\boldsymbol{P}_X,\boldsymbol{P}_S]^\top\boldsymbol{Q}\right\|_F^2+\lambda_1\|\boldsymbol{P}_X-\boldsymbol{P}_S\|_F^2+\lambda_2(\|\boldsymbol{P}_X\|_1+\|\boldsymbol{P}_S\|_1+\|\boldsymbol{Q}\|_1). \tag{4.8}$$

最后固定 $\boldsymbol{S}$、$\boldsymbol{P}_X$、$\boldsymbol{P}_S$ 和 $\boldsymbol{Q}$，优化式 (4.6) 等价于优化式 (4.9)，从而得到 $\boldsymbol{Z}$ 在 $\mathcal{J}$ 上的表示，即更新 $\boldsymbol{P}_Z$：

$$\min_{\boldsymbol{P}_Z\geq 0} \left\|\boldsymbol{Z}^\top-\boldsymbol{P}_Z^\top\boldsymbol{Q}\right\|_F^2+\lambda_2\|\boldsymbol{P}_Z\|_1. \tag{4.9}$$

第一阶段优化完成后，得到 $\boldsymbol{P}_X^*$ 和 $\boldsymbol{P}_S^*$。然后在第二阶段的优化中，通过求解式 (4.7) 得到最优分类器 $\boldsymbol{w}^*$。

此外，为了热启动算法，通过解式 (4.10) 初始化 $\boldsymbol{P}_X$、$\boldsymbol{P}_Z$ 和 $\boldsymbol{Q}$：

$$\min_{\boldsymbol{P}_X\geq 0,\boldsymbol{P}_Z\geq 0,\boldsymbol{Q}\geq 0} \left\|[\boldsymbol{X},\boldsymbol{Z}]^\top-[\boldsymbol{P}_X,\boldsymbol{P}_Z]^\top\boldsymbol{Q}\right\|_F^2+\lambda_2(\|\boldsymbol{P}_X\|_1+\|\boldsymbol{P}_Z\|_1+\|\boldsymbol{Q}\|_1). \tag{4.10}$$

算法 4.1总结了 `AMIV-lss` 方法。

在具体实现中，通过交替优化求解式 (4.8)：固定 $\boldsymbol{P}_X$ 和 $\boldsymbol{Q}$，通过求解式 (4.11) 更新 $\boldsymbol{P}_S$：

$$\min_{\boldsymbol{P}_S\geq 0}\|\boldsymbol{S}^\top-\boldsymbol{P}_S^\top\boldsymbol{Q}\|_F^2+\lambda_1\|\boldsymbol{P}_S-\boldsymbol{P}_X\|_F^2+\lambda_2\|\boldsymbol{P}_S\|_1. \tag{4.11}$$

固定 $\boldsymbol{P}_S$ 和 $\boldsymbol{Q}$，通过求解式 (4.12) 更新 $\boldsymbol{P}_X$：

$$\min_{\boldsymbol{P}_X\geq 0}\|\boldsymbol{X}^\top-\boldsymbol{P}_X^\top\boldsymbol{Q}\|_F^2+\lambda_1\|\boldsymbol{P}_X-\boldsymbol{P}_S\|_F^2+\lambda_2\|\boldsymbol{P}_X\|_1. \tag{4.12}$$

最后，固定 $\boldsymbol{P}_X$ 和 $\boldsymbol{P}_S$，通过求解式 (4.13) 更新 $\boldsymbol{Q}$：

$$\min_{\boldsymbol{Q}\geq 0}\|[\boldsymbol{X},\boldsymbol{S}]^\top-[\boldsymbol{P}_X,\boldsymbol{P}_S]^\top\boldsymbol{Q}\|_F^2+\lambda_2\|\boldsymbol{Q}\|_1. \tag{4.13}$$

类似地，优化求解式 (4.10)，迭代地固定 $[\boldsymbol{P}_X, \boldsymbol{P}_Z]$，通过求解式 (4.14) 更新 $\boldsymbol{Q}$：

$$\min_{\boldsymbol{Q}\geq 0} \ \left\|[\boldsymbol{X}, \boldsymbol{Z}]^\top - [\boldsymbol{P}_X, \boldsymbol{P}_Z]^\top \boldsymbol{Q}, \right\|_F^2 + \lambda_2 \|\boldsymbol{Q}\|_1; \tag{4.14}$$

固定 $\boldsymbol{Q}$，通过求解式 (4.15) 更新 $[\boldsymbol{P}_X, \boldsymbol{P}_Z]$：

$$\min_{[\boldsymbol{P}_X, \boldsymbol{P}_Z]\geq 0} \ \left\|[\boldsymbol{X}, \boldsymbol{Z}]^\top - [\boldsymbol{P}_X, \boldsymbol{P}_Z]^\top \boldsymbol{Q}\right\|_F^2 + \lambda_2 \left\|[\boldsymbol{P}_X, \boldsymbol{P}_Z]\right\|_1. \tag{4.15}$$

为了优化各个子问题式 (4.13)–(4.15)，采用贪心坐标轴下降方法[100]。对每个单变量元素的更新规则如下：

$$\begin{aligned} P_{i,r} &\leftarrow \max(0, P_{i,r} - G_{Pi,r}/H_{Pi,i}), \\ Q_{i,r} &\leftarrow \max(0, Q_{i,r} - G_{Q_{i,r}}/H_{Q_{i,i}}), \end{aligned}$$

其中 $G$ 为目标函数 $f(\boldsymbol{P}, \boldsymbol{Q})$ 的梯度矩阵，$H$ 为 $f(\boldsymbol{P}, \boldsymbol{Q})$ 的海森矩阵。

为了求解式 (4.7)，构造了新的数据集：$\left([\boldsymbol{P}_X, \boldsymbol{P}_S], [\boldsymbol{y}; \boldsymbol{y}]^\top\right)$。然后使用 `LIBLINEAR` 工具包 [67] 得到 $\boldsymbol{w}^*$。

## 4.3 实验测试

本节介绍了实验设置，并汇报了实验结果，验证了 `AMIV-lss` 方法利用增广多示例视图学习的有效性。

### 4.3.1 实验设置

首先介绍 TechPaper 和 WebPage 数据集：

**TechPaper 数据集：** 首先选择机器学习领域的 12 个主题，包括主动学习 (AL)、聚类 (CLST)、深度学习 (DPL)、度量学习 (MTCL)、多示例学习 (MIL)、多标记学习 (MLL)、多任务学习 (MTL)、多视图学习 (MVL)、在线学习 (OL)、强化学习 (RL)、半监督学习 (SSL) 和迁移学习 (TFL)。然后应用微软学术搜索 API 下载这些主题的学术论文的摘要和它们相应参考文献的摘要。对每一篇文章，提取 TF-IDF 特征构造示例，总共构造了 46,531 个示例，并采用 one-vs-rest 策略为每个主题构造数据集。构造数据集时，选取等量的正负示例构成单示例视图，然后把每个示例对应文章参考文献提取的示例打包构成该示例的增广多

示例视图。平均每个数据集有 416 个对象，7,732 个示例。

**WebPage 数据集：** 在构造 WebPage 数据集时，把 MILWEB 数据集[102] 重构成 AMIV 形式。原始数据集包括 113 个网页主页和 3,423 个链接网页。这些网页由 9 名志愿者根据他们各自的兴趣标上类别标记。换言之，一共有 9 个子数据集。这里，仍然提取 TF-IDF 特征为每个网页构造示例。因此，单示例视图由原始 113 个网页示例构成，而多示例视图由相应链接网页示例包组成。具体的，WebPage 数据集的每个子数据集都包括 113 个对象，3,536 个示例。

实验将 AMIV-lss 与 Standard、SV-mil)、Concatenation、MV-sil、Amil、ALapSVM 以及 ASVM+ 对比，以验证本文所提方法的有效性。这些方法将已有的传统学习方法改造以适用于 AMIV 学习问题：

(1) Standard: 仅仅在单示例视图上训练线性 SVM；

(2) SV-mil: 仅仅在多示例视图上应用常用多示例学习方法 miSVM [93]；

(3) Concatenation (Conc): 首先将每个单示例视图上的的示例和它对应增广多示例视图包中的示例拼成一个特征向量。然后用新构成的数据训练线性 SVM。由于包中示例数目未必相同，拼接而成的特征向量长度可能不同。此时，需要在长度短的特征向量后补 0，以保证特征向量长度一致；

(4) MV-sil: 把每个增广多示例包中的示例求平均，这样增广多示例视图就可以转换成为一个普通的单示例视图，即把 AMIV 形式数据集转换成为一个传统的多视图学习数据集。然后就可以应用 NMF 方法 [100] 学习多视图之间的语义子空间。最后在这个子空间上训练一个线性 SVM。MV-sil 方法可以看成是 AMIV-lss 的一个退化算法（固定关键原型为包的中心点）；

(5) Amil: 把单示例视图上的每个示例当做是一个特殊的包（每个包中只有 1 个示例），然后把这些包与增广多示例视图的包一起作为新的多示例学习数据集。这样把 AMIV 形式数据集转换成为一个普通多示例学习数据集。最后在新构成的数据集上训练 miSVM 分类器 [93]；

(6) AlapSVM: AMIV 形式的数据，把其单示例视图数据作为标记示例，把增广多示例视图数据当作未标记示例，然后在新构成的半监督数据集上训练 Laplacian SVM (LapSVM) 分类器 [54]。

(7) ASVM+: 把增广多示例视图按包求平均，得到的向量作为辅助信息，然后利用 SVM+ 方法，用这些辅助信息为优化提供额外的约束，从而训练分类器 [13, 103]。

**表 4.1** TechPaper 数据集上的准确率 (均值 ± 标准差)。斜体表示 AMIV-lss 显著优于对比方法（t 检验，$\alpha = 0.05$）。粗体显示的是对应数据集上的最佳结果。

| | AMIV-lss | Standard | SV-mil | Conc | MV-sil | Amil | AlapSVM | ASVM+ |
|---|---|---|---|---|---|---|---|---|
| AL | **.952±.018** | *.924±.034* | *.831±.026* | *.922±.016* | *.944±.012* | *.851±.038* | *.931±.036* | *.829±.016* |
| CLST | **.979±.017** | *.944±.023* | *.895±.058* | *.963±.018* | *.966±.016* | *.895±.062* | *.947±.019* | *.915±.026* |
| DPL | .918±.017 | *.915±.014* | *.831±.039* | **.932±.032** | .918±.017 | *.845±.041* | .915±.010 | *.895±.016* |
| MTCL | **.981±.008** | *.962±.014* | *.888±.016* | *.945±.016* | *.971±.011* | *.892±.016* | *.964±.015* | *.932±.015* |
| MIL | **.909±.023** | *.852±.034* | *.799±.042* | *.877±.035* | .902±.036 | *.840±.033* | *.851±.029* | *.842±.044* |
| MLL | **.955±.032** | *.927±.026* | *.892±.015* | .950±.037 | .947±.021 | *.914±.018* | *.935±.013* | *.930±.016* |
| MTL | **.885±.038** | *.868±.042* | *.804±.052* | .875±.029 | .873±.057 | *.830±.033* | .883±.028 | *.848±.032* |
| MVL | **.987±.015** | *.959±.008* | *.878±.027* | *.943±.015* | *.976±.008* | *.913±.010* | *.961±.009* | *.960±.008* |
| OL | **.998±.005** | *.980±.005* | *.890±.008* | *.948±.020* | *.986±.016* | *.927±.024* | *.986±.005* | *.963±.015* |
| RL | .888±.033 | .888±.018 | .880±.049 | **.918±.016** | .888±.018 | .888±.046 | .888±.020 | .888±.018 |
| SSL | **.971±.013** | *.958±.029* | *.839±.024* | *.933±.036* | *.962±.027* | *.862±.028* | .967±.029 | *.968±.017* |
| TL | **.825±.013** | *.800±.028* | *.805±.010* | *.808±.028* | .823±.013 | *.792±.015* | .823±.031 | *.799±.015* |

### 4.3.2 实验结果

表 4.1和 4.2分别总结了前述所有方法在 TechPaper 和 WebPage 数据集上的结果。从表中可见，AMIV-lss 方法在 16 个数据集上取得了最好结果，在剩下的 5 个数据集上取得了次好结果（合计 21 个数据集）。

Concatenation 方法把两个视图拼接到一起，同时利用了两个视图的信息，但取得的结果与两个单视图方法相比，时好时坏。采用 MV-sil 方法时也得到相同的表现。这种性能的起伏可能是由于增广信息不仅仅引入了有用的信息，还带来了噪声。而 AMIV-lss 方法通过利用局部线性设计关键原型，所学习到的语义表示可以减轻噪声带来的负面影响，从而更加鲁棒。AMIV-lss 比仅利用单示例视图的 Standard 方法以及只利用多示例视图信息的 SV-mil 方法得到的结果均有显著提高，且在大部分情况下，AMIV-lss 优于 Concatenation 和 MVsil。

Amil 同时利用了单示例视图和增广多示例视图的信息，总是比仅使用多示例视图的 SV-mil 方法要好。同样利用了两个视图，但是整体上 AMIV-lss 优于 Amil，因为 Amil 方法独立对待两个视图，而 AMIV-lss 方法利用了视图之间的结构信息。

AlapSVM 在 AMIV 数据上应用 LapSVM。它把增广多示例视图上的示例当做未标记示例，利用了示例之间的流形结构，并取得了比单视图学习方法更好的效果。但是，对 AMIV 数据而言，AlapSVM 并不理想。因为在多示例视图中的示例并不仅仅是未标记示例，而是存在着包结构以及包上的标记信息，

**表 4.2** WebPage 数据集上的准确率 (均值 ± 标准差)。斜体表示 AMIV-lss 显著优于对比方法（t 检验，$\alpha = 0.05$）。粗体显示的是对应数据集上的最佳结果。

| | AMIV-lss | Standard | SV-mil | Conc | MV-sil | Amil | AlapSVM | ASVM+ |
|---|---|---|---|---|---|---|---|---|
| V1 | **.871±.019** | *.839±.017* | *.846±.019* | *.786±.045* | *.849±.029* | *.861±.022* | .866±.027 | *.821±.014* |
| V2 | **.868±.019** | *.819±.031* | .866±.026 | *.780±.043* | *.841±.030* | *.841±.046* | *.840±.035* | *.841±.015* |
| V3 | **.861±.031** | *.812±.034* | *.840±.043* | *.780±.061* | *.832±.042* | **.861±.067** | *.840±.051* | *.823±.016* |
| V4 | .911±.031 | .911±.014 | .909±.022 | **.924±.032** | *.787±.051* | .911±.072 | .911±.056 | *.806±.018* |
| V5 | .894±.036 | *.884±.036* | *.877±.042* | .897±.020 | *.859±.086* | **.917±.066** | .884±.081 | *.798±.018* |
| V6 | .897±.034 | .897±.029 | *.883±.033* | .897±.029 | *.849±.086* | **.911±.064** | .897±.085 | *.838±.019* |
| V7 | **.789±.082** | *.735±.080* | *.756±.068* | *.782±.041* | .788±.066 | *.749±.077* | *.735±.059* | .787±.028 |
| V8 | **.804±.088** | *.781±.098* | *.736±.057* | .802±.051 | *.702±.106* | *.769±.069* | .799±.069 | *.761±.048* |
| V9 | .786±.050 | *.774±.092* | *.749±.019* | **.799±.043** | *.742±.041* | *.749±.012* | .786±.053 | .790±.048 |

AlapSVM 方法忽略了这些信息。 AMIV-lss 方法考虑了包的信息，从而取得了比 AlapSVM 好的结果。

ASVM+ 将增广特征视图作为辅助特征信息，用以进一步约束分类器。但是由于辅助信息是由多示例包中所有示例平均得到，将噪声引入了特征，因此得到的结果有时不如仅使用原始视图信息的分类结果。除此之外，那些辅助信息不参与决策，即无法用于预测。因此，整体上来看，AMIV-lss 取得了比 SVM+ 更好的性能。

## 4.4 本章小结

为了在学习任务中充分利用增广特征信息，减少噪声的影响，提出了增广多示例视图 (AMIV) 学习框架，并设计了 AMIV-lss 方法，通过在两个异构视图之间建立公共隐藏语义子空间，从而求解 AMIV 学习问题。实验结果表明了 AMIV-lss 方法的有效性，它比单视图方法或是利用多视图/多示例的方法能取得更好的结果。

本工作已成文发表：

- Y. Zhu and J. Wu and Y. Jiang and Z.-H. Zhou. Learning with augmented multi-instance view. In: **Proceedings of the 6th Asian Conference on Machine Learning (ACML'14)**, Nha Trang, Vietnam, 2014, JMLR: W&CP 39, pp.234-249.（机器学习领域重要国际会议）

# 第五章　样本增广学习

## 5.1　引言

样本增广学习研究根据动态增广的样本高效地更新已有模型，避免从头训练。实际应用场景中，数据往往同时表现为多个视图的形式，例如，每分钟都有总计长达数百小时时长的视频上传到 YouTube 上，包括图像、音频和文本视图；每天都有数小时双语新闻进行报道，每种语言的描述都是一个视图；每年许多学术论文被发表出来，其中文章内容和参考文献可以看作是对一篇论文的不同视图的描述。针对这种多视图样本增广学习问题，本章工作主要研究如何根据新样本高效地利用多个视图之间的结构信息更新模型、提升学习性能。

多视图学习是处理这种一个对象由多个视图描述（表示为多个特征向量）的重要学习范式，研究者们提出了很多方法[23, 30–33]。例如在网页分类上，co-training 通过在文本内容视图和网页链接视图上进行组合标记传播取得了比在单个视图上直接分类更好的结果 [23, 25]；在多语言文本分类任务中，SCMV 在不同语言视图间学到的公共子空间上进行分类 [31]，效果显著优于仅使用单个视图方法。

但是利用多视图之间的潜在关系总是伴随着高计算成本。大多数多视图方法都被应用于少样本低维度的静态小数据集上：训练样本通常小于 5000、特征数不超过 1000，且需要反复扫描数据集。然而，许多实际应用都涉及到大规模的多视图数据，例如图像与文本、网页与链接、双语新闻等，且新的数据源源不断地产生。在这样的情况下，需要多视图学习算法能够高效、即时处理这些新增的数据，即增广样本，更新模型，而不是反复将数据载入内存重复训练，这对多视图学习带来了巨大挑战。在线学习是一种非常高效地用以构建大规模学习系统的策略，且能够对动态增广样本及时调整现有模型。但是目前在线学习的研究工作主要集中在单视图学习上 [9–12]，而无法直接扩展到多视图学习。

针对多视图样本增广学习问题，本工作提出了一种单趟多视图学习 OPMV (One Pass Multi-View learning) 方法，只需扫描数据一次，即可实时更新模型，无需额外存储，对于新增样本也无需重新训练。具体地，这个问题可被形式化

为多视图一致性约束下组合目标函数联合优化问题，其中，一致性约束表现为一个线性等式，要求同一对象不同视图上的预测一致。OPMV 可以看成是在线交替方向乘子法 (online ADMM) 方法的推广。

ADMM，首次由 Gabay 和 Mercier 于 1976 年提出[104]。在实际应用中，ADMM 有很多很好的性质：易于实现[80, 105]、便于并行化[80, 106] 以及优秀的性能。Wang 和 Banerjee 提出了第一个在线 ADMM 方法[107]，随后很多在线 ADMM 的变体版本被提出[108–110] 。但是所有的这些在线 ADMM 的工作研究的都是系数固定的线性等式约束，而在 OPMV 中的线性约束是随着扫描不同样本而不断改变的。针对 OPMV，本工作分析了它的后悔界，并在 27 个数据集上验证了 OPMV 的有效性和高效性。

## 5.2 本文方法

本节提出了 OPMV 方法，将多视图样本增广学习问题形式化为多视图一致性约束下组合目标函数联合优化问题，并从理论上分析了 OPMV 的收敛率。

### 5.2.1 OPMV 方法形式化

在多视图学习中，每个样本 $\boldsymbol{x}$ 由几个不相交的特征空间描述。为了简化问题且不失一般性，在此主要讨论 2 个视图的设置。令 $\mathcal{X} = \mathcal{X}^1 \times \mathcal{X}^2$ 表示样本空间，其中 $\mathcal{X}^1$ 和 $\mathcal{X}^2$ 表示两个视图空间；$\mathcal{Y} = \{+1, -1\}$ 表示标记空间。观察到的训练集合为 $\boldsymbol{S}_n = \{(\boldsymbol{x}_1^1, \boldsymbol{x}_1^2; y_1), (\boldsymbol{x}_2^1, \boldsymbol{x}_2^2; y_2), \ldots, (\boldsymbol{x}_n^1, \boldsymbol{x}_n^2; y_n)\}$，其中每个样本都是从 $\mathcal{X} \times \mathcal{Y}$ 上独立同分布 (i.i.d.) 采样得到的。

令 $\mathcal{H}^1$ 和 $\mathcal{H}^2$ 分别表示每个视图上的函数空间。为了简化符号，定义 $[n] = \{1, 2, \ldots, n\}$，其中 $n$ 为正整数，并假设所有向量都在有限内积空间中 $\langle \cdot, \cdot \rangle$。对于两个大小相同的向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$，令 $\boldsymbol{u} \otimes \boldsymbol{v}$ 表示它们的外积矩阵，标记 $^\top$ 表示转置操作。

给定训练集 $\boldsymbol{S}_n$，多视图学习的目的是学到两个函数 $h^1 \in \mathcal{H}^1$ 和 $h^2 \in \mathcal{H}^2$ 在视图一致性约束下最小化经验 0/1 损失：

$$\begin{aligned} \min_{h^1 \in \mathcal{H}^1, h^2 \in \mathcal{H}^2} \quad & \sum_{i=1}^{n} \mathbb{I}(h^1(\boldsymbol{x}_i^1) \neq y_i) + \mathbb{I}(h^2(\boldsymbol{x}_i^2) \neq y_i) \\ \text{s.t.} \quad & h^1(\boldsymbol{x}_i^1) = h^2(\boldsymbol{x}_i^2) \quad \text{for} \quad i \in [n]. \end{aligned}$$

其中，$\mathbb{I}(\cdot)$ 表示指示函数，当输入为真返回 1，否则返回 0。

上式是对多视图学习的一般形式化，其中指示函数 (即 0/1 损失) 是一个非凸不连续函数，因此直接对它进行优化是一个 NP 难问题。实际上，考虑一些代替损失 $\ell$(例如铰链损失、指数损失等)，可使得优化能够高效求解。为了表述简单，研究线性函数空间，即

$$\mathcal{H}^1 = \{\boldsymbol{w}^1 \colon \|\boldsymbol{w}^1\| \le \mathcal{B}\} \quad 和 \quad \mathcal{H}^2 = \{\boldsymbol{w}^2 \colon \|\boldsymbol{w}^2\| \le \mathcal{B}\},$$

它能够扩展到非线性分类器。多视图学习的优化问题可以写作式 (5.1)：

$$\begin{aligned} \min_{\boldsymbol{w}^1 \in \mathcal{H}^1, \boldsymbol{w}^2 \in \mathcal{H}^2} \quad & \sum_{i=1}^{n} \ell(\langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle, y_i) + \ell(\langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle, y_i) \\ \text{s.t.} \quad & \operatorname{sign}(\langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle) = \operatorname{sign}(\langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle), \quad i \in [n], \end{aligned} \tag{5.1}$$

其中 $\operatorname{sign}(\langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle)$ 和 $\operatorname{sign}(\langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle)$ 分别表示不同视图上的预测标记。

在一致性约束 $\operatorname{sign}(\langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle) = \operatorname{sign}(\langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle),\ i \in [n]$ 中存在 sign(.) 函数，使得直接优化式 (5.1) 非常困难。出于计算上的考虑，将一致性约束简化为 $\langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle = \langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle,\ i \in [n]$。这样可得优化式 (5.2)：

$$\min_{\boldsymbol{w}^1 \in \mathcal{H}^1, \boldsymbol{w}^2 \in \mathcal{H}^2} \sum_{i=1}^{n} \ell(\langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle, y_i) + \ell(\langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle, y_i) : \langle \boldsymbol{w}^1, \boldsymbol{x}_i^1 \rangle = \langle \boldsymbol{w}^2, \boldsymbol{x}_i^2 \rangle,\ i \in [n]. \tag{5.2}$$

为了能够高效处理大规模多视图学习任务，处理新增样本时能够及时更新模型，设计了一种单趟学习算法优化式 (5.2)。令 $(\boldsymbol{x}_t^1, \boldsymbol{x}_t^2; y_t) \in S_n$ 表示第 $t \in [T]$ 时刻出现的标记样本，其中 $T$ 表示迭代轮数。对于第 $t$ 轮的优化任务如式 (5.3) 所示。

$$\min_{\boldsymbol{w}^1 \in \mathcal{H}^1, \boldsymbol{w}^2 \in \mathcal{H}^2} \phi_t(\boldsymbol{w}^1) + \psi_t(\boldsymbol{w}^2) \ : \ \langle \boldsymbol{w}^1, \boldsymbol{x}_t^1 \rangle = \langle \boldsymbol{w}^2, \boldsymbol{x}_t^2 \rangle \tag{5.3}$$

其中 $\phi_t(\boldsymbol{w}^1) = \ell(\langle \boldsymbol{w}^1, \boldsymbol{x}_t^1 \rangle, y_t) + \lambda \mathcal{R}(\boldsymbol{w}^1)$，$\psi_t(\boldsymbol{w}^2) = \ell(\langle \boldsymbol{w}^2, \boldsymbol{x}_t^2 \rangle, y_t) + \lambda \mathcal{R}(\boldsymbol{w}^2)$，$\lambda > 0$，$\mathcal{R}$ 是正则化项。

### 5.2.2 优化求解

注意到式 (5.3) 是一个线性等式约束下的组合目标函数优化，与在线 ADMM 的形式类似但是不同：在线 ADMM 中，线性等式约束中的系数在整个优化过程中是固定不变的；但在式 (5.3) 中，约束中的系数随着不同时间出现的不同样本

而不断变化。因此式 (5.3) 可以看作是在线 ADMM 框架的推广。

具体而言，本工作考虑了 $L_2$ 范数正则化项，即 $\mathcal{R}(\boldsymbol{w}^1) = \|\boldsymbol{w}^1\|_2^2$、$\mathcal{R}(\boldsymbol{w}^2) = \|\boldsymbol{w}^2\|_2^2$。在实现中，选择铰链损失作为替代损失 $\ell$ (这里可以替代为任何其它梯度或者次梯度存在的损失函数)。式 (5.3) 的增广拉格朗日函数为：

$$L_t(\boldsymbol{w}^1, \boldsymbol{w}^2, u_t) = \phi_t(\boldsymbol{w}^1) + \psi_t(\boldsymbol{w}^2) + u_t(\langle \boldsymbol{w}^1, \boldsymbol{x}_t^1 \rangle - \langle \boldsymbol{w}^2, \boldsymbol{x}_t^2 \rangle) + \frac{\rho}{2}\Big(\langle \boldsymbol{w}^1, \boldsymbol{x}_t^1 \rangle - \langle \boldsymbol{w}^2, \boldsymbol{x}_t^2 \rangle\Big)^2$$

其中 $\boldsymbol{w}^1$ 和 $\boldsymbol{w}^2$ 是原始变量，$u_t$ 是对偶变量，$\rho$ ($\rho > 0$) 是惩罚参数。引入 $\alpha_t = u_t/\rho$ 可得：

$$L_t(\boldsymbol{w}^1, \boldsymbol{w}^2, \alpha_t) = \phi_t(\boldsymbol{w}^1) + \psi_t(\boldsymbol{w}^2) - \alpha_t^2 + \frac{\rho}{2}(\langle \boldsymbol{w}^1, \boldsymbol{x}_t^1 \rangle - \langle \boldsymbol{w}^2, \boldsymbol{x}_t^2 \rangle + \alpha_t)^2.$$

为了简化计算，采用了线性化策略：

$$\phi_t(\boldsymbol{w}^1) = \phi_t(\boldsymbol{w}_t^1) + \langle \nabla\phi_t(\boldsymbol{w}_t^1), \boldsymbol{w}^1 - \boldsymbol{w}_t^1 \rangle,$$

其中 $\nabla\phi_t(\boldsymbol{w}_t^1)$ 表示 $\phi_t(\boldsymbol{w}^1)$ 在 $\boldsymbol{w}^1 = \boldsymbol{w}_t^1$ 的梯度。通过式 (5.4) 更新 $\boldsymbol{w}_{t+1}^1$ ：

$$\boldsymbol{w}_{t+1}^1 \leftarrow \arg\min_{\boldsymbol{w}^1} \langle \nabla\phi_t(\boldsymbol{w}_t^1), \boldsymbol{w}^1 \rangle + \frac{1}{\eta} B_\Psi(\boldsymbol{w}^1, \boldsymbol{w}_t^1) + \frac{\rho}{2}\Big(\langle \boldsymbol{w}^1, \boldsymbol{x}_t^1 \rangle - \langle \boldsymbol{w}_t^2, \boldsymbol{x}_t^2 \rangle + \alpha_t\Big)^2, \quad (5.4)$$

其中，Bregman 散度 $B_\Psi$ 用于控制 $\boldsymbol{w}_t^1$ 和 $\boldsymbol{w}_{t+1}^1$ 之间的距离，$\eta$ 是学习率。此工作主要考虑欧氏距离，即 $B_\Psi(\boldsymbol{w}^1, \boldsymbol{w}_t^1) = \frac{1}{2}\|\boldsymbol{w}^1 - \boldsymbol{w}_t^1\|_2^2$。

通过最小化式 (5.4)，$\boldsymbol{w}_{t+1}^1$ 的更新如式 (5.5) 所示：

$$\boldsymbol{w}_{t+1}^1 \leftarrow \left(\frac{1}{\eta}\mathbf{I} + \rho \mathbf{x}_\mathbf{t}^\mathbf{1} \otimes \mathbf{x}_\mathbf{t}^\mathbf{1}\right)^{-\mathbf{1}} \times \left(\frac{\mathbf{1}}{\eta}\mathbf{w}_\mathbf{t}^\mathbf{1} - \nabla\phi_\mathbf{t}(\mathbf{w}_\mathbf{t}^\mathbf{1}) + \rho(\langle \mathbf{w}_\mathbf{t}^\mathbf{2}, \mathbf{x}_\mathbf{t}^\mathbf{2} \rangle - \alpha_\mathbf{t})\mathbf{x}_\mathbf{t}^\mathbf{1}\right), \quad (5.5)$$

其中 $\mathbf{I}$ 是 $d^1 \times d^1$ 的单位矩阵，$d^1$ 是 $\mathcal{X}^1$ 的维度。注意到这里的更新包括了一个 $d^1 \times d^1$ 的矩阵求逆操作，当维度很大的时候，将导致很大的计算和存储开销。为了解决这个问题，引入了 Sherman–Morrison 公式[111]。这样一来 $\boldsymbol{w}_{t+1}^1$ 的更新如式 (5.6) 所示：

$$\boldsymbol{w}_{t+1}^1 \leftarrow \eta \boldsymbol{v}_t^1 - \beta_t^1 \boldsymbol{w}_t^1, \quad (5.6)$$

其中，

$$\boldsymbol{v}_t^1 \;=\; -\nabla\phi_t(\boldsymbol{w}_t^1) + \frac{1}{\eta}\boldsymbol{w}_t^1 + \rho(\langle \boldsymbol{w}_t^2, \boldsymbol{x}_t^2 \rangle - \alpha_t)\boldsymbol{x}_t^1 \quad (5.7)$$

**算法 5.1** OPMV

**输入**: $\lambda > 0$、$\rho > 0$、$\eta > 0$、观察数据。
**输出**: $\boldsymbol{w}^1$ 和 $\boldsymbol{w}^2$。

初始化：$\boldsymbol{w}_0^1 = \mathbf{0}$，$\boldsymbol{w}_0^2 = \mathbf{0}$，$\alpha_0 = 0$;
**for** $t = 0, 1, \cdots, T-1$ **do**
  接受一个标记样本 $(\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, y_t)$;
  通过式 (5.6)-(5.8) 更新 $\boldsymbol{w}_{t+1}^1$;
  通过式 (5.9)-(5.11) 更新 $\boldsymbol{w}_{t+1}^2$;
  通过式 (5.12) 更新 $\alpha_{t+1}$;
**end for**
输出 $\boldsymbol{w}^1 = \boldsymbol{w}_T^1$，$\boldsymbol{w}^2 = \boldsymbol{w}_T^2$。

$$\beta_t^1 = \frac{\rho\eta^2\langle \boldsymbol{x}_t^1, \boldsymbol{v}_t^1\rangle}{1+\rho\eta\langle \boldsymbol{x}_t^1, \boldsymbol{x}_t^1\rangle}. \tag{5.8}$$

$\boldsymbol{v}_t^{1(2)}$ 和 $\beta_t^{1(2)}$ 是两个中间结果，前者是一个向量，后者为一个常数，因此无需直接计算和存储 $(\mathbf{I}/\eta + \rho\mathbf{x_t^1} \otimes \mathbf{x_t^1})^{-1}$，从而提高了算法效率，且能够应用于高维数据。同样地，$\boldsymbol{w}_{t+1}^2$ 的更新如式 (5.9) 所示：

$$\boldsymbol{w}_{t+1}^2 \leftarrow \eta\boldsymbol{v}_t^2 - \beta_t^2\boldsymbol{w}_t^2, \tag{5.9}$$

其中，

$$\boldsymbol{v}_t^2 = -\nabla\psi_t(\boldsymbol{w}_t^2) + \frac{1}{\eta}\boldsymbol{w}_t^2 + \rho(\langle \boldsymbol{w}_{t+1}^1, \boldsymbol{x}_t^1\rangle + \alpha_t)\boldsymbol{x}_t^2 \tag{5.10}$$

$$\beta_t^2 = \frac{\rho\eta^2\langle \boldsymbol{x}_t^2, \boldsymbol{v}_t^2\rangle}{1+\rho\eta\langle \boldsymbol{x}_t^2, \boldsymbol{x}_t^2\rangle}. \tag{5.11}$$

最后令 $L_t(\boldsymbol{w}^1, \boldsymbol{w}^2, \alpha_t)$ 关于 $\alpha_t$ 的梯度为 0，可得 $\alpha_{t+1}$ 的更新如式 (5.12) 所示：

$$\alpha_{t+1} \leftarrow \alpha_t + \langle \boldsymbol{w}_{t+1}^1, \boldsymbol{x}_t^1\rangle - \langle \boldsymbol{w}_{t+1}^2, \boldsymbol{x}_t^2\rangle. \tag{5.12}$$

算法5.1总结了 OPMV 算法。在测试阶段，新样本 $(\boldsymbol{x}_{new}^1, \boldsymbol{x}_{new}^2)$ 的预测标记为 $y = \text{sign}(\langle \boldsymbol{w}_T^1, \boldsymbol{x}_{new}^1\rangle + \langle \boldsymbol{w}_T^2, \boldsymbol{x}_{new}^2\rangle)$.

### 5.2.3 理论分析

本节分析了 OPMV 的收敛速率，首先引入引理 5.1：

**引理 5.1** $f(\boldsymbol{w})$ 为一个凸函数，对于任意数值 $r>0$ 和任意向量 $\boldsymbol{u}$，令

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w}) + r\|\boldsymbol{w}-\boldsymbol{u}\|_2^2.$$

对 $f(\boldsymbol{w}^*)$ 的梯度或任意次梯度 $\boldsymbol{g}\in\partial f(\boldsymbol{w}^*)$ 有

$$\langle \boldsymbol{g}, \boldsymbol{w}^*-\boldsymbol{w}\rangle \le r(\|\boldsymbol{w}-\boldsymbol{u}\|_2^2 - \|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2 - \|\boldsymbol{u}-\boldsymbol{w}^*\|_2^2)$$

成立。

**证明**: 对于凸函数 $f(\boldsymbol{w})$，最优解 $\boldsymbol{w}^*$ 满足

$$\langle \boldsymbol{g}+2r(\boldsymbol{w}^*-\boldsymbol{u}), \boldsymbol{w}-\boldsymbol{w}^*\rangle \ge 0.$$

结合

$$2\langle \boldsymbol{w}^*-\boldsymbol{u}, \boldsymbol{w}^*-\boldsymbol{w}\rangle = \|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2 + \|\boldsymbol{u}-\boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w}-\boldsymbol{u}\|_2^2,$$

即得证。 □

假设分类器和次梯度或者梯度在每一轮迭代 $t\in[T]$ 有界，即

**假设 5.2** $\|\boldsymbol{x}_t^1\|\le\mathcal{B}_0$ ， $\|\boldsymbol{x}_t^2\|\le\mathcal{B}_0$；

**假设 5.3** $\|\boldsymbol{w}_t^i\|\le\mathcal{B}_1$ ， $\|\boldsymbol{w}_*^i\|\le\mathcal{B}_1, i\in\{1,2\}$；

**假设 5.4** $\|\nabla\phi_t(\boldsymbol{w}_t^1)\|\le\mathcal{B}_2$ ， $\|\nabla\psi_t(\boldsymbol{w}_t^2)\|\le\mathcal{B}_2$；

给出关于后悔界的理论结果如下：

**定理 5.5** 令 $\{\boldsymbol{w}_t^1, \boldsymbol{w}_t^2, \alpha_t\}$ 表示由算法5.1生成的序列，设定 $\rho = T^{-3/2}$ 、$\eta = T^{-1/2}$，在假设5.2- 5.4的条件下，有

$$\sum_{t=1}^{T}\Big(\phi_t(\boldsymbol{w}_t^1)+\psi_t(\boldsymbol{w}_t^2)\Big) - \min_{(\boldsymbol{w}_*^1,\boldsymbol{w}_*^2)\in C}\sum_{t=1}^{T}\Big(\phi_t(\boldsymbol{w}_*^1)+\psi_t(\boldsymbol{w}_*^2)\Big) \le (\mathcal{B}_1+\mathcal{B}_2+4\mathcal{B}_0^2\mathcal{B}_1^2)T^{1/2}+4\mathcal{B}_0\mathcal{B}_1^2/T,$$

成立，其中 $C=\{(\boldsymbol{w}_*^1,\boldsymbol{w}_*^2)\colon \mathrm{sign}(\langle\boldsymbol{w}_*^1,\boldsymbol{x}_t^1\rangle)=\mathrm{sign}(\langle\boldsymbol{w}_*^2,\boldsymbol{x}_t^2\rangle),\ t\in[T]\}$.

**证明**: 由于 $\boldsymbol{w}_{t+1}^1$ 是式 (5.4) 的最优解，由引理 5.1，有下式成立：

$$\left\langle \nabla\phi_t(\boldsymbol{w}_t^1) + \rho(\langle \boldsymbol{w}_{t+1}^1, \boldsymbol{x}_t^1\rangle - \langle \boldsymbol{w}_t^2, \boldsymbol{x}_t^2\rangle + \alpha_t)\boldsymbol{x}_t^1, \boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_*^1 \right\rangle \\ \leq \left(\|\boldsymbol{w}_*^1 - \boldsymbol{w}_t^1\|_2^2 - \|\boldsymbol{w}_*^1 - \boldsymbol{w}_{t+1}^1\|_2^2 - \|\boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_t^1\|_2^2\right)/2\eta.$$

对于凸函数 $\phi_t$，有 $\phi_t(\boldsymbol{w}_t^1) - \phi_t(\boldsymbol{w}_*^1) \leq \langle \nabla\phi_t(\boldsymbol{w}_t^1), \boldsymbol{w}_t^1 - \boldsymbol{w}_*^1\rangle$ 成立，联合上述两个不等式可得：

$$\rho(\langle \boldsymbol{w}_{t+1}^1, \boldsymbol{x}_t^1\rangle - \langle \boldsymbol{w}_t^2, \boldsymbol{x}_t^2\rangle + \alpha_t)\langle \boldsymbol{x}_t^1, \boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_*^1\rangle + \phi_t(\boldsymbol{w}_t^1) - \phi_t(\boldsymbol{w}_*^1) \\ \leq \langle \nabla\phi_t(\boldsymbol{w}_t^1), \boldsymbol{w}_t^1 - \boldsymbol{w}_{t+1}^1\rangle + \left(\|\boldsymbol{w}_*^1 - \boldsymbol{w}_t^1\|_2^2 - \|\boldsymbol{w}_*^1 - \boldsymbol{w}_{t+1}^1\|_2^2 - \|\boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_t^1\|_2^2\right)/2\eta.$$

根据 Young 不等式，可得：

$$\langle \nabla\phi_t(\boldsymbol{w}_t^1), \boldsymbol{w}_t^1 - \boldsymbol{w}_{t+1}^1\rangle \leq \eta\|\nabla\phi_t(\boldsymbol{w}_t^1)\|_2^2/2 + \|\boldsymbol{w}_t^1 - \boldsymbol{w}_{t+1}^1\|_2^2/2\eta.$$

关于视图 1 可以推得：

$$\rho(\langle \boldsymbol{w}_{t+1}^1, \boldsymbol{x}_t^1\rangle - \langle \boldsymbol{w}_t^2, \boldsymbol{x}_t^2\rangle + \alpha_t)\langle \boldsymbol{x}_t^1, \boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_*^1\rangle + \phi_t(\boldsymbol{w}_t^1) - \phi_t(\boldsymbol{w}_*^1) \\ \leq \eta\|\nabla\phi_t(\boldsymbol{w}_t^1)\|_2^2/2 + \left(\|\boldsymbol{w}_*^1 - \boldsymbol{w}_t^1\|_2^2 - \|\boldsymbol{w}_*^1 - \boldsymbol{w}_{t+1}^1\|_2^2\right)/2\eta. \quad (5.13)$$

同样地，关于视图 2 可以推得：

$$\rho(\langle \boldsymbol{w}_{t+1}^2, \boldsymbol{x}_t^2\rangle - \langle \boldsymbol{w}_{t+1}^1, \boldsymbol{x}_t^1\rangle - \alpha_t)\langle \boldsymbol{x}_t^2, \boldsymbol{w}_{t+1}^2 - \boldsymbol{w}_*^2\rangle + \psi_t(\boldsymbol{w}_t^2) - \psi_t(\boldsymbol{w}_*^2) \\ \leq \eta\|\nabla\psi_t(\boldsymbol{w}_t^2)\|_2^2/2 + \left(\|\boldsymbol{w}_*^2 - \boldsymbol{w}_t^2\|_2^2 - \|\boldsymbol{w}_*^2 - \boldsymbol{w}_{t+1}^2\|_2^2\right)/2\eta. \quad (5.14)$$

联合式 (5.13) 和 (5.14)，从 $t = 0$ 一直加到 $t = T - 1$，即可得：

$$\sum_{t=0}^{T-1} \phi_t(\boldsymbol{w}_t^1) + \psi_t(\boldsymbol{w}_t^2) - \phi_t(\boldsymbol{w}_*^1) - \psi_t(\boldsymbol{w}_*^2) \\ \leq \frac{1}{2\eta}\left(\|\boldsymbol{w}_T^2\| + \|\boldsymbol{w}_T^1\|\right) + \frac{\eta}{2}\sum_{t=0}^{T-1}\left(\|\nabla\phi_t(\boldsymbol{w}_t^1)\|_2^2 + \|\nabla\psi_t(\boldsymbol{w}_t^2)\|_2^2\right) \\ +\rho\sum_{t=0}^{T-1}\langle \boldsymbol{w}_t^2 - \boldsymbol{w}_{t+1}^2\rangle\langle \boldsymbol{x}_t^1, \boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_*^1\rangle + \rho\sum_{t=0}^{T-1}\alpha_{t+1}(\langle \boldsymbol{x}_t^2, \boldsymbol{w}_{t+1}^2 - \boldsymbol{w}_*^2\rangle - \langle \boldsymbol{x}_t^1, \boldsymbol{w}_{t+1}^1 - \boldsymbol{w}_*^1\rangle).$$

根据式 (5.12)，有

$$\alpha_t = \sum_{i=1}^{t-1}\langle \boldsymbol{w}_{i+1}^1, \boldsymbol{x}_i^1\rangle - \langle \boldsymbol{w}_{i+1}^2, \boldsymbol{x}_i^2\rangle,$$

**表 5.1** 数据集信息，$d^1$ 和 $d^2$ 分别代表两个视图的维度。

| 数据集 | 样本数 | $d^1$ | $d^2$ | 数据集 | 样本数 | $d^1$ | $d^2$ | 数据集 | 样本数 | $d^1$ | $d^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rt.EN-FR | 18,758 | 21,531 | 24,892 | Rt.GR-FR | 29,953 | 34,279 | 24,892 | Rt.SP-GR | 12,342 | 11,547 | 34,262 |
| Rt.EN-GR | 18,758 | 21,531 | 34,215 | Rt.GR-IT | 29,953 | 34,279 | 15,505 | Rt.SP-IT | 12,342 | 11,547 | 15,500 |
| Rt.EN-IT | 18,758 | 21,531 | 15,506 | Rt.GR-SP | 29,953 | 34,279 | 11,547 | Cora | 2,708 | 2,708 | 1,433 |
| Rt.EN-SP | 18,758 | 21,531 | 11,547 | Rt.IT-EN | 24,039 | 15,506 | 21,517 | IMDB | 617 | 1,878 | 1,398 |
| Rt.FR-EN | 26,648 | 24,893 | 21,531 | Rt.IT-FR | 24,039 | 15,506 | 24,892 | NG.M2 | 500 | 2,000 | 2,000 |
| Rt.FR-GR | 26,648 | 24,893 | 34,287 | Rt.IT-GR | 24,039 | 15,506 | 34,278 | NG.M5 | 500 | 2,000 | 2,000 |
| Rt.FR-IT | 26,648 | 24,893 | 15,503 | Rt.IT-SP | 24,039 | 15,506 | 11,547 | NG.M10 | 500 | 2,000 | 2,000 |
| Rt.FR-SP | 26,648 | 24,893 | 11,547 | Rt.SP-EN | 12,342 | 11,547 | 21,530 | NG.NG1 | 400 | 2,000 | 2,000 |
| Rt.GR-EN | 29,953 | 34,279 | 21,531 | Rt.SP-FR | 12,342 | 11,547 | 24,892 | NG.NG2 | 1,000 | 2,000 | 2,000 |

可得 $|\alpha_t| \leq 2(t-1)\mathcal{B}_0\mathcal{B}_1$。因此，有下式成立：

$$\sum_{t=0}^{T-1} \phi_t(\boldsymbol{w}_t^1) + \psi_t(\boldsymbol{w}_t^2) - \phi_t(\boldsymbol{w}_*^1) - \psi_t(\boldsymbol{w}_*^2) \leq B_1/\eta + \eta T\mathcal{B}_2 + \rho(4T\mathcal{B}_0\mathcal{B}_1^2 + 4T^2\mathcal{B}_0^2\mathcal{B}_1^2). \tag{5.15}$$

令 $\rho = T^{-3/2}$ and $\eta = T^{-1/2}$ 带入式 (5.15) 即得证。 □

## 5.3 实验测试

本节在 27 个数据集上验证了 OPMV 处理多视图样本增广学习的有效性和高效性。

### 5.3.1 实验设置

在 27 个常用多视图真实数据集上进行实验，包括 Cora [112] 、IMDB [113]、News Group [114] 、Reuter [115] 。对于多分类数据集 Cora、IMDB 和 Reuter，通过随机将多个类别分为两类将其转换为二分类任务。数据集的细节信息如表5.1所示。

在实验中，将 OPMV 与 1 个单视图学习方法、4 个多视图方法 (非在线学习) 以及 OPMV 的非在线变体比较。所有的对比方法都需要将整个训练数据集载入内存，并且在训练过程中需要反复扫描整个数据集。而 OPMV 则只需要在每个样本出现的时候更新模型，无需存储整个训练集。对比方法的更多细节如下所示：

(1) SV：将两个视图拼成一个单视图，然后应用 SVM 进行分类；

(2) CCAMV：首先应用 CCA 提取两个视图公共子空间上的表达，然后在得到的

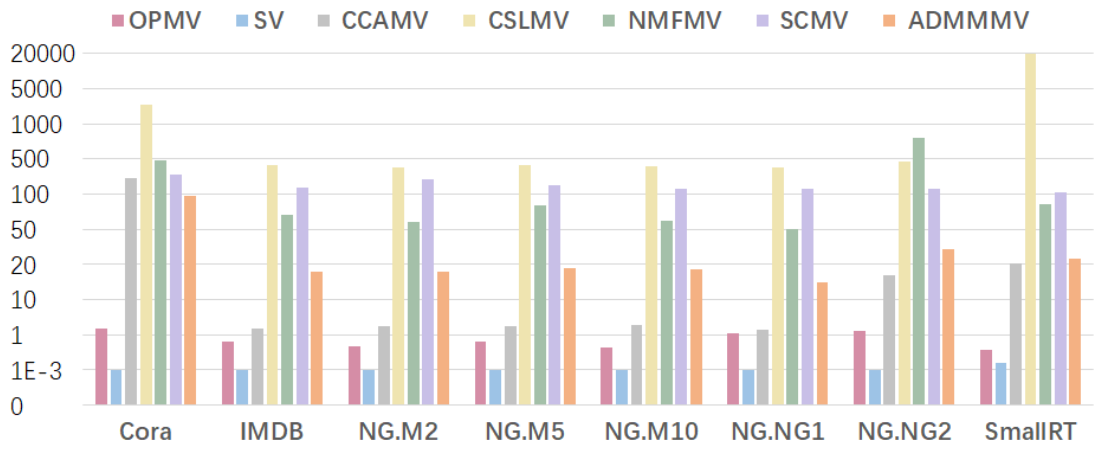



**图 5.1** 算法时间比较 (单位：s)。

公共子空间上进行分类；

(3) CSLMV：首先应用 CSL[32] 学到两个视图之间的公共子空间，然后在上面应用 SVM 分类；

(4) NMFMV：首先应用 NMF[33] 学习隐藏表示，然后应用 SVM 分类；

(5) SCMV：用 SCMV[31] 同时学习子空间投影和分类器；

(6) ADMMMV：OPMV 的非在线学习版本，通过 ADMM 优化式 (5.2)。

### 5.3.2 实验结果

比较结果如表 5.2所示；算法平均耗时如图 5.1所示。由于非在线多视图学习方法在大规模数据集（如 Reuter）均不能在 8 小时之内返回一次运行的结果，因此在随机选择的这些数据集子集上进行性能比较。其中，每个数据子集包括 3,000 个样本和 400 个特征 (对应出现频率最高的 400 个单词)。对于这些采样后的数据集 (简称为 SmallRt)，对比结果见表 5.3，算法平均耗时如图 5.1中最后一列所示。实验结果验证了 OPMV 的有效性和高效性。

首先，OPMV 优于 SV 方法。SV 方法将两个视图合并成一个视图，没有利用到视图之间的结构。OPMV 在所有数据上的性能都优于 SV，验证了利用不同视图之间的结构可以有效提升学习性能 (在 OPMV，体现为多视图一致性约束)。此外 OPMV 在线学习的性能与它的非在线版本 ADMMMV 是可比的，但在时间效率上有显著提高。

其次，OPMV 在较小数据集上的性能也超过了 CCAMV、CSLMV、NMFMV 以及 SCMV 等非在线多视图学习方法。在对 Reuter 进行采样后的小数据集上，与

**表 5.2** 预测任务准确率比较。'N/A' 表示 8 小时之内不能返回单次运行结果。斜体表示 OPMV 显著优于对比方法（t 检验，$\alpha = 0.05$）。

| 数据集 | OPMV | SV | CCAMV | CSLMV | NMFMV | SCMV | ADMMMV |
|---|---|---|---|---|---|---|---|
| Cora | .902±.013 | *.882±.009* | *.880±.020* | *.890±.013* | .901±.026 | *.860±.003* | .903±.007 |
| IMDB | .602±.003 | *.593±.003* | .598±.003 | *.587±.010* | .607±.005 | *.586±.004* | .618±.003 |
| NG.M2 | .940±.026 | .935±.017 | *.880±.061* | .946±.014 | .941±.022 | *.890±.040* | .945±.017 |
| NG.M5 | .933±.030 | .936±.014 | .940±.046 | .940±.035 | .911±.044 | .924±.049 | .942±.024 |
| NG.M10 | .877±.038 | *.849±.039* | *.862±.042* | *.866±.032* | *.861±.043* | *.856±.037* | .871±.028 |
| NG.NG1 | .951±.030 | *.943±.028* | .949±.031 | .952±.030 | *.932±.026* | *.920±.044* | .960±.020 |
| NG.NG2 | .921±.020 | .915±.019 | .919±.035 | .920±.020 | *.920±.019* | *.910±.024* | .935±.018 |
| Rt.EN-FR | .936±.003 | *.926±.007* | N/A | N/A | N/A | N/A | N/A |
| Rt.EN-GR | .933±.004 | *.923±.005* | N/A | N/A | N/A | N/A | N/A |
| Rt.EN-IT | .933±.004 | *.924±.006* | N/A | N/A | N/A | N/A | N/A |
| Rt.EN-SP | .932±.004 | *.924±.004* | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-EN | .905±.004 | *.891±.003* | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-GR | .904±.005 | *.894±.005* | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-IT | .904±.004 | *.891±.003* | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-SP | .903±.004 | *.888±.003* | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-EN | .926±.004 | *.899±.002* | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-FR | .927±.004 | *.899±.005* | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-IT | .923±.004 | *.903±.004* | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-SP | .925±.003 | *.902±.002* | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-EN | .897±.003 | *.877±.006* | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-FR | .898±.003 | *.877±.005* | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-GR | .895±.004 | *.878±.005* | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-SP | .895±.003 | *.874±.005* | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-EN | .953±.004 | *.922±.007* | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-FR | .953±.004 | *.921±.007* | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-GR | .953±.005 | *.925±.010* | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-IT | .952±.003 | *.919±.079* | N/A | N/A | N/A | N/A | N/A |

对比方法相比，OPMV 取得了更好的性能。其可能原因如下：1) 多数对比方法主要通过无监督学习得到公共子空间，由于没有监督信息指导，无法保证学到的子空间适合相应分类任务；2) 对比方法均为非凸优化，可能会收敛到局部最优解；3) 对比方法对参数选择较为敏感。

在运行时间的比较中，由图5.1所示，OPMV 远远快于现有的非在线多视图学习方法。例如在 Cora 数据集上，有 2,708 个样本，2 个视图的维度分别为 2,708 和 1,433，OPMV 比 ADMMMV 快了 50 倍；比 CCAMV 快了 110 倍；比 SCMV 快了 130 倍；比 NMFMV 快了 180 倍；甚至比 CSLMV 快了 1000 倍。在

**表 5.3** 采样数据集 SmallRt 上的预测准确率比较。'N/A' 表示 8 小时之内不能返回单次运行结果。斜体表示 OPMV 显著优于对比方法（t 检验，$\alpha = 0.05$）。（t 检验，$\alpha = 0.05$）。

| 数据集 | OPMV | SV | CCAMV | CSLMV | NMFMV | SCMV | ADMMMV |
|---|---|---|---|---|---|---|---|
| Rt.EN-FR | .876±.005 | *.849±.005* | *.865±.008* | N/A | *.854±.010* | *.841±.013* | .882±.005 |
| Rt.EN-GR | .881±.005 | *.852±.008* | *.849±.017* | N/A | *.852±.008* | *.868±.017* | .889±.006 |
| Rt.EN-IT | .872±.011 | *.852±.003* | *.865±.021* | N/A | *.860±.005* | *.867±.009* | .885±.007 |
| Rt.EN-SP | .881±.002 | *.852±.004* | .874±.035 | N/A | .868±.007 | *.861±.001* | .884±.003 |
| Rt.FR-EN | .842±.009 | *.791±.005* | *.800±.014* | N/A | *.810±.010* | *.796±.003* | .840±.003 |
| Rt.FR-GR | .830±.008 | *.790±.005* | *.795±.017* | N/A | *.792±.011* | *.797±.008* | .840±.004 |
| Rt.FR-IT | .836±.004 | *.789±.017* | *.795±.011* | N/A | *.828±.003* | *.794±.015* | .844±.008 |
| Rt.FR-SP | .833±.009 | *.789±.002* | *.801±.019* | N/A | .827±.002 | *.807±.017* | .845±.005 |
| Rt.GR-EN | .882±.001 | *.820±.010* | *.820±.010* | N/A | *.865±.004* | *.863±.013* | .883±.002 |
| Rt.GR-FR | .878±.004 | *.819±.005* | *.800±.022* | N/A | *.850±.015* | *.834±.012* | .885±.004 |
| Rt.GR-IT | .880±.004 | *.820±.011* | *.809±.017* | N/A | *.866±.004* | *.856±.005* | .887±.005 |
| Rt.GR-SP | .878±.001 | *.823±.003* | *.810±.017* | N/A | *.868±.013* | *.830±.016* | .889±.001 |
| Rt.IT-EN | .831±.004 | *.791±.009* | *.800±.015* | N/A | *.800±.009* | *.794±.010* | .839±.003 |
| Rt.IT-FR | .830±.006 | *.792±.004* | *.795±.020* | N/A | *.800±.009* | *.795±.006* | .827±.001 |
| Rt.IT-GR | .833±.003 | *.783±.006* | *.790±.020* | N/A | *.809±.008* | *.797±.007* | .838±.004 |
| Rt.IT-SP | .830±.003 | *.793±.005* | *.795±.017* | N/A | *.802±.009* | *.800±.005* | .834±.002 |
| Rt.SP-EN | .917±.003 | *.883±.011* | *.887±.009* | N/A | *.900±.010* | *.895±.007* | .917±.005 |
| Rt.SP-FR | .915±.004 | *.883±.002* | *.879±.014* | N/A | *.901±.016* | *.897±.011* | .917±.002 |
| Rt.SP-GR | .910±.001 | *.881±.012* | *.891±.015* | N/A | .909±.005 | *.883±.011* | .923±.002 |
| Rt.SP-IT | .917±.004 | *.880±.011* | *.906±.017* | N/A | *.899±.007* | *.906±.015* | .923±.005 |

Reuter 数据集上，OPMV 单趟运行只需要不到 4 分钟，而对比多视图方法均无法在 8 小时内返回结果。即使在采样后的 SmallRt 数据集上，OPMV 也比对比多视图方法高效。

## 5.4 本章小结

多视图学习是一种重要的利用增广特征的机器学习范式，但是以往的方法计算复杂度较高，需要反复扫描数据，因而多应用于较小的数据集，无法满足在大数据情况下对增广样本及时更新模型的需求。为了能够高效解决多视图样本增广学习问题，提出了 OPMV 方法，只需要扫描一遍数据，而无需将所有数据载入内存，即能根据新增多视图样本对模型进行高效地更新。该方法是基于多视图一致性约束下的组合目标函数优化。理论分析证明了 OPMV 的收敛速率为 $O(1/\sqrt{T})$，实验验证了 OPMV 的有效性和高效性。

本工作已总结成文：

- Y. Zhu and W. Gao and Z.-H. Zhou. One-pass multi-view learning. In **Proceedings of the 7th Asian Conference on Machine Learning (ACML'15)**, Hong Kong, 2015, JMLR: W&CP 45, pp.407-422.（机器学习领域重要国际会议）

# 第六章　综合增广学习

## 6.1　引言

第二到五章的工作大部分基于某一种增广信息，在实际应用中还有多种信息增广的形式同时存在的情况：例如在多标记图片分类任务中，预测图片中的新标记（标记增广）；可以获得额外的文本描述信息作为原始特征的补充（特征增广）；同时面对样本动态增加的情形，要求能够对模型即时更新（样本增广）。本工作主要研究了这种同时进行标记、特征、样本增广学习的综合增广学习问题。

直接在多标记学习的框架下解决标记增广学习，检测新标记是非常困难的。现有一类多标记弱标记学习工作 [116] 考虑标记部分缺失的情况：某些样本上缺失的标记在另一些样本上可以观测到。在这种情况下，可以利用低秩结构特性，通过优化矩阵核范数 [18, 37] 或者最小化重构误差进行低秩矩阵分解 [19, 117]，从而对标记矩阵进行补全。但是在多标记新标记学习中，由于新标记在整个训练集中均未被标注，即真实标记矩阵整列缺失，无法利用低秩矩阵补全。

如果直接采用异常检测的方法判断是否有新标记出现也可能失败 [118, 119]：(1) 一个样本可能不仅仅有新标记，还可能同时与多个已知标记相关联，因而很难从特征上将同时有新标记和已知标记的样本与那些没有新标记只有相同已知标记的样本区分开；(2) 即使一个样本被检测算法认为是一个异常样本，可能由于该样本真的与新标记相关联，也有可能是因为出现了训练集中没有出现过的已知标记的组合。

为了处理该综合增广学习中的标记增广问题，考虑应用表示能力更强的学习范式，主要包括结构和特征两个层面。从结构上，学习框架采用多示例多标记学习框架 [65, 68, 70, 71, 74, 120]。其中，每个对象由一个多示例包表示，包中的每个示例对应某一个语义概念；包的标记即为包中样本标记的并集（在训练过程中，只能观察到包的标记，而具体的包中示例的标记不可见）。以图像标注任务为例，一个多示例包对应的是一幅图像，而包中的示例则对应图像分割成的

各个小块。直觉上看，在多示例多标记学习框架下，只要能够预测各个示例的标记，如果其中某个示例不属于已知标记，则该示例以及它所在的包存在新标记。但是通过穷举方式搜索所有样本-标记组合的复杂度关于包中示例的个数和标记的个数均为指数级增长。因此，实际应用多示例多标记学习框架需要精心设计高效算法。大部分现有的多示例多标记学习算法不能处理新标记学习问题，且由于较高的计算复杂度，很难处理大规模数据。从利用表示能力更强的特征角度，希望能够利用深度学习的分层卷积网络结构，自动学习出好的特征表示。除此之外，深度学习小批量更新的方式可以随着样本增多即时更新模型，可以直接处理样本增广的情况。之前的深度学习的研究主要集中在监督学习、半监督学习和无监督学习 [121–123]，而对于这种特殊的整列标记缺失的弱监督学习问题 [116] 很少见相关工作。

在利用增广特征时，通过多视图学习框架 [23, 25, 30, 31, 124]，利用辅助信息进一步提升模型性能。以图片标注任务为例，直觉上，相较于原始图片，这些文本描述中有更加简明凝练的语义信息，甚至可能包含了隐藏语义所对应的描述。把图片作为原始视图，获取相应的文本描述作为增广视图，需要针对多标记新标记学习问题，设计合适的网络结构。

为了考虑这种同时进行标记、特征及样本增广学习的综合增广学习问题，提出了一种端到端的多视图多示例标记新标记学习方法 EM3NL (End-to-end Multi-view Multi-Instance Multi-label learning with New Labels), 将深度神经网络与多视图、多示例多标记学习结合，输入原始图片和文本，输出已知标记和新标记的预测。主要贡献如下:

(1) 将深度神经网络与多视图、多示例多标记学习结合，设计了一种直接输入图像-文本视图，输出已知标记和新标记预测的端到端统一解决方案；
(2) 提出了一种多示例包级的损失函数、以及对新标记预测的正则化项;
(3) 改变了 DeepMIML 工作 [125] 中深度模型的上层网络结构，使其适用于新标记检测。

## 6.2 本文方法

首先，对本文研究的特征、测试集标记、样本综合增广学习作如下形式化：令 $\boldsymbol{\upsilon} = \{1, 2, \cdots, l\}$ 表示已知的标记集合，$\mathcal{X}$ 和 $\mathcal{X}^*$ 分别表示原始视图和增广视图的样本空间。给定一个输入序列 $\{\boldsymbol{X}_i, \boldsymbol{X}_i^*, \boldsymbol{y}_i\}_{i=1}^T$，其中 $\boldsymbol{X}_i \in \mathcal{X}$ 表示原始

视图的一个样本，$\boldsymbol{X}_i^* \in \mathcal{X}^*$ 表示相应的增广视图，$\boldsymbol{y}_i \in \mathcal{Y} = \{0,1\}^l$ 表示观察到的标记向量，$y_{i,j} = 1$ 表示第 $i$ 个样本有标记 $v_j$，否则没有。需要特别指出，这里的原始视图和增广视图可以是但不限于原始的图片和文本输入。本工作虽以图片与文本视图为例，但可以通过改变特征学习的网络结构扩展到其它输入类型。令 $\boldsymbol{v}' = \{1,2,\cdots,l,l+1\}$ 表示测试集标记集合，其中，$v'_{l+1} \notin \boldsymbol{v}$ 表示存在新标记。令 $\mathcal{Y}' = \{0,1\}^{l+1}$ 表示测试集的标记空间。目标是学习一个映射 $f : \mathcal{X} \times \mathcal{X}^* \to \mathcal{Y}'$，预测测试样本的标记 (包括新标记)。

然后提出了 EN3NL 方法，将深度神经网络与多视图、多示例多标记学习结合，输入原始图片和文本，输出已知标记和新标记的预测。后文主要介绍 EN3NL 方法的细节，包括目标函数设计、多示例多标记学习网络结构以及特征提取网络结构设计，分别对应第6.2.1-6.2.3节。

### 6.2.1 目标函数设计

在 EM3NL 中，利用卷积及池化操作建立深度神经网络，分别为图像视图和文本视图学得多示例特征，记为 $\boldsymbol{B}_i = \phi(\boldsymbol{X}_i)$ 和 $\boldsymbol{B}_i^* = \psi(\boldsymbol{X}_i^*)$。令 $\hat{\boldsymbol{y}}_i = \mathcal{H}(\boldsymbol{B}_i)$ 和 $\hat{\boldsymbol{y}}_i^* = \mathcal{H}^*(\boldsymbol{B}_i)$ 分别表示图像视图和文本视图多示例包上标记的预测 (包括新语义)；令 $\Pi_i^j(.)$ 表示取矩阵的第 $i$ 到第 $j$ 列，则 EM3NL 优化如式 (6.1) 所示：

$$\min_{\phi,\psi,\mathcal{H},\mathcal{H}^*} \sum_{i=1}^{T} \mathcal{L}(\Pi_1^l(\hat{\boldsymbol{y}}_i), \boldsymbol{y}_i) + \mathcal{L}(\Pi_1^l(\hat{\boldsymbol{y}}_i^*), \boldsymbol{y}_i) + \lambda_1(\mathcal{R}_1(\hat{\boldsymbol{y}}_i) + \mathcal{R}_1(\hat{\boldsymbol{y}}_i^*)) + \lambda_2 \mathcal{R}_2(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_i^*), \quad (6.1)$$

其中，$\mathcal{L}(\cdot,\cdot)$ 表示在已知标记上的分类误差损失函数；$\mathcal{R}_1(\cdot)$ 是对预测值的正则化项，由于没有关于新标记的监督信息，希望通过该正则化项利用标记之间的关系进一步约束标记的预测值；$\mathcal{R}_2(\cdot,\cdot)$ 是视图一致性正则化项，约束同一对象的不同视图上的预测值尽可能接近。

由于 $\mathcal{Y}' \in \{0,1\}^{l+1}$，且一个对象可能同时与多个对象相关联，因此对于 $\mathcal{L}(\cdot,\cdot)$，采用深度学习中常用的二值交叉熵损失函数，即

$$\mathcal{L}(\Pi_1^l(\hat{\boldsymbol{y}}_i), \boldsymbol{y}_i) = -\sum_{j=1}^{l} y_{i,j} \log \hat{y}_{i,j} + (1 - y_{i,j}) \log(1 - \hat{y}_{i,j});$$

将 $\hat{\boldsymbol{y}}_i$ 替换为 $\hat{\boldsymbol{y}}_i^*$ 即可得 $\mathcal{L}(\Pi_1^l(\hat{\boldsymbol{y}}_i^*), \boldsymbol{y}_i)$。

进一步地，通过 $\mathcal{R}_1(\cdot)$ 利用标记排序约束对新标记的预测值：假设 $\forall v'_j \in \boldsymbol{v}_i'^{+} \geqslant v'_{l+1} \geqslant \forall v'_j \in \boldsymbol{v}_i'^{-}$ 对 $i \in [T]$ 成立，其中 $\boldsymbol{v}_i'^{+} = \{v'_j | y_{i,j} = 1\}$、$\boldsymbol{v}_i'^{-} = \{v_j | y_{i,j} = 1\}$

分别表示第 $i$ 个样本上观察到的相关标记集合和不相关标记集合，$a \geqslant b$ 表示 $a$ 排在 $b$ 之前。这样令新标记排序在正标记与不相关标记之间，无论新标记预测为 1 或 0，均不会使得标记排序损失增大。为了对新标记排在正标记之前或是不相关标记之后进行惩罚，定义：

$$\mathcal{R}_1(\hat{\boldsymbol{y}}_i) = -\sum_{j=1}^{l}(1-y_{i,j})\log(\sigma(\hat{y}_{i,l+1}-\hat{y}_{i,j}))+y_{i,j}\log(1-\sigma(\hat{y}_{i,l+1}-\hat{y}_{i,j})),$$

其中 $\sigma$ 是 sigmoid 函数。

最后，通过最小化预测值的欧氏距离约束同一对象两个视图上的预测尽可能相近，定义：

$$\mathcal{R}_2 = \|\hat{y}_i - \hat{y}_i^*\|^2.$$

## 6.2.2 包级预测与示例级预测

为了方便描述使符号简洁，本节中省略了多示例包以及相应预测值的下标。对于示例级的预测，类似 DeepMIML [125]，首先将示例映射到子标记语义：假设每个标记 (包括新标记) 有 $k$ 个子语义，采用 3 层全连接层 (FC，Fully connected layer)$\mathcal{F}$，将多示例包中的每个示例 $\boldsymbol{b}_i \in \boldsymbol{B}$ 映射到 $\boldsymbol{g} = \mathcal{F}(\boldsymbol{b}_i) \in \mathbb{R}^{(l+1)k}$。

前两层全连接层均采用 Relu 函数作为激活函数，与 DeepMIML 不同的是，最后一层采用 Softmax 函数作为激活函数而非 Sigmoid 函数。应用 Softmax 函数相比于 Sigmoid 函数更加适合新标记检测任务：应用 Softmax 函数由于要求所有子语义上的预测之和为 1，通过优化往往会得到在某些语义上有比较大的预测值，而在其它语义上的预测值均非常接近 0。当已知标记的语义的预测值都非常小的时候，则更有可能将示例预测为包含新标记的示例。

同样地，利用 Reshape 操作将 $\boldsymbol{g}$ 转为一个 $(l+1)\times k$ 的矩阵作为 2D-子语义层 (2D-subconcept layer)。由于利用 Softmax 函数得到的预测值天然有概率的含义，表示属于某个子语义的概率，则根据加法原理，2D-子语义层按行相加则得到了对该示例每个标记上的预测值 $\boldsymbol{f}_i$。单个示例的预测如图 6.1所示。

在现有的多示例学习方法中，根据示例级预测得到包级预测的方法主要包括：(1) 将包中所有示例的预测值平均作为最终对包的标记预测；(2) 将包中所有示例在每个标记上预测的最大值作为最终对包的标记预测。对于 (1)，如果对于某个标记，无关示例的数量远大于相关示例的数量，那么通过平均得到的预

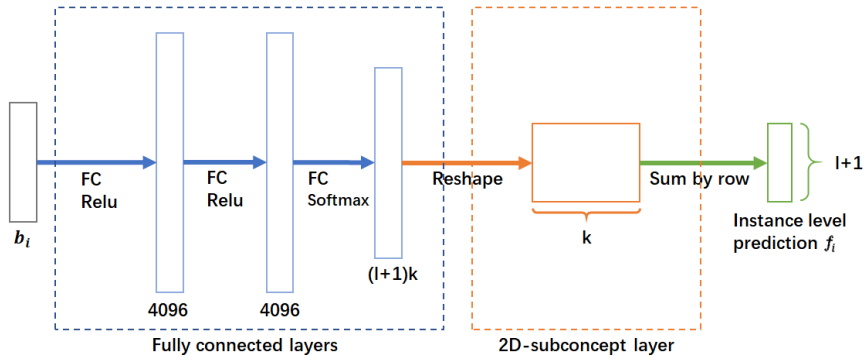



图 6.1 示例级标记预测

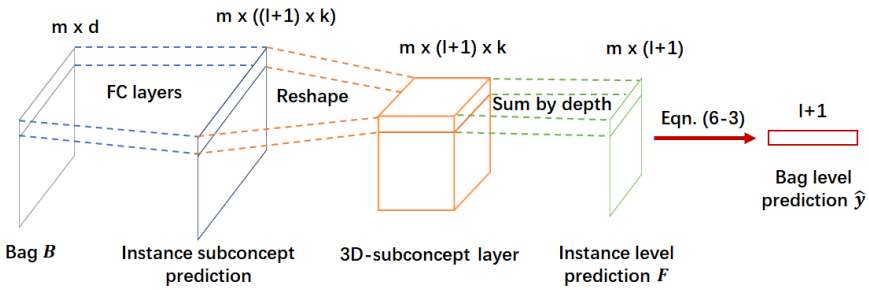



图 6.2 包级标记预测

测值将接近 0，与真实情况相反。对于 (2)，只有部分示例对包的标记预测做出了贡献，而实际上，包中与标记相关的示例可能不止一个，它们对包的标记预测的贡献都应该考虑进去。

由于每个包与多个标记相关联，且假设包中每个示例对应一个标记，本工作对包标记预测的基本思想是着重考虑每个示例上最大预测值对应的标记预测对包的标记预测的贡献：$\boldsymbol{F} = [\boldsymbol{f}_1; \cdots, \boldsymbol{f}_m]$ 表示多示例包 $\boldsymbol{B}$ 中所有示例的示例级预测矩阵，其中预测值经由全连接层和 2D-子语义层输出，$m$ 是包中示例的个数。定义示例预测值权重矩阵为：

$$\boldsymbol{W} = [\frac{1}{m}]^{m\times(l+1)} + \boldsymbol{S}, \tag{6.2}$$

其中 $\boldsymbol{S} = [\boldsymbol{s}_1; \cdots; \boldsymbol{s}_m]$，$\boldsymbol{s}_i = \boldsymbol{e}_{\arg\max_j f_{i,j}}$，$\boldsymbol{e}_k$ 是第 $k$ 位为 1，其余位为 0 的行向量。式 (6.2) 中的第一项是每个示例对每个标记的预测值对包上标记预测的基本贡献权重，第二项代表每个样本在预测值最大的标记上对包上相应标记预测作出更多的贡献。则对包中示例预测加权求和后可得包上的预测为：

$$\hat{\boldsymbol{y}} = \frac{\mathbf{1}^\top(\boldsymbol{W} \circ \boldsymbol{F})}{\boldsymbol{z}}, \tag{6.3}$$

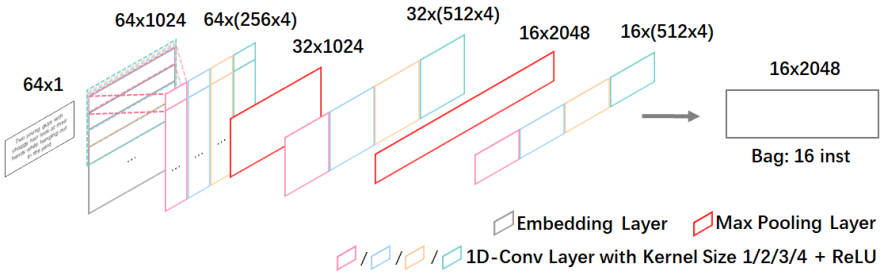



**图 6.3** 辅助文本视图特征学习网络结构

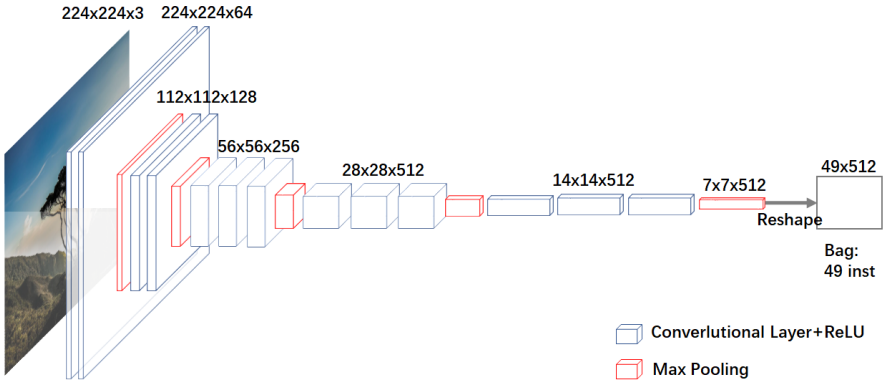



**图 6.4** 原始图像视图特征学习网络结构

其中 $\boldsymbol{z} = \mathbf{1}^{\top}\boldsymbol{W}$ 为归一化系数向量，$\circ$ 表示矩阵对应元素相乘，$\div$ 表示对应元素相除。

综合样本级和包级预测，设计了用于多示例包标记预测的顶层网络结构 $\mathcal{H}$，将多示例包映射到 $\mathcal{Y}'$，如图6.2所示，其中 $\boldsymbol{B}$ 到 $\boldsymbol{F}$ 映射网络结构的横截面即为图6.1展示的示例级标记预测网络结构，且对于包中每个示例子语义 (subconcept) 的预测中，全连接层对所有包中示例共享权重。增广视图多示例包到标记预测网络 $\mathcal{H}^{*}$ 的设计与 $\mathcal{H}$ 类似，只需根据输入示例的维度调整每一层的大小即可。

### 6.2.3 特征提取网络

对文本特征提取，为了便于获取多示例包，没有采用常用的 LSTM 结构，而是基于一维卷积设计了卷积神经网络结构。对于每段文本输入，采用定长为 64 的截断策略，即对单词串长度超过 64 的部分进行截断，不足 64 的进行补 0 操作。利用 Embedding 层将其转换为 $64 \times 1024$ 的实值矩阵。考虑 4 种不

同大小卷积核的一维卷积层，分别对应在每 1-4 个单词对应的向量上做卷积操作，然后将输出拼接作为该层的表示。这样做考虑了单词以及不同长度短语的结构（不同卷积核卷积在边界处可能长度不足，此时利用 Padding 操作进行补 0）。交替增加一维卷积层和最大池化层 (Max Pooling Layer)，得到 16 个 2048 维示例组成的示例包，如图 6.3所示。其中粉色/蓝色/橙色/绿色的矩形分别表示在 1/2/3/4 个单词对应的向量上做 1 维卷积，红色矩形代表最大池化层，输出每两个单词对应表示的最大值。最后得到的每个示例为一个长度为 4 的单词子串的相关语义的表示。

对图像特征提取，采用标准的 VGG 网络结构，输入为一张 $224 \times 224 \times 3$ 的彩色图像，输出为 49 个 512 维示例组成的多示例包。其中每个示例对应的是原始图片 1/49 的一个小块。图像特征提取网络结构如图 6.4所示。

## 6.3 实验测试

本节在 MS-COCO 数据集上验证了 EM3NL 方法处理综合增广学习问题的有效性。

### 6.3.1 实验设置

实验数据集为 MS-COCO [126] 数据集，一共包括超过 12 万张图片，且每张图片关联 5 句文本描述。MS-COCO 数据中一共有覆盖了 80 个不同标记，平均每个样本有 2.91 个相关标记。随机将 MS-COCO 划分训练集和测试集，其中 2/3 样本作为训练数据，剩余的 1/3 作为测试数据。为了验证 EM3NL 发现新标记的性能，分别将不同数量的标记从训练集样本标记中去除 (随机去除 1 个/5 个/10 个标记)，而测试集中仍保留这些标记，将去除标记后的数据集分别命名为 MS-COCO-1，MS-COCO-5，MS-COCO-10。在训练过程中，每一轮迭代乱序读取 16 条数据，以模拟动态样本增广。

从 2 个方面评价模型的性能，包括：(1) 新标记检测的性能；(2) 在已知标记集合上的性能。对于 (1)，如果样本包含训练集标记集合中所去除的任意标记，则该样本被认为存在新标记，采用 AUC 值作为评价指标。对于 (2)，分别采用多标记学习中的四个常用指标 [42] Hamming loss、Ranking loss、Coverage 和 Average Precision (AP) 对模型进行评价，其中 Hamming loss、Ranking loss 和 Coverage 越小越好，Average Precision 越高越好。

由于之前没有端到端的多视图多示例标记新标记学习研究工作（由于 MIMLNC 和 DMNL 方法无法处理大规模的数据，因而也无法在提取的特征上应用它们作为对比），为了验证 EM3NL 的有效性，对比以下方法[①]：

(1) EM3NL-gt：采用 EM3NL 方法进行训练，但是在训练时能够看到新标记作为监督信息。此方法用于显示 EM3NL 没有新标记监督信息进行训练与利用了真实新标记标记信息之间的性能差距；

(2) EM3NL-im：仅使用 EM3NL 的图像视图训练模型，优化：

$$\min_{\phi,\mathcal{H}} \sum_{i=1}^{T} \mathcal{L}(\Pi_1^l(\hat{\boldsymbol{y}}_i), \boldsymbol{y}_i) + \lambda_1 \mathcal{R}_1(\hat{\boldsymbol{y}}_i);$$

(3) EM3NL-txt：仅使用 EM3NL 的文本视图训练模型，优化：

$$\min_{\psi,\mathcal{H}^*} \sum_{i=1}^{T} \mathcal{L}(\Pi_1^l(\hat{\boldsymbol{y}}_i^*), \boldsymbol{y}_i) + \lambda_1 \mathcal{R}_1(\hat{\boldsymbol{y}}_i^*);$$

(4) DeepMIML-mv：利用 $\mathcal{R}_2$ 将 DeepMIML 扩展至多视图模型；由于 DeepMIML 方法不考虑新标记，则将多示例包中所有预测值均低于 0.5 的示例当作新标记示例，认为包含该示例的包有新标记。

### 6.3.2 实验结果

新标记预测的结果 (AUC) 如图 6.5所示。EM3NL 在新标记预测任务上的结果仅次于 EM3NL-gt，它超过两个单视图版本 EM3NL-im 和 EM3NL-txt，远超 DeepMIML-mv。DeepMIML-mv 的设计不考虑新标记，因此它倾向于将包中示例预测为已知标记，因此对新标记的预测接近随机猜测。EM3NL-gt 从训练阶段就能利用真实新标记信息，因此在新标记预测的结果最好，而 EM3NL 尽管在训练时没有新标记的监督信息，仍然在测试集上对新标记进行预测时，取得了接近 EM3NL-gt 的性能，充分验证了 EM3NL 在新标记预测任务上的有效性。

进一步观察 MS-COCO-1、MS-COCO-5 和 MS-COCO-10 在示例级对新标记的预测结果，并以此绘制热度图。图 6.6 展示了 3 例原始图片和相应新标记预测的热度图，每个数据集对应一例，热度图颜色由深蓝至浅黄表示预测值从 0 到 1。尽管示例级的新标记预测与原图的相关语义的位置并不完全吻合，但是

[①]所有方法的优化求解均采用反向传播算法和 Adadelta 算法 [127]。

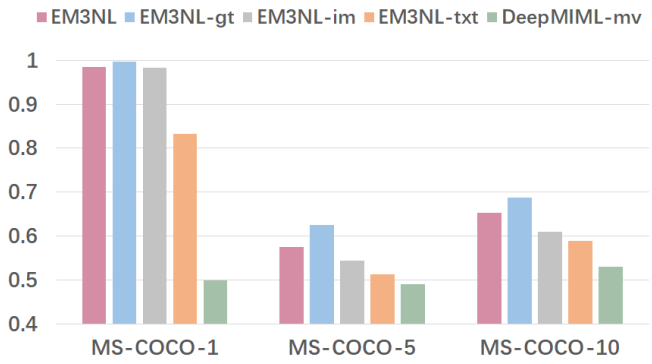
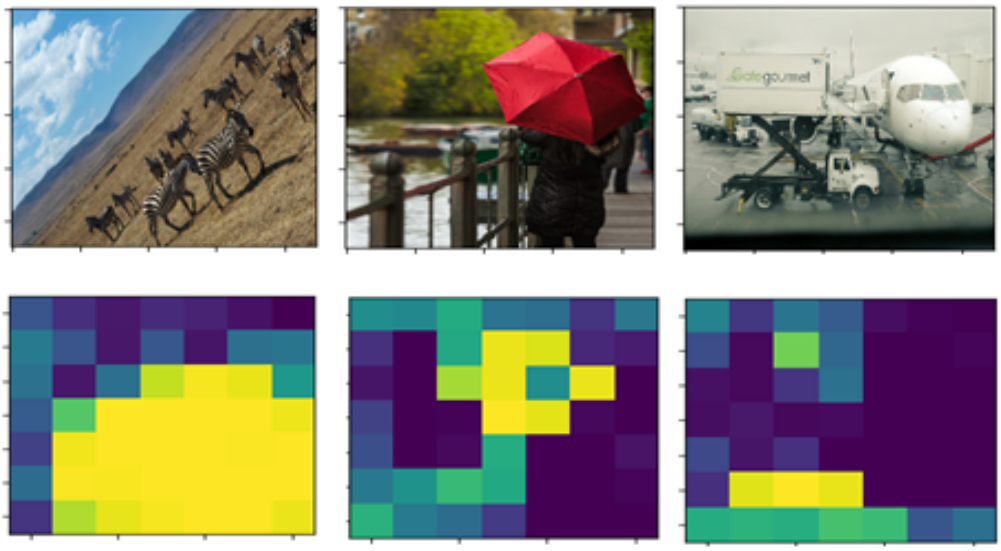



**图 6.5** 新标记检测 AUC 对比结果

二者之间有着非常强的相关性。考虑到这些新标记在训练的时候完全未被标注出来，说明 EM3NL 能够相对准确的在图片中定位到新语义的位置。



MS-COCO-1 样例
真实标记为“Zebra”

MS-COCO-5 样例
真实标记为“Umbrella”

MS-COCO-10 样例
真实标记为“Truck”

**图 6.6** 上部为原始图片，下部为示例级的新标记预测热度图，颜色由深蓝至黄色代表热度由 0 到 1

在已知标记上评价指标的结果包括 Hamming loss、Ranking loss、Coverage 和 Average Precision（AP）分别如图 6.7所示。同样在已知标记上，EM3NL 在大部分指标上取得了除 EM3NL-gt 外最好的性能。这说明 EM3NL 预测新标记，并不会损害对已知标记的预测效果。这进一步说明了 EM3NL 的有效性。

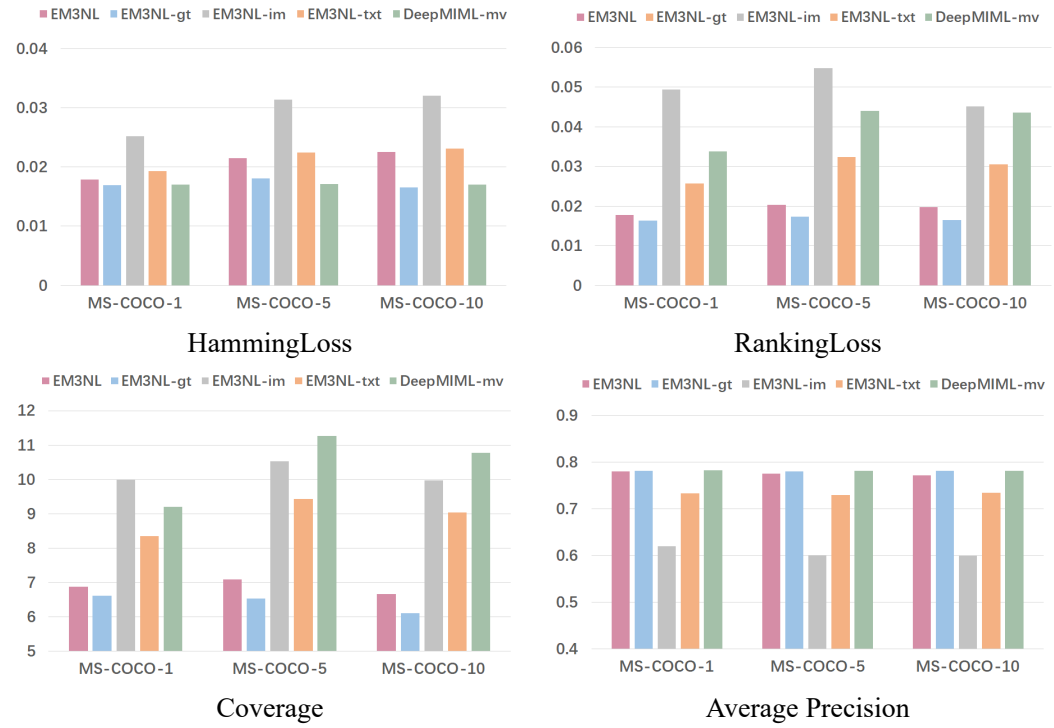



**图 6.7** 在已知标记上的性能比较

## 6.4 本章小结

针对同时进行标记、特征、样本增广学习的综合增广学习问题，提出了一种基于多视图多示例多标记新标记学习方法 EM3NL，将深度神经网络与多视图、多示例多标记学习相结合，设计了一种直接输入图像-文本视图（特征增广），对已知标记预测和新标记检测的端到端统一解决方案（标记增广），并能够通过小批量优化即时对模型更新（样本增广）。在 MS-COCO 数据集上验证了本方法处理标记、特征、样本综合增广学习问题的有效性。

本工作已总结成文：

- 朱越, 姜远, 周志华. 一种基于多示例多标记学习的新标记学习方法. **中国科学：信息科学**, 已录用.（国内核心期刊）

# 第七章　结束语

本文工作主要涉及两个国家自然科学基金项目“开放动态环境下的在线机器学习理论与方法”(61333014) 和“新型深度学习模型与方法的研究”(61751306) 的内容。

传统监督学习通常假设训练数据类别标记恒定、特征信息充分、样本充足。但很多现实的机器学习任务不满足这些假设条件，导致学习效果不尽人意。为此，本文考虑引入增广信息，提出了增广信息学习。增广信息包括传统学习方法中未考虑的信息以及动态过程中出现的新信息，例如新的类别标记、额外的特征、动态增加的样本等。本文的研究工作主要取得了以下创新成果：

在第二章中，针对训练集标记增广学习问题提出了 GLOCAL 方法。GLOCAL 能够通过学习隐标记表示及优化标记流形同时恢复缺失标记，训练分类器，探索与利用全局和局部标记关系。与之前的工作相较，它同时利用全局和局部标记关系，且直接通过数据学习标记拉普拉斯矩阵，而不需要其它关于标记关系的先验知识。此外，GLOCAL 为标记完整和有标记缺失的多标记学习提供了一个统一的解决方案。实验结果显示本文方法能够有效探索全局和局部标记关系，恢复训练集缺失标记，并在预测指标上有明显提高。

在第三章中，分别针对静态和动态测试集标记增广学习问题提出了 DMNL 和 MuENL 方法。DMNL 基于多示例多标记框架，将静态测试集标记增广学习问题形式化成一个有非负正交约束的优化问题，包括示例包损失项和一个聚类正则化项，使得已知标记和新标记可以被同时建模。MuENL 同时利用特征与标记预测值构造新标记检测器，并为新标记设计了鲁棒的更新模型。实验结果表明 DMNL 和 MuENL 方法分别在静态数据集和动态数据流上能够有效地预测出新标记，在整体性能上相较其它方法也取得了更好的结果。

在第四章中，针对多示例特征增广学习问题提出了 AMIV-lss 方法。AMIV-lss 首先将带噪增广特征问题形式化为增广多示例视图学习问题，利用多示例学习框架的特性减少噪声对学习性能带来的影响。AMIV-lss 通过在两个异构视图（原始单示例视图和一个增广多示例视图）之间建立公共隐藏语义子空间，从而利用视图之间的结构关系。然后在所学到的子空间中完成分类任

务。实验结果表明本文方法显著提升了学习性能，优于对比的仅利用原始特征的学习方法和其它利用增广特征的学习方法。

在第五章中，针对多视图样本增广学习提出了 OPMV 方法。传统的多视图学习方法复杂度高、需要反复扫描数据，对于新增样本需要从头训练，OPMV 只需要扫描一遍数据，而无需将所有数据载入内存即能根据每个新增样本对模型进行高效地更新。该方法基于多视图一致性约束下的组合目标函数优化。理论分析证明了 OPMV 的收敛速率为 $O(1/\sqrt{T})$。实验结果表明该方法有着很好的泛化性能，而且在计算效率上比其它方法有巨大提升。

在第六章中，针对同时进行标记、特征、样本增广学习的综合增广学习问题提出了 EM3NL 方法。EM3NL 是一种多视图多示例多标记新标记学习方法，将深度卷积神经网络与多视图、多示例多标记学习相结合，设计了一种直接输入图像-文本视图（特征增广），对已知标记预测和新标记检测的端到端统一解决方案（标记增广），并能够通过小批量优化即时对模型更新（样本增广）。实验结果表明本文方法能够检测出新标记，并能够在找到该新标记的对应区域，在整体性能指标上也有显著提升。

在将来的工作中主要有以下有待进一步研究的问题：

**信息增广的过程中发生了概念漂移：** 本文工作假设增广信息不会影响关于已知信息的数据分布，例如第二章和第三章中，已知标记上的数据真实分布是不变的；第四章特征增广学习中，已知特征上的数据分布也保持不变；第五章样本增广学习中所有多视图样本都采自同一分布。但在有些应用中，随着新的信息的出现和加入，还会对已有数据分布造成影响，即发生了概念漂移现象。那么针对这种概念漂移如何进行学习，是有待将来进一步思考和研究的问题。

**更多类型增广信息综合的学习问题：** 本工作考虑的综合增广信息学习同时进行测试集标记、特征和样本增广学习，但是假设训练集上的已知标记没有缺失。除了这种设定外，还有更多类型的综合增广信息学习问题有待研究。例如在考虑测试集标记、特征、样本综合增广学习时，同时考虑观察标记存在部分缺失的情况，对训练集标记进行增广。

# 参考文献


[1] Z.-H. Zhou and Z.-Q. Chen. Hybrid decision tree. *Knowledge-Based Systems*, 2002, 15(8):515–528.

[2] B.-J. Hou, L. Zhang, and Z.-H. Zhou. Learning with feature evolvable streams. In *Proceedings of the Advances in Neural Information Processing Systems 30*, Long Beach, CA, 2017, pp.1416–1426.

[3] C. Hou and Z.-H. Zhou. One-pass learning with incremental and decremental features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, abs/1605.09082, in press.

[4] I. Kuzborskij, F. Orabona, and B. Caputo. From n to n+1: multiclass transfer incremental learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp.3358–3365.

[5] Q. Da, Y. Yu, and Z.-H. Zhou. Learning with augmented cass by exploiting unlabeled data. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada, 2014, pp.1760–1766.

[6] X. Xu, W. Wang, and J. Wang. A three-way incremental-learning algorithm for radar emitter identification. *Frontiers of Computer Science*, 2016, 10(4):673–688.

[7] S. Rüping. Incremental learning with support vector machines. In *Proceedings of the 1st IEEE International Conference on Data Mining*, San Jose, CA, 2001, pp.641–642.

[8] X. Mu, K. M. Ting, and Z.-H. Zhou. Classification under streaming emerging new classes: A solution using completely random trees. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(8):1605–1618.

[9] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[10] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2007, 69(2-3):169–192.

[11] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Haifa, Israel, 2010, pp.14–26.

[12] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 2010, 11:2543–2596.

[13] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 2009, 22(5-6):544–557.

[14] D. Pechyony and V. Vapnik. On the theory of learnining with privileged information. In *Proceedings of the Advances in Neural Information Processing Systems 24*, Vancouver, Canada, 2010, pp.1894–1902.

[15] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, pp.825–832.

[16] S. You, C. Xu, Y. Wang, C. Xu, and D. Tao. Privileged multi-label learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2017, pp.3336–3342.

[17] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *Proceedings of the 24th Conference on Artificial Intelligence*, Atlanta, GA, 2010, pp.593–598.

[18] M. Xu, R. Jin, and Z.-H. Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Proceedings of the Advances in Neural Information Processing Systems 26*, Lake Tahoe, NV, 2013, pp.2301–2309.

[19] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31th International Conference on Machine Learning*, Beijing, China, 2014, pp.593–601.

[20] G. Tsoumakas, I. Katakis, and L. Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 2011, 23(7):1079–1089.

[21] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 2001, 13(7):1443–1471.

[22] F. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining*, Pisa, Italy, 2008, pp.413–422.

[23] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Conference on Computational Learning Theory*, Madison, WI, 1998, pp.92–100.

[24] W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *Proceedings of the 18th European Conference on Machine Learning*, Warsaw, Poland, 2007, pp.454–465.

[25] W. Wang and Z.-H. Zhou. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp.1135–1142.

[26] Z.-H. Zhou, D.-C. Zhan, and Q. Yang. Semi-supervised learning with very few labeled training examples. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2007, pp.675–680.

[27] W. Wang and Z.-H. Zhou. Multi-view active learning in the non-realizable case. In *Proceedings of the Advances in Neural Information Processing Systems 23*, Vancouver, Canada, 2010, pp.2388–2396.

[28] W. Wang and Z.-H. Zhou. On multi-view active learning and the combination with semi-supervised learning. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008, pp.1152–1159.

[29] K. Chaudhuri, S.-M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009, pp.129–136.

[30] Y.-H. Guo. Convex subspace representation learning from multi-view data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, WA, 2013, pp.387–393.

[31] Y.-H. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, UK, 2012, pp.1615–1622.

[32] M. White, Y. Yu, X. Zhang, and D. Schuurmans. Convex multi-view subspace learning. In *Proceedings of the Advances in Neural Information Processing Systems 25*, Lake Tahoe, NV, 2012, pp.1673–1681.

[33] S.-Y. Li, Y. Jiang, and Z.-H. Zhou. Partial multi-view clustering. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada, 2014, pp.1968–1974.

[34] S.-J. Huang, S. Chen, and Z.-H. Zhou. Multi-label active learning: Query type matters. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp.946–952.

[35] N. Gao, S.-J. Huang, and S. Chen. Multi-label active learning by model guided distribution matching. *Frontiers of Computer Science*, 2016, 10(5):845–855.

[36] W. Bi and J. T. Kwok. Multilabel classification with label correlations and missing labels. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada, 2014, pp.1680–1686.

[37] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *Proceedings of the 14th IEEE International Conference on Data Mining*, Shenzhen, China, 2014, pp.1067–1072.

[38] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010, 20(4):1956–1982.

[39] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th International Conference on Machine Learning* Montreal, Canada, 2009, pp.457–464,.

[40] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. Transduction with matrix completion: Three birds with one stone. In *Proceedings of the Advances in Neural Information Processing Systems 23*, Vancouver, Canada, 2010, pp.757–765.

[41] J. Lee, S. Kim, G. Lebanon, Y. Singer, and S. Bengio. Llorma: Local low-rank matrix approximation. *The Journal of Machine Learning Research*, 2016, 17(1):442–465.

[42] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(8):1819–1837.

[43] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 2004, 37(9):1757–1771.

[44] J. Fürnkranz, E. Hüllermeier, E. Mencía, and K. Brinker. Multilabel classifica-

tion via calibrated label ranking. *Machine Learning*, 2008, 73(2):133–153.

[45] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 2011, 85(3):333–359.

[46] K. H. Huang and H. T. Lin. Cost-sensitive label embedding for multi-label classification. *Machine Learning*, 2017, 106:1725–1746.

[47] X.-Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang. Multi-label dictionary learning for image annotation. *IEEE Transactions on Image Processing*, 2016, 25(6):2712–2725.

[48] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. Wang. Learning deep latent space for multi-label classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp.2838–2844.

[49] S.-J. Huang and Z.-H. Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, Toronto, Canada, 2012, pp.949–955.

[50] W. Weng, Y. Lin, S. Wu, Y. Li, and Y. Kang. Multi-label learning based on label-specific features and local pairwise label correlation. *Neurocomputing*, 2018, 273:385–394.

[51] K. Punera, S. Rajan, and J. Ghosh. Automatically learning document taxonomies for hierarchical classification. In *Proceedings of the 14th International Conference on WWW*, Chiba, Japan, 2005, pp.1010–1011.

[52] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 2010, pp.999–1008.

[53] J. Petterson and T. Caetano. Submodular multi-label learning. In *Proceedings of the Advances in Neural Information Processing Systems 24*, Granada, Spain, 2011, pp.1512–1520.

[54] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006, 7:2399–2434.

[55] S. Melacci and M. Belkin. Laplacian support vector machines trained in the primal. *Journal of Machine Learning Research*, 2011, 12:1149–1184.

[56] D. Luo, C. Ding, H. Huang, and T. Li. Non-negative laplacian embedding. In

*Proceedings of the 9th IEEE International Conference on Data Mining*, Miami, FL, 2009, pp.337–346.

[57] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102(43):15545–15550.

[58] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 2007, 3(1):140–149.

[59] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated green's function. In *Proceedings of the 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp.2029–2034.

[60] F. Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

[61] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 2014, 15:1455–1459.

[62] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *Proceedings of the Advances in Neural Information Processing Systems 15*, Vancouver, Canada], 2002, pp.721–728.

[63] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[64] P. Duygulu, K. Barnard, J. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, 2002, pp.97–112.

[65] M.-L. Zhang. A k-nearest neighbor based multi-instance multi-label learning algorithm. In *Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence*, Arras, France, 2010, pp.207–212.

[66] Z.-H. Zhou and M.-L. Zhang. Multi-label learning. In *C. Sammut, G. I. Webb, eds. Encyclopedia of Machine Learning and Data Mining, Berlin: Springer*, 2017, pp.875–881.

[67] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X-R. Wang, and C.-J. Lin. LIBLINEAR:

A library for large linear classification. *Journal of Machine Learning Research*, 2008, 9:1871–1874.

[68] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. *Artificial Intelligence*, 2012, 176(1):2291–2320.

[69] N. Nguyen. A new svm approach to multi-instance multi-label learning. In *Proceedings of the 10th IEEE International Conference on Data Mining*, Sydney, Australia, 2010, pp.384–392, .

[70] F. Briggs, X. Z. Fern, and R. Raich. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp.534–542, .

[71] S.-J. Huang, W. Gao, and Z.-H. Zhou. Fast multi-instance multi-label learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada, 2014, pp.1868–1874.

[72] Z.-H. Zhou. Learnware: On the future of machine learning. *Frontiers of Computer Science*, 2016, 10(4):589–590.

[73] A. T. Pham, R. Raich, X. Z. Fern, and J. P. Arriaga. Multi-instance multi-label learning in the presence of novel class instances. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015, pp.2427–2435.

[74] A. T. Pham, R. Raich, and X. Z. Fern. Dynamic programming for instance annotation in multi-instance multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(12):2381–2394.

[75] M.-L. Zhang and Z.-H. Zhou. M3miml: A maximum margin method for multi-instance multi-label learning. In *Proceedings of the 8th IEEE International Conference on Data Mining*, Pisa, Italy, 2008, pp.688–697.

[76] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009, pp.1249–1256.

[77] Z.-H. Zhou and X.-Y. Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 2010, 26(3):232–257.

[78] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-

factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, 2006, pp.126–135.

[79] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon. Spectral relaxation for k-means clustering. In *Proceedings of the Advances in Neural Information Processing Systems 13*, Vancouver, Canada, 2001, pp.1057–1064.

[80] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 2011, 3(1):1–122.

[81] S. Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks*, Hong Kong, China, 2008, pp.1828–1832.

[82] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the 2005 IEEE Conference on Computer Vision*, Beijing, China, 2005, pp.1800–1807.

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

[84] Q. Lou, R. Raich, F. Briggs, and X. Z. Fern. Novelty detection under multi-label multi-instance framework. In *Proceedings of the 23rd IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, 2013, pp.1–6.

[85] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, and Z. Lei. The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In *IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, 2013, pp.1–8.

[86] D. Turnbull, L. Barrington, D. A. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transaction on Audio, Speech & Language Processing*, 2008, 16(2):467–476.

[87] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *Proceedings of the 9th International Conference on Music Information Retrieval*, Philadelphia, PA, 2008, pp.325–330.

[88] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification.

In *Proceedings of the Advances in Neural Information Processing Systems 14*, Vancouver, Canada, 2001, pp.681–687.

[89] X.-Z. Wu and Z.-H. Zhou. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp.3780–3788.

[90] V. Jain, N. Modhe, and P. Rai. Scalable generative models for multi-label learning with missing labels. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp.1636–1644.

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

[92] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Proceedings of the Advances in Neural Information Processing Systems 10*, Denver, CO, 1997, pp.570–576.

[93] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Proceedings of the Advances in Neural Information Processing Systems 15*, Vancouver, Canada, 2002, pp.561–568.

[94] Y.-X. Chen, J.-B. Bi, and J.-Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(12):1931–1947.

[95] J.-R. Foulds and E. Frank. A review of multi-instance learning assumptions. *Knowledge Engineering Review*, 2010, 25(1):1–25.

[96] M. Mayo and E. Frank. Experiments with multi-view multi-instance learning for supervised image classification. In *Proceedings of the 26th International Conference Image and Vision Computing*, Pattaya, Bangkok, 2011, pp.363–369.

[97] W. Li, L.-X. Duan, I. W. Tsang, and D. Xu. Co-labeling: a new multi-view learning approach for ambiguous problems. In *Proceedings of the 12th IEEE International Conference on Data Mining*, Brussels, Belgium, 2012, pp.419–428, .

[98] D. Zhang, J.-R. He, and R.-D. Lawrence. MI2LS: Multi-instance learning from multiple informationsources. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL,

2013, pp.149–157.

[99] P.-W. Foltz, W. Kintsch, and T.-K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 1998, 25(2-3):285–307.

[100] C.-J. Hsieh and I.-S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, 2011, pp.1064–1072.

[101] S.-T. Roweis and L.-K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290(5500):2323–2326.

[102] Z.-H. Zhou, K. Jiang, and M. Li. Multi-instance learning based web mining. *Applied Intelligence*, 2005, 22(2):135–147.

[103] W. Li, D. Dai, M. Tan, D. Xu, and L. V. Gool. Fast algorithms for linear and kernel SVM+. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp.2258–2266.

[104] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 1976, 2(1):17–40.

[105] D. Yogatama and N. Smith. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014, pp.656–664.

[106] R. Zhang and J. Kwok. Asynchronous distributed admm for consensus optimization. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014, pp.1701–1709.

[107] H. Wang and A. Banerjee. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, 2012, pp.1119–1126.

[108] H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, 2013, pp.80–88.

[109] T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference*

*on Machine Learning*, Atlanta, GA, 2013, pp.392–400.

[110] W. Zhong and J. Kwok. Fast stochastic alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014, pp.46–54.

[111] J. Sherman and W. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 1950, 21(1):124–127.

[112] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 2000, 3(2):127–163.

[113] G. Bisson and C. Grimal. Co-clustering of multi-view datasets: A parallelizable approach. In *Proceedings of the 12th IEEE International Conference on Data Mining*, Brussels, Belgium, 2012, pp.828–833.

[114] S. Hussain, C. Grimal, and G. Bisson. An improved co-similarity measure for document clustering. In *Proceedings of the 9th International Conference on Machine Learning and Applications*, Washinton, DC, 2010, pp.190–197.

[115] M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *Proceedings of the Advances in Neural Information Processing Systems 22*, Vancouver, Canada, 2009, pp.28–36.

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

[117] Y. Zhu, J.T. Kwok, and Z.-H. Zhou. Multi-label learning with global and local correlation. *IEEE Transactions on Knowledge and Data Engineering*, 2018, 30(6):1081–1094.

[118] Y. Zhu, K. M. Ting, and Z.-H. Zhou. Discover multiple novel labels in multi-instance multi-label learning. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp.2977–2984.

[119] Y. Zhu, K. M. Ting, and Z.-H. Zhou. Multi-label learning with emerging new labels. *IEEE Transactions on Knowledge and Data Engineering*, 10.1109/TKDE.2018.2810872, in press.

[120] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with applica-

tion to scene classification. In *Proceedings of the Advances in Neural Information Processing Systems 19*, Vancouver, Canada, 2007, pp.1609–1616.

[121] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(8):1798–828.

[122] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 2015, 61:85–117.

[123] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 2015, 521(7553):436–444.

[124] Y. Zhu, W. Gao, and Z.-H. Zhou. One-pass multi-view learning. In *Proceedings of the 7th Asian Conference on Machine Learning*, Hong Kong, China, 2015, pp.407–422, .

[125] J. Feng and Z.-H. Zhou. Deep MIML network. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp.1884–1890.

[126] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, and D. Ramanan. Microsoft COCO: common objects in context. In *Proceedings of the 13th European Conference on Computer Vision*, 2014, pp.740–755.

[127] M. D. Zeiler. Adadelta:an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

# 致　　谢

毕业论文写到了致谢一节，七年南京大学生活的终点线已近在眼前。此刻的我，抑制不住思绪在脑海旋转，眼前浮现了回忆的走马灯，七年的一点一滴一幕一幕晃过，清晰地仿若发生在昨日：有过茫然不知所措的彷徨无助，有过奋不顾身的勇往直前，有过屡败屡战的锲而不舍，也有过成功的欢欣若狂……一时之间百感交集，然后渐渐平复，最终沉淀成一种情感，感激。感谢所有出现在我身边的人，深深地感谢他们对我的帮助和鼓励。

首先要感谢我的导师周志华老师。周老师的睿智，对专业领域的博闻强识与深刻洞察，令我折服。在平时的讨论班中，周老师经常为我们梳理出知识的历史发展脉络，使我们能够了解问题的前因后果，更好地体会作者的逻辑思路；通过讲读文章，教给我们如何看待学术问题、如何思维，同时也培养了我们的学术品味；对于上台作报告的我们，通过周老师的点评，也逐渐提升了报告水平。在讨论的过程中，周老师对问题总能提出深刻而独到的见解，使我深受启发。在平时的交流中，发给周老师的每一封邮件，总能得到非常及时的回复，在第一时间里得到周老师的指导。尤其是在修改论文的时候，有的时候发初稿发得比较晚，周老师会熬夜逐字逐句进行修改，还注明修改的逻辑。经由周老师修改后的语句，简明凝练，我从中学到了很多。除了学术之外，在平时生活中，周老师也教给我们许多做人做事的道理。我做了错事后，周老师在严厉批评的同时，也是非常耐心的指导，督促我改正。可以说，我在读博期间取得的每一点进步都离不开周老师的帮助，非常感谢周老师对我的指导、耐心和宽容。

其次，感谢 Kai Ming Ting 老师和 James Kwok 老师的指导。除了周老师外，我在这两位老师的指导下也学到了很多东西。和他们合作期间，几乎每周都有一到两次的单独讨论。Kai Ming Ting 老师做工作总是从大量实验入手，每个实验设计都非常精妙，有理有据地支撑科研论述。从他那，我学到了很多具体的实验设计、分析、写作的思路和思想。James Kwok 老师做工作谋定而后动，把所有的问题形式化、优化都推敲精细后，实验总能取得预期效果，非常高效。从他那，我学到了一些解决问题的思路方法以及优化技术。非常感谢两

位老师。

感谢 LAMDA 实验室的其他老师们、同学们。身处这样一个充满活力的集体中，倍感温暖。在学术上，很多老师、师兄都给过我学术上的帮助；在生活上，大家平时的笑闹给科研生活增加了许多调味，令人难忘。特别感谢高尉师兄，记得 2015 年年前的整整两周，几乎没有怎么休息，带着我从无到有一起拼出一个工作。正是这个工作，正是这么一拼，让我找到了点科研的感觉，带着我走出了低谷。现在想起来除了感激还是感激。感谢姜远老师对我学习生活的悉心关照。感谢黎铭老师、詹德川老师、俞扬老师，李楠师兄、李宇峰师兄、王魏师兄等老师、师兄师姐们在科研和其他事务、工作上给过我的帮助。感谢钱超师兄、黄圣君师兄、庞明师弟带着我一起打球；感谢徐淼师姐、侯博建师弟、吴西竹师弟等和我一起健身的小伙伴们；也要感谢周宇航师弟、刘冲师弟组织远足活动。因为你们，我的生活不再单调。同样感谢和我一届的博士生慕鑫同学，漫漫学术路上砥砺同行。

最后要感谢家人的理解支持，他们是我坚强的后盾。七年以来，他们一直默默分享着我所经历的酸甜苦辣，为状态低迷时的我而焦虑，为熬夜拼搏时的我担忧，为取得研究成果时的我高兴。父母所有的这些情感，最终化成的是我回家时的一桌好菜，两三杯好酒，四五句关怀。七年了，儿子终于也要毕业了，非常感谢父母的默默付出，希望他们能够好好享受退休生活，不用再为我起伏不定的状态操心。

# 附录 A　攻读博士学位期间的学术成果和获奖情况

## 一、攻读博士学位期间发表的论文

1. Y. Zhu and J. T. Kwok and Z.-H. Zhou. Multi-Label Learning with Global and Local Correlation. **IEEE Transactions on Knowledge and Data Engineering**, 2018, 39(6):1081–1094.（CCF-A 类期刊）
2. Y. Zhu and K. M. Ting and Z.-H. Zhou. Multi-Label Learning with Emerging New Labels. **IEEE Transactions on Knowledge and Data Engineering**, in press.（CCF-A 类期刊）
3. 朱越, 姜远, 周志华. 一种基于多示例多标记学习的新标记学习方法. **中国科学：信息科学**, 已录用.（国内核心期刊）
4. Y. Zhu and K. M. Ting and Z.-H. Zhou. Discover multiple novel labels in multi-instance multi-label learning. In: **Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)**, San Francisco, CA, 2017, pp.2977-2983.（CCF-A 类会议）
5. Y. Zhu and K. M. Ting and Z.-H. Zhou. New class adaptation via instance generation in one-pass class incremental learning.. In: **Proceedings of the 17th IEEE International Conference on Data Mining (ICDM'17)**, Orleans, LA, 2017, pp.1207-1212.（CCF-B 类会议）
6. Y. Zhu and K. M. Ting and Z.-H. Zhou. Multi-label learning with emerging new labels. In: **Proceedings of the 16th IEEE International Conference on Data Mining (ICDM'16)**, Barcelona, Spain, 2016, pp.1371-1376.（CCF-B 类会议）
7. Y. Zhu and W. Gao and Z.-H. Zhou. One-pass multi-view learning. In **Proceedings of the 7th Asian Conference on Machine Learning (ACML'15)**, Hong Kong, 2015, JMLR: W&CP 45, pp.407-422.（机器学习领域重要国际会议）
8. Y. Zhu and J. Wu and Y. Jiang and Z.-H. Zhou. Learning with augmented multi-

instance view. In: **Proceedings of the 6th Asian Conference on Machine Learning (ACML'14)**, Nha Trang, Vietnam, 2014, JMLR: W&CP 39, pp.234-249.（机器学习领域重要国际会议）

9. K. M. Ting, Y. Zhu, Z.-H. Zhou. Isolation kernel and its effect to SVM. In: **Proceedings of the 24nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'18)**, London, UK, 2018, pp.xx-xx. （CCF-A 类会议）
10. K. M. Ting, Y. Zhu, M. Carman, Y. Zhu, and Z.-H. Zhou. Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure. In: **Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'16)**, San Francisco, CA, 2016, pp.1205-1214. （CCF-A 类会议）

## 二、攻读博士学位期间获得的主要奖励

1. 南京大学优秀博士生培养计划，2016 年，2017 年
2. AAAI 学生旅行奖，2017 年

## 三、攻读博士学位期间参加的主要科研项目

1. 国家自然科学基金项目“开放动态环境下的在线机器学习理论与方法”(61333014)
2. 国家自然科学基金项目“新型深度学习模型与方法的研究”(61751306)