arXiv:2305.05518v2 [cs.LG] 4 Dec 2024

多标签学习的最小学习机

Joonas Hämäläinen于韦斯屈莱大学信息技术学院芬兰
joonas.k.ha
malainen@jyu.fi

Antoine Hubermont那慕尔大学，比利时那慕尔；Telespazio，比利时巴斯托涅
antoine.hubermont@unamur.be

Amauri Souza
塞阿拉联邦教育、科学与技术学院—IFCE 计算机科学系 巴西
amauriholanda@ifce.edu.br

César L. C. Mattos
塞阿拉联邦大学—UFC 计算机科学系
巴西
cesarlincoln@dc.ufc.br

João P. P. Gomes
塞阿拉联邦大学—UFC 计算机科学系
巴西
jpaulo@dc.ufc.br

Tommi Kärkkäinen
于韦斯屈莱大学信息技术学院
芬兰
tommi.karkkainen@jyu.fi

摘要

基于距离的监督方法——最小学习机通过学习输入距离矩阵与输出距离矩阵之间的映射，从数据中构建预测模型。本文提出了新方法，并评估其核心组件——距离映射如何适用于多标签学习。所提出的方法将距离映射与反距离加权相结合。尽管该方法是多标签学习文献中最简单的方法之一，但在中小规模的多标签学习问题上取得了最先进的性能。除了简单之外，所提出的方法还具有完全确定性：其超参数可通过一种具有闭式形式的基于排序损失的统计量进行选择，从而避免采用传统的基于交叉验证的超参数调优。此外，由于该方法基于简单的线性距离映射构建，我们证明它能够评估多标签分类预测的不确定性；对于以数据为中心的机器学习流水线而言，这是一项很有价值的能力。

关键词 多标签学习 · 多标签分类 · 逆距离加权 · 最小学习机 · 不确定性

1 引言

多标签分类（MLC， [1, 2, 3]）是一种监督式机器学习任务，其中一个实例可以关联多个类别。这与更常见的机器学习任务——单标签分类（SLC）——不同；在单标签分类中，每个实例只关联一个类别，并且仅由一个活跃标签表示。这种简化意味着所有其他可能的类别标签都不相关，但这可能是一种天真的假设，因为许多现实世界的问题本质上都是多标签的。

事实上，多标签分类在许多领域都有相关应用，包括但不限于文本和文档分类 [4, 5, 6, 7, 8, 9, 10], 医学影像 [11, 12, 13, 14], 健康以及

生物信息学 [15, 16, 17, 18, 19], 遥感 [20], 网络安全 [21],脑研究 [22], 电力负荷监测[23], 以及计算化学领域的应用 [24, 25]。

形式上，在多标签分类中，每个实例都与可能标签的一种二分划分相关联，划分为相关标签集和无关标签集。通常，多标签分类方法通过对标签进行评分和/或排序并进行阈值化来预测这种二分划分。因此，多标签分类与多标签排序密切相关。这两项任务共同隶属于多标签学习这一总括概念 [26]。

通常，传统的监督方法并不直接适用于多标签分类问题。处理这些问题的两种主要方法是 [1] i) 算法适配和 ii) 问题转换。第一种方法会对单标签分类方法进行修改，以便在训练过程中以特定方式处理多标签分类中的相关标签。第二种方法则转换标签空间，使其适用于任意单标签分类方法。除了单标签分类方法的核心超参数外，这一步还可能涉及第二层超参数。

本文评估了用于多标签分类问题的基于距离回归的方法。本文的核心观点是，线性距离映射是多标签学习的一种有效构造。我们的主要提议是构建一种多标签分类器，将最小学习机（MLM）的距离回归步骤 [27] 与反距离加权（IDW）方法 [28]相结合。这种组合称为多标签最小学习机（ML-MLM），它构成了一种简单的多标签分类方法，完全具有确定性，其超参数可通过一种具有闭式形式的基于排序损失的统计量来选择。此外，我们还展示了 ML-MLM 能够直观地评估预测不确定性；我们认为，对于以数据为中心的机器学习流水线而言，这是一项很有价值的特性。

为了支持我们的核心观点并证明所提方法的可行性，我们给出了理论结果和大量实验结果。首先，我们展示了理论结果，以支持并阐释将线性距离映射与 IDW 方案相结合的方法。其次，我们针对四种基于距离回归的变体开展了广泛的实验研究（十个基准数据集、八项核心多标签分类性能指标），并将结果与最先进水平进行了比较。我们重点将所提方法与一种基于随机森林的方法（预测聚类树随机森林） [29], 进行比较；在大量实验比较中，该方法一直是表现最佳的方法之一 [26, 3]。实验结果表明，对于规模较小至中等的多标签分类问题，ML-MLM 及其他基于距离回归的方法能够与最先进水平相媲美，同时也是多标签分类文献中最简单的方法之一。此外，根据实验结果，基于距离回归的方法明显优于多标签 k近邻（ML-kNN） [30]。

Bemporad [31]将反距离加权与径向基函数（RBF）相结合，用于探测全局优化领域中代价高昂的目标函数评估。这种反距离加权与RBF的结合在方法上与ML-MLM有些相似，因为RBF与MLM密切相关 [32]。另一项与我们的工作密切相关的研究来自Zhang和Zhou [30]。该论文提出了ML-kNN，它与我们的方案相似，因为ML-kNN也利用特征空间中的距离进行多标签学习。在多标签分类分类体系中，ML-kNN与我们的方法属于同一类别；然而，Bemporad的研究贡献与我们的工作更为接近。此外，这篇论文的灵感来源于MLM在多标签分类中的初步成果所展现出的潜力 [33]。

与所提出的主要方法最接近的是NN-MLM方法 [34]。然而，所提出的方法与NN-MLM有明显区别。首先，所提出的方法能够对标签进行排序，从而将MLM扩展到多标签学习。其次，超参数选择采用基于排序损失的统计量进行设计，而此前尚未提出过这种做法。第三，所提出的方法能够预测新的标签组合，而NN-MLM仅限于训练数据中出现过的标签集。这些都是与基于反距离加权的MLM方法相关的贡献。此外，我们还针对基于MLM的多标签分类提出了许多较小但新颖的方法改进和方案。我们提出： i) 基于距离回归模型的逐实例阈值化（局部RCut）； ii) 提议将与最小BAN对应的参考点用于LLS-MLM； iii) 为NN-MLM在多标签分类中的应用提供理论见解； iv) 详细比较基于MLM的多标签分类方法； v) 分析基于距离回归的方法的时间复杂度； vi) 提供关于如何使用基于距离回归的方法评估不确定性的见解。此外，我们还全面回顾了相关文献，并对最先进的多标签分类方法进行了比较。总而言之，我们的主要贡献如下：

- 一种将反距离加权与 MLM 融合用于多标签分类的新方法
- 多标签分类问题中 MLM 的理论结果
- MLM 的方法改进
- 多标签分类的全面文献综述
- 对最先进水平的多标签分类方法进行全面的实验比较

本文其余部分的结构如下。首先，在第2节中，我们回顾多标签分类方法和实验比较，并总结与我们工作最密切相关的文章。然后，在第3节中，符号表示和基本

介绍方法。此外，我们还给出了 MLM 的理论结果。然后，在第3.2节中介绍本文的主要方案 ML-MLM。接下来是论文的实验部分，见第4节。最后，第5节和第6节对论文进行了讨论和总结。

2 背景

本节首先在第2.1节回顾多标签分类中最核心的方法，为本文提供必要的背景知识。随后，我们在第2.2节回顾与本文主要贡献最为密切相关的文章。最后，我们在第2.3节总结多标签分类的实验结果，为本文的结果提供背景。

2.1 多标签分类简述

解决多标签分类（MLC）问题最直接的方法是所谓的二元关联（BR）问题转换方法 [35, 1]。在 BR 方法中，采用一对其余方法，针对每个标签分别训练模型。随后，直接根据这些针对各标签的单标签分类模型的预测结果确定预测标签。BR 存在一些众所周知的缺点。当标签数量较大时，BR 以及其他问题转换方法的计算复杂度可能很高。此外，BR 假设训练时无需考虑标签依赖关系的信息，即每个预测标签仅由输入数据决定。尽管存在这些缺点，BR 仍是多标签分类（MLC）中重要的方法基线。

如 [1] 和第 1 节所示，多标签分类（MLC）技术基本可分为问题转换（*PT*）和算法适配（*AA*）方法。除了二元相关性（BR），标签幂集（LP）也是一种经典的问题转换方法 [36],，它将训练集中每一种不同的标签组合都视为一个单独且不同的类别。事实上，存在许多基于问题转换技术的开源软件包，它们通过调用标准的单标签方法来实现多标签分类算法 [37, 38, 39]。分类器链方法 [40] 通过利用二元相关性模型的预测结果，以链式结构扩展特征空间。在这种链式结构中，训练二元相关性模型时，会使用完整的特征集和一部分预测标签（至少排除目标标签）来预测目标标签。

根据 [2], , 算法适配方法还可进一步分为一阶方法和二阶方法；经典的 k 近邻（kNN）多标签分类扩展 ML-kNN [30], 是前者的例子，而采用修正核的排序支持向量机 [41], 则是后者的例子。一阶算法适配方法也是对单标签多类分类器的直接扩展，能够通过构造生成反映类别概率的输出向量，例如采用 1-of-k 类别编码的浅层和深层神经网络，不过需要采用修正后的标签策略 [42]。在最简单的情况下，可以只返回所有置信度足够高的标签：这些标签可通过阈值 [41, 42, 43], 选出，也可以完全拒绝不可靠的预测 [44], , 或像 ML-kNN [30]那样依据贝叶斯规则，使用估计的后验概率来确定。 [45] 中的 BPNN 方法是对反向传播算法的适配，用于优化神经网络的权重，以实现多标签分类。

除算法适配（AA）和问题转换（PT）外，还可以将多标签分类（MLC）方法分为第三类，即集成方法[26]。这些方法指的是对常见的问题转换方法与算法适配方法进行不同形式的组合。一个典型例子是 Tsoumakas 等人[1]提出的 RAKEL（随机 k 标签集）方法。HOMER 方法 [46] 基于一种构造方式：利用具有层次结构的标签空间的标签幂集。在该结构中，采用分而治之的范式训练一组二元分类器，以预测元标签（标签子集的标签幂集）。在训练阶段，通过 k-means 平衡聚类得到标签的层次结构。集成分类器链（ECC）为 PT 提供了更好地应对多标签分类问题多尺度特性的方法，同时考虑标签依赖关系；ECC 即单标签模型的级联组合 [40, 47, 48]。这些集成方法及其他集成方法都基于一个基分类器，该分类器可以针对不同标签子集 [49, 50], 、不同训练样本 [46, 51], 以及不同特征子集[52, 53, 54], 进行训练，也可以结合这些划分策略 [55]。特征和/或标签也可以被转换到低维空间中，既可用于单个 MLC 模型，也可用于集成 [56, 57, 58, 59]。集成中整合不同基分类器的方法有多种，例如直接对输出取均值 [49], 、选择并使用最突出的分类器 [60], , 以及加权堆叠策略 [61]。RF-PCT 方法 [51] 使用具有层次聚类结构的决策树，作为随机森林（RF）集成方法的基分类器。在 RF 中，集成通过在观测样本和特征两个维度上随机化学习过程来构建，分别采用自助采样和随机特征选择。RF-PCT 的聚类结构通过自顶向下最大化方差缩减来获得。值得注意的是，如 [62]所示，所提出的新型 MLC 分类器 MLM 也可以自然地用作集成中的基分类器。

如 [63, 32], 所示，MLM 技术可以与随机神经网络（NN）技术相关联；针对多标签分类问题的相关研究见 [64]。该研究采用函数链接网络和宽度学习系统作为神经网络架构，每种架构各有两个变体。训练完成后的最终结果通过训练得到的阈值函数计算，类似于 [41]。研究使用 12 个基准数据集，将该方法与另外三种随机基技术进行比较，并采用五种评估指标（汉明损失、单一错误、覆盖率、排序损失和平均精度）。实验结果表明，该方法取得了高质量的结果，并提高了计算效率。
近年来，人们提出了许多新的多标签分类方法和途径。与一般的机器学习类似（见 [65]），一个热门方向是采用并开发更多特征选择（FS）方法，例如过滤式 [66, 67],、包装式 [68], 或混合式／嵌入式方法 [69]。还可以同时选择特征和实例，例如依据它们在潜在主题空间中的依赖关系进行选择 [70, 71]。 [72, 73, 52], 中对多标签分类的特征选择进行了综述，其总体回顾见 [65, 第 2 节6]。总之， i) 大多数特征选择方法属于过滤式； ii) 多标签分类问题的特征选择方法可从四个角度分类：标签、搜索策略、与学习算法的交互方式以及数据格式； iii) 多标签分类问题的过滤式、包装式和嵌入式特征选择方法，分别与单一及内部／外部二元相关性技术存在直接联系。例如, [54, 74] 提出了针对特定标签的特征选择技术。针对多标签分类问题采用迭代搜索策略的特征选择方法也已被提出，例如进化方法 [75] 或多目标优化方法 [76, 77] 。 [78] 提出了利用粗糙集对特征进行排序和选择。

总而言之，多年来，人们针对多标签分类问题提出并试验了大量不同的方法，将新颖组件与异构数据源相结合，和/或对现有技术进行整合与改进。PT和AA这两类基本方法不断发展并相互融合。例如，可以构造新标签[72], 处理标签信息不完整的半监督场景 [79], 处理流式数据以及在线/增量学习 [80, 81], 应用主动学习技术，通过增量查询最相关的实例来提升模型性能 [82],并提取更高层次的特征 [83, 84]。

2.2 距离回归与距离加权方案

本小节将更深入地回顾与本文方法密切相关的研究。我们考察了采用基于距离回归的构造方法进行多标签分类的密切相关工作，也回顾了反距离加权领域最相关的研究。

我们提出的方法采用了来自MLM的距离回归步骤 [27]。因此，与之最密切相关的工作是[33], 该研究通过简单的最近邻启发式方法，将MLM应用于多标签分类问题。在该方法中，通过近似求解多边定位问题来获得分类结果。该方法与NN-MLM [34]的区别仅在于标签空间的假设。命题1（第3节）将进一步讨论这一差异。此外， [33], 提出了一种基于聚类的NN-MLM方法，用于多标签分类，使MLM能够扩展至大规模数据集。初步结果表明，所提出的方法在准确率和汉明损失指标上优于ML-kNN，且性能可与基于随机森林的最先进方法相媲美。本文不再进一步讨论这种基于聚类的NN-MLM多标签分类方法，因为我们将重点放在多个多标签分类指标下，对所提出的方法在小型至中型数据集上进行全面评估。需要注意的是，大规模多标签分类问题的最先进水平与其他问题有所不同。近年来，大多数针对大规模多标签分类问题（包括极端多标签分类）的最先进方法都基于深度学习 [85, 86]。不过，一些基于聚类的方法（例如SLEEC） [87], 在比较中仍表现良好 [88]。

由于简单易用，反距离加权已被用于若干数据驱动方法和应用中 [89, 90, 91, 92, 31, 93]。然而，与我们提出的方法不同，这些工作仅在特征空间中使用反距离加权。在我们提出的方法中，我们在输出空间中使用反距离加权，将经排序损失优化的回归层添加到该方法中，用于对输出空间向量进行局部加权，以实现分类。

Joseph 和 Kang [89] 将反距离加权与线性回归相结合，以提高预测准确率。所提出的方法专为大规模数据集和高维输入空间中的回归任务设计，并能够为预测提供置信区间。Shi 等人 [93] 将反距离加权用于检测分布式光伏电站中的异常。该方法基本沿用原始形式，并采用文献中建议的幂参数选择方式 $P = 2$ （例如 [28]）。该方法的主要思路是使用反距离加权预测电站的输出，再将预测值与实际值进行比较，通过阈值化判断电站是否存在异常运行。Chen 和 Liu [90] 将反距离加权应用于空间回归任务（降雨预测），并研究了幂参数和影响半径参数的表现。幂参数的取值通过交叉验证进行调优，最优值位于区间 $[0, 5]$。

Chen 等人 [91] 将反距离加权与混合效应模型相结合，构建了一种处理缺失数据的两步方法。随后，该方法又与极端梯度提升相结合，开发出一种稳健的缺失数据处理方法

，用于回归。该方法成功应用于在缺失数据比例较高（近 90%）的情况下，预测地面层环境颗粒物的低浓度。在 [92] Cevik 提出了一种基于反距离加权的人脸识别局部描述子。与 [91], 中的目标类似，该描述子旨在更稳健地表示特征空间。具体而言，它旨在解决特征空间中的旋转变化和噪声效应。该反距离加权方法增加了一个距离衰减参数，以便根据距离和像素强度更好地匹配局部像素区域。与最先进水平相比，所提出的方法在分类准确率方面表现良好。Bemporad [31]将反距离加权与径向基函数（RBF）相结合，开发了一种基于该全局优化算法的采集函数，以改进对昂贵目标函数的评估。仅从方法论角度来看，这种反距离加权与 RBF 的组合与本文提出的方法有些相近，因为 MLM 与 RBF 之间存在相似之处。例如，RBF 的通用逼近证明也可应用于 MLM [32]。

2.3 实证结果总结

接下来，我们不求面面俱到，而是总结过去十年中不同方法最相关的实验比较，并与第4节报告的我们自己的实验进行对照。所回顾的实验自2016年起开展，并根据以下两个标准被视为全面： *i*) 使用至少十个数据集，其中至少一个包含超过10 000个实例； *ii*) 比较中至少纳入五种其他算法。

在 [43], 中，作者提出了一种采用支持向量机基分类器的基于层次树的集成，称为*ML-Forest*。该研究基于汉明损失进行参数调优，采用五折交叉验证（CV），并与另外八种最先进水平的算法进行了比较。实验使用了12个数据集，训练实例数最多约为21 500（'Tmc2007'）。根据九项基于二分的质量指标，所提出的技术、基于平衡聚类层次结构的多标签分类器（HOMER， [46]）以及预测聚类树随机森林（RF-PCT， [51]）表现最佳。然而，在四项基于排序的指标上，与RF-PCT和ML-Forest方法相比，两阶段架构（TSA， [94]）和 [1] 中的BR方法表现略优。就训练时间而言，RF-PCT（模型数量固定为50）在表现最佳的方法中速度最快。

在 [50], 集成学习中，研究提出了两种标签选择策略变体，以获得平衡标签子集，并采用以 RBF 核支持向量机为基分类器的一对多多类分类策略。广泛的实验及与其他方法的比较基于 30 个数据集，其中规模最大的数据集包含约 44,000 个样本（"Mediamill"），特征数最多的则接近 13,000 个（"EukaryoteGo"）。研究采用十折交叉验证并重复五次，使用四项质量标准，将提出的两种算法与另外 12 种多标签分类相关方法进行比较：基于标签的准确率（1−汉明损失）、基于样例的 F 度量，以及基于标签的宏平均/微平均 F 度量。总之，各种方法在准确率方面表现相当且令人满意；但在基于样例和基于标签的 F 度量方面，所提方法优于传统 PT 方法，且质量与最先进水平的集成方法相近。有趣的是，在几个数据集（"CS"和"Chess"，仅含二元特征）上，没有任何一种方法取得令人满意的结果。此外，研究还得出结论：与宏平均 F 度量相比，微平均 F 度量在评估和比较不同方法时更加稳定。

在介绍并试验标签特定特征选择方法 [74], 时，基于汉明损失、单错误率、排序损失、平均精度和宏平均曲线下面积等质量指标，将其与 16 个基准数据集及其他七种参考方法进行了比较。使用五折交叉验证计算比较结果及其变异性。实验结果支持所提出的 WRAP（采用标签特定特征生成的多标签分类）方法。

[3] 报告了一项近期非常全面的比较研究，使用 42 个基准数据集比较了 26 种方法。研究使用 20 项评估指标评估了这些方法的预测性能。比较结果表明，基于随机森林的基于树的集成，尤其是 RF-PCT 和 RFDTBR（基于决策树随机森林的二元关联）问题转换方法（本文称之为 BR-RF），始终是表现最佳的策略之一。研究强调，基于多种评估标准和视角选出最佳方法并非易事。特征选择综述 [65] 中通常也会得出这一结论。

最新的实验报告见 [60], ，其中提出了一种方法（MLDE），通过评估集成中动态生成的基分类器的准确率和排序损失，实现基分类器的动态选择与组合。所进行的比较基于 24 个数据集（均来自 http://www.uco.es/kdis/mllresources/），并采用了另外 9 种基准方法：RAKEL [1],ML-kNN [30],ECC [40],ACkELo [50],MLWSE[61], LF-LELC [95], ELIFT [96], BOOMER [97],和 DECC [98]。采用逻辑回归、决策树、贝叶斯网络和支持向量机作为基分类器，并使用准确率、F1度量、汉明损失、单一错误和平均精度

作为质量指标。将数据按 2/3 和 1/3 的比例划分为训练集和测试集，并重复五次后，研究得出结论：所提出的 MLDE 方法始终表现最佳之一。

3 多标签最小学习机

本节安排如下。在第3.1节中，我们首先给出 MLM 的基本表述以及 NN-MLM 近似的理论结果。在第3.2节中，我们基于 MLM 的距离回归映射和反距离加权方案推导主要方案。最后，在第3.3节中，我们推导出基于排序损失的闭式解，用于主要方案的模型选择。

3.1 基本表述及近似结果

多标签分类任务对应于一类监督学习问题，其中输入可以关联多个类别标签。形式上，给定一个数据集 $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$，其中包含输入—输出对 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^M \times \mathbb{Y}$，它们表示来自未知联合分布 $p^\star(\mathbf{x}, \mathbf{y})$ 的独立同分布样本。集合 $\mathbb{Y}$ 表示所有可能多标签赋值的全集，我们将其记为 $\mathbb{Y} = \{0,1\}^L$ — L 是类别总数；若某个样本属于类别 c，则输出 $\mathbf{y}$ 的第 c 个分量设为 1，否则设为 0。我们希望找到一个映射/模型 $h\colon \mathbb{R}^M \to \mathbb{Y}$，使得对于任意 $(\mathbf{x}, \mathbf{y}) \sim p^\star$，$h(\mathbf{x}) = \mathbf{y}$ 以较高概率成立。

最小学习机（MLM，[27]）是一种监督式方法，其训练过程包括拟合一个多响应线性回归模型，用于描述在输入空间和输出空间中计算得到的距离。对于新输入的预测，是通过使用底层线性模型估计距离，然后在可能的输出空间中进行搜索来实现的。

形式上，设 $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ 是一组输入向量 $\mathbf{x}_i \in \mathbb{R}^M$，$\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$ 是其对应的输出 $\mathbf{y}_i$。此外，设 $\mathcal{R} = \{\mathbf{r}_i\}_{i=1}^K \subset \mathcal{X}$ 是 K 长度的子集 $\mathcal{X}$，其对应输出为 $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^K \subset \mathcal{Y}$。对于输入 $\mathbf{x}$，使用参数$\mathbf{B} \in \mathbb{R}^{K\times K}$ 的 MLM 给出的预测 $\hat{}\mathbf{y}$ 为

$$\hat{\mathbf{y}} = h(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{Y}} \left\{ J(\mathbf{y}) = \sum_{k=1}^{K} \left[\|\mathbf{y} - \mathbf{t}_k\|^2 - \left(\sum_{i=1}^{K} B_{i,k} \|\mathbf{x} - \mathbf{r}_i\| \right)^2 \right]^2 \right\}. \tag{1}$$

项 $\sum_{i=1}^K B_{i,k}\|\mathbf{x} - \mathbf{r}_i\|$ 表示$\mathbf{y}$ 与$\mathbf{t}_k$ 之间距离的估计——我们也将其记为 $\hat{}\delta_k$。值得注意的是，公式 1 可视为一个多边定位问题——根据到固定（锚点）位置的距离估计，求出查询点的坐标。

在求解公式 1 中描述的优化问题之前，我们需要确定参数 $\mathbf{B}$。我们将 $\mathbf{D}_\mathbf{x} \in \mathbb{R}^{N\times K}$ 定义为 $\mathcal{X}$ 与 $\mathcal{R}$之间的成对欧几里得距离矩阵，即 $D_{ij} = \|\mathbf{x}_i - \mathbf{r}_j\|$。类似地，$\mathbf{D}_\mathbf{y} \in \mathbb{R}^{N\times K}$ 表示 $\mathcal{Y}$ 与 $\mathcal{T}$之间的成对距离矩阵。在其原始形式中，MLM 采用 $\hat{}\mathbf{B} = (\mathbf{D_x}^T\mathbf{D_x})^{-1}\mathbf{D_x}^T\mathbf{D_y}$，这对应于输入距离与输出距离之间线性模型的普通最小二乘解。

对于类别均衡的 SLC 任务，[34] 表明，公式 1 中的最小化等价于选择由距离回归模型近似得到的最近输出参考点（或类别标签），即我们只需计算

$$\hat{\mathbf{y}} = \mathbf{t}_{k^\star} \quad \text{with} \quad k^\star = \operatorname*{argmin}_{k=1,\ldots,K} \left\{ \hat{\delta}_k = \sum_{i=1}^{K} \hat{B}_{i,k} \|\mathbf{x} - \mathbf{r}_i\| \right\}. \tag{2}$$

这种策略称为最近邻 MLM（NN-MLM），[33] 针对多标签设置提出了该策略。在这方面，命题1表明，对于 δ_k 的一般$\hat{}$值，这一最近邻过程无法为多标签分类产生最优解。然而，当输出距离的估计准确时，NN-MLM 确实对应于公式 1 中多边定位问题的解。这解释了 [33] 中报告的 NN-MLM 在多标签分类方面的高预测性能。我们在附录 A 中给出了证明。

命题1（用于多标签学习的最近邻 MLM）。设 $L \geq 2$ 表示一个满足 $\mathbb{Y} = \{0,1\}^L$ 的多标签分类问题中的类别数。此外，假设输出参考点集合 $\mathcal{T}$ 包含所有可能的多标签赋值。那么，以下结论成立：

$\hat{}$*1.* 对于 δ_k 的任意取值，*NN-MLM* 并不总能为公式 *1* 中的优化问题得出最优解。

ˆ*2.* 若 $\delta_k = \|\mathbf{y}' - \mathbf{t}_k\|$ 对某个 $\mathbf{y}' \in \mathbb{Y}$ 成立，则 *NN-MLM* 会返回公式 *1* 中最小化问题的最优解。

在 MLM 术语中，选择集合 $\mathcal{T}$和 $\mathcal{R}$ 的问题称为参考点选择。这个问题与 Nyström方法中地标点的选择密切相关 [32]。为简化实验和分析，本文重点使用完整的输出空间参考点集合 $\mathcal{T}$，并将输入空间中的所有不同点作为输入空间参考点 $\mathcal{R}$。请注意，MLM文献中提出了多种参考点选择方案 [99, 100, 101, 32] ，以提升模型性能。

3.2 多标签算法

在实践中，满足确保 NN过程最优性（命题1）的要求往往很困难。此外，贪心选择最近标签的方式限制了 NN-MLM 的适用范围，使其只能用于已见过的标签集合；也就是说，预测仅限于训练期间观察到的情况。因此，我们在此采用另一种方案，利用完整的标签集合，将 1-NN 扩展为 N-NN。为此，我们需要定义一种组合标签的方法。总之，我们的方案依赖于计算凸组合，其中重要性权重是标签空间中ˆ距离的函数（使用 δ 的值估计）。受最近邻在多标签设置中的成功启发 [33], ，我们旨在采用这样的加权方案：第一个最近邻的重要性应高于第二个，依此类推。

一种自然的实现方式是采用反距离加权 [28]。这种方案只需一个参数——幂参数——即可控制标签集凸组合的权重。下文将阐述这一思路，并介绍 ML-MLM（多标签 MLM）：一种用于多标签分类、将标准 MLM 中的距离回归步骤与反距离加权相结合的新型基于距离的方法。

与 MLM 类似，ML-MLM 假设输入和输出/标签数据空间中计算得到的距离之间存在线性映射。在最小二乘准则下，我们可以得到线性系数的闭式估计。考虑集合 $\mathcal{X}$、 $\mathcal{Y}$、 $\mathcal{R}$ 和 $\mathcal{T}$，与原始 MLM 中相同，只是 $\mathcal{T} = \mathcal{Y}$——所有观测到的目标也都是输出参考点。此外，设$\mathbf{D_x} \in \mathbb{R}_+^{N\times K}$为 $\mathcal{X}$ 与$\mathcal{R}$之间的成对 ℓ_2距离矩阵，而$\mathbf{D_y} \in \mathbb{R}_+^{N\times N}$为 $\mathcal{Y}$中各元素之间的距离矩阵。ML-MLM 的训练只需计算 ˆ$\mathbf{B}$ $= (\mathbf{D_x}^T\mathbf{D_x})^{-1}\mathbf{D_x}^T\mathbf{D_y}$，其中 ˆ $\mathbf{B}$ $\in \mathbb{R}^{K\times N}$ 表示模型参数。

为了对新的输入$\mathbf{x}$进行多标签预测，ML-MLM 首先应用距离回归模型，得到相对于已观测多标签目标集合 $\mathcal{Y}$ 的距离估计 $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_N]$，即:

$$\hat{\boldsymbol{\delta}} = [\|\mathbf{x} - \mathbf{r}_1\|_2, \cdots, \|\mathbf{x} - \mathbf{r}_K\|_2]\,\hat{\mathbf{B}}, \tag{3}$$

其中，$\mathbf{r}_i$ 表示 $\mathcal{R}$的第 i个元素。接着，我们利用这些距离估计，为 $\mathcal{Y}$中的每个目标定义重要性权重 [28] ，并对已观测的多标签目标进行加权平均:

$$\bar{\mathbf{y}} = \frac{1}{Z}\sum_{i=1}^{N} w_P(\hat{\delta}_i)\mathbf{y}_i, \quad \text{with} \quad w_P(\hat{\delta}_i) = \begin{cases} \hat{\delta}_i^{-P}, & \text{if } \hat{\delta}_i > 0, \\ 1, & \text{if } \hat{\delta}_i = 0, \end{cases} \tag{4}$$

其中， $Z = \sum_{i=1}^{N} w_P$(ˆδ_i)，且 $0 < P \in \mathbb{R}$ 为超参数。具体而言，变量 P控制重要性权重随距离

δ_i $= [y_1 \ \ldots \ y_L] \in$ 增加而衰减的速度。最后，

$\{0,1\}^L$

我们通过对$\mathbf{y}$的值进行阈值化，得到多标签预测ˆ$\mathbf{y}$ˆ,,ˆ:

$$\hat{y}_c = \mathbb{1}[\bar{y}_c > t] \qquad \forall c = 1, \ldots, L, \tag{5}$$

其中， $\mathbb{1}[\cdot]$ 表示指示函数，当其参数为真时返回 1，否则返回 0； t 是一个超参数。算法 1 和算法2概述了 ML-MLM 的训练和预测过程。图1通过一个玩具数据集展示了算法2的预测流程。

算法1距离回归（训练）

要求： 输入数据 $\mathcal{X}$，输出标签 $\mathcal{Y}$，输入空间参考点 $\mathcal{R}$ˆ**确保：** 距离

回归模型**B1:** $\mathbf{D_x} \leftarrow$ 计算 $N \times K$ 输入空间距离矩阵，介于 $\mathcal{X}$和

$\mathcal{R}$

2: $\mathbf{D_y} \leftarrow$ 计算 $N \times N$ 距离矩阵以用于 $\mathcal{Y}$3: ˆ

$\mathbf{B} \leftarrow$ 求解 $(\mathbf{D_x}^T\mathbf{D_x})^{-1}\mathbf{D_x}^T\mathbf{D_y}$

算法2ML-MLM预测

要求： 输入向量 $\mathbf{x}^*$，距离回归模型 $\hat{\mathbf{B}}$、参考点 $\mathcal{R}$、输出标签 $\mathcal{Y}$、幂参数 P，分类阈值 t

确保： 标签向量$\mathbf{y}^*$

$\mathbf{d} \leftarrow [\, \|\mathbf{x}^* - \mathbf{r}_1\|, \ldots\ldots, \|\mathbf{x}^* - \mathbf{r}_K\| \,]$ // 计算输入空间距离

$\boldsymbol{\delta} \leftarrow \mathbf{d}\hat{\mathbf{B}}$ // 预测输出空间距离

$\bar{\mathbf{y}} \leftarrow \sum_i w_P(\hat{\delta}_i)\mathbf{y}_i / Z$ // 计算标签得分（公式 4）

$\mathbf{y}^* \leftarrow \mathbb{1}[\bar{\mathbf{y}} > t]$ // 相关标签阈值

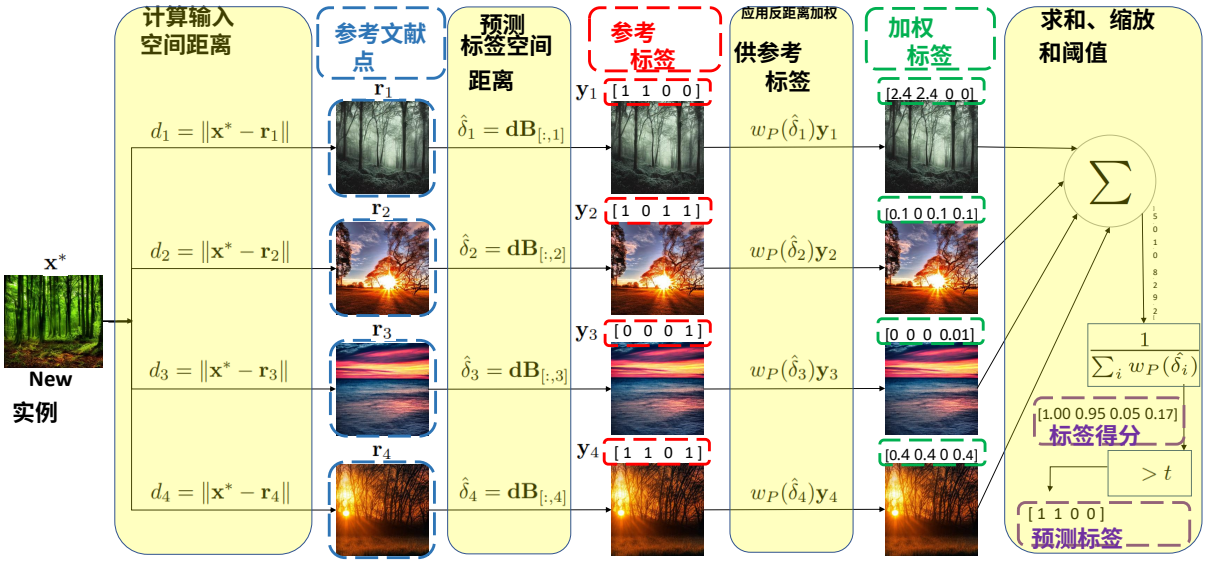


图1：以玩具数据集训练的模型为例，展示了 ML-MLM 进行预测时的主要组成部分。该数据集由四个实例组成，分别对应四个唯一的标签集。距离回归模型$\mathbf{B}$ 使用四个输入空间参考点 $\{\mathbf{r}_i\}_{i=1}^4$ 进行了训练，并对应有标签空间向量 $\{\mathbf{y}_i\}_{i=1}^4$ （在 MLM 的语境中视为输出空间参考点）。IDW 权重使用 $P = 2$计算。此处的示例图像由 Stable Diffusion 创建。

3.3 模型选择

现在我们讨论如何调整超参数 P（幂）和 t（全局类别阈值）。关于P的选择，我们采用多标签问题中广泛使用的评估指标：排序损失 [2]。排序损失适用于像 ML-MLM 这样提供标签得分的方法，因为它衡量了排序的整体质量。我们将该损失与交叉验证方案结合使用，更具体地说，是留一交叉验证（LOOCV）。值得注意的是，由于 ML-MLM 距离回归步骤具有线性特性，我们可以利用预测平方和（PRESS）统计量[102] 以闭式形式计算留一误差。因此，我们只需运行一次 MLM 的训练算法。

在算法1中，系数矩阵$\hat{}\mathbf{B}$ 可以通过Moore-Penrose伪逆求解。对于$\mathbf{U} = {\mathbf{D}_\mathbf{x}}^T\mathbf{D}_\mathbf{x} + \alpha\mathbf{I}$，我们有$\hat{}\mathbf{B} = \mathbf{U}^\dagger {\mathbf{D}_\mathbf{x}}^T\mathbf{D}_\mathbf{y}$。对于训练集中的每个观测值 $i \in \{1, \ldots, N\}$，我们可以确定相应的样本外距离预测$\hat{}\boldsymbol{\delta}_i^{LOO}$为

$$\hat{\boldsymbol{\delta}}_i^{LOO} = \frac{\hat{\mathbf{D}}_{\mathbf{y}(i,:)} - \mathbf{H}_{(i,i)}\mathbf{D}_{\mathbf{y}(i,:)}}{1 - \mathbf{H}_{(i,i)}}, \tag{6}$$

其中，帽子矩阵 $\mathbf{H} = \mathbf{D}_\mathbf{x}\mathbf{U}^\dagger {\mathbf{D}_\mathbf{x}}^T$，且 $\hat{}\mathbf{D}_\mathbf{y} = \mathbf{H}\mathbf{D}_\mathbf{y}$。因此，ML-MLM 的留一交叉验证排序损失（LRL）统计量为

$$\text{LRL} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\{(j,k) \mid \bar{\mathbf{y}}_{i(j)} < \bar{\mathbf{y}}_{i(k)},\ j \in g_i^+,\ k \in g_i^0\}|}{|g_i^+||g_i^0|}, \tag{7}$$

其中，$\bar{\mathbf{y}}_i = \sum_i w_P(\hat{\boldsymbol{\delta}}_i^{LOO})\mathbf{y}_i/Z$ 是由幂参数值和样本外距离预测确定的得分向量，$g_i^+ = \{j:\ \mathbf{g}_{i(j)} = 1\}$ 是一个标签索引集合，指向某个实例i的相关真实标签；相应地，$g_i^0 = \{j:\ \mathbf{g}_{i(j)} = 0\}$ 是一个标签索引集合，指向某个实例 i的不相关真实标签。因此，ML-MLM 训练的目标是求出使 LRL 最小的 $P^* = \text{argmin}_P$。在 ML-MLM 中采用 Allen 的 PRESS 公式，与传统用法不同：传统用法中，每个超参数值对应的均方误差计算都需要重新求解最小二乘问题 [103, 104]。在 ML-MLM 训练中，只需求解一次最小二乘问题即可调整超参数。如果采用朴素的留一交叉验证实现，这一区别会更加显著，因为每个超参数值都需要求解 N 次最小二乘问题。

多标签分类模型的阈值化通常可以采用逐标签阈值化或全局阈值化。在 [105], 作者得出结论，多阈值化方法与单阈值化方法之间没有显著差异。通过将预测的标签基数 [1] 与真实标签基数相匹配，可以直观地选择全局阈值。

找到 P^*后，可以针对 $\mathcal{Y}_t = \mathbb{1}[\bar{\mathcal{Y}} > t]$调整阈值化 t，其中 $\bar{\mathcal{Y}} = \{\bar{\mathbf{y}}_i\}_{i=1}^N$ 使得

$$\textsc{Card}(\mathcal{Y}_t) = \textsc{Card}(\mathcal{Y}), \tag{8}$$

其中 $\textsc{Card}(\cdot)$ 计算给定标签向量集合的标签基数（标签基数指标 CARD 见公式（9））.

4 实验与结果

本节将利用十个基准数据集，对所提出的方法（ML-MLM）、另外三种基于距离回归（DRB）的方法以及 13 种参考方法进行比较。由于我们无法获得所有方法的排序分数，因此基于排序的比较和基于二分划分的比较采用了不同的方法集合。在基于排序的比较中，我们比较三种 DRB 方法和四种参考方法。在基于二分划分的比较中，我们比较全部四种 DRB 方法和 13 种参考方法。新方法的实现可在 https://github.com/jookriha/ml-mlm 获取。在实验中，我们重点将所提出的方法与一种最先进的方法 RF-PCT 进行比较；Bogatinovski 等人 [3] 在一项大规模比较中得出结论，RF-PCT 属于表现最佳的方法之一（见第 2 节）。对于 DRB 方法，我们尤其关注 ML-MLM 与 LLS-MLM 的比较。此外，我们还分析了该方法预测的不确定性及其时间复杂度。第 4.1 节介绍实验设置。第 4.2 节通过四种基于排序的指标和四种基于二分的指标给出实验比较结果。第 4.3 节展示如何评估预测的不确定性。最后，第 4.4 节分析并比较计算成本。

4.1 实验设置

在多标签分类（MLC）中，数据集通常用标签基数和标签密度指标来描述 [1], ，这两个指标概括了每个实例平均有多少个相关标签。标签基数定义为

$$\textsc{Card} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}^T\mathbf{y}_i, \tag{9}$$

其中 $\mathbf{1} \in \{1\}^L$。标签密度定义为

$$\textsc{Dens} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}^T\mathbf{y}_i}{L}. \tag{10}$$

实验所用的数据集汇总于表1，其中 CARD 和 DENS 指标是根据训练集和测试集标签向量的并集计算得出的。除了这些指标和常见的数据特征外，我们还计算了训练集和测试集并集中唯一标签向量的数量，记为L_u，以及测试集中不同标签组合（#新标签向量）的数量，记为 L_n。所有数据集均可从 http://mulan.sourceforge.net/datasets-mlc.html（Mulan 仓库）获取。

为了理解与多标签分类问题相关的实证研究，需要了解相关背景，并选择和理解用于评估和比较不同方法的质量指标。显然，多标签分类问题本身具有多准则性质，因为我们希望获得最简单的模型，同时它的训练速度最快，并且在所有可能的指标上都能取得最佳表现。这一领域已得到全面

评估， [106], 我们和其他人的实验中使用的最相关指标也已汇总于附录 B。指标的实现可在 https://github.com/jookriha/ml-mlm 获取。
由于我们将结果与 [26] 中基于排序的指标结果进行比较，因此选取的数据集相同，但排除了Bookmarks数据集，因为该数据集缺少部分参考结果。此外，对于该数据集，使用几乎完整的参考点集训练 BR-MLM 的成本会非常高。请注意，Scene数据集接近单标签分类问题，因为其标签基数为 1。074。因此，该数据集对于多标签分类比较的相关性存在一定疑问。
为了进行比较，我们计算了以下四种 DRB 方法的结果：

• **BR-MLM** 使用三次方程多标签多边测量（C-MLM） [34] 作为BR方法的基模型。在C-MLM中，逐标签多边测量问题采用闭式解。每个C-MLM模型都会直接输出用于标签排序和分类的实值分数。由于我们采用局部RCut阈值化，因此该方法无需调整超参数。本文提出了这三个要素的整合方案。

• **LLS-MLM**通过定位线性系统（LLS） [32] 方法求解 MLM 的多边定位问题，其中预测距离、基准锚节点（BAN）和输出空间中的参考点（BAN 除外）共同构成一个线性方程组。线性方程组的解经由 BAN 转移后，为所有标签输出实值分数。这些得分直接用于标签排序和分类。我们将与最小预测距离对应的参考点作为 BAN。由于该方法也采用局部 RCut 阈值化，因此无需进行超参数调优。本文提出了这三项要素的整合方案。

• **ML-MLM** 是本文的主要提案。该方法采用反距离加权（IDW） [28] 方法，对训练标签向量集合进行凸组合。两个超参数——幂参数和全局阈值化——都可以直接选定。样本外距离预测可根据 PRESS 统计量 [102]的相关思想，通过闭式解求得。这些由闭式解得到的预测使一次性训练过程成为可能，并可使用 LRL 统计量来选择幂参数和基于标签基数的全局阈值参数。

• **NN-MLM** 根据对应参考点直接预测一组标签，该参考点具有最小的预测距离。该方法在 [33]中提出。请注意，该方法是将原始 NN-MLM [34]直接改编为多标签设置的结果。因此，我们在此沿用其在单标签领域中的名称。主要区别在于，对于多标签分类（MLC）中的 NN-MLM，在最优求解多边定位问题方面并不存在理论保证（命题1）。

对于所有这四种 DRB 方法，我们使用相同的一组输入空间参考点 $\mathbf{R}$。我们选择所有唯一的训练实例作为参考点 $\mathbf{R}$。我们保留了完整的输出参考点集合，因为奇异性问题仅涉及矩阵 $\mathbf{D_x}^T\mathbf{D_x}$ 的逆矩阵。在 6 中，我们将 α 参数设定为输入空间参考点两两距离的较低 1000 分位数，并排除了参考点到其自身的距离（所有两两距离均非零）。我们观察到，这种启发式方法提高了 $\mathbf{D_x}^T\mathbf{D_x}$ 逆矩阵的唯一求解能力。对于 LLS-MLM、ML-MLM 和 NN-MLM，我们始终使用相同的距离回归模型。

在 LLS-MLM 预测算法中（ [32], 见算法1），我们选择基准锚节点（BAN）作为输出空间参考点，其对应的预测距离最小。我们针对 BAN 选择进行了广泛的实验研究。结果发现，在多标签分类问题中，所提出的选择策略优于最初提出的基于随机的选择。因此，在我们的实验中，LLS-MLM 预测具有确定性。由此，这里使用的四种 DRB 方法在训练和预测方面完全确定。对于 BR-MLM，我们对每个单标签分类模型使用 C-MLM [34]，因为无法从最近邻 MLM [34] 方法中获得标签评分/排序。对于 ML-MLM，我们使用留一交叉验证排序损失统计量来调整幂参数。我们对排序损失值随幂参数变化的初步探索显示，较低端取值的变化率最大。因此，我们从以 2 为底的 2^s 值中搜索幂参数取值，其中 $s = 0$，0。1，0。2，.....，8。

对于 BR-MLM 和 LLS-MLM 方法，我们发现采用固定的 0.5 阈值化时，分类结果并不理想。因此，我们提出了一种新的阈值方法，称为“局部排序截断”（局部RCut）：将得分最高的 K_{cut} 个标签标记为相关。与局部RCut中的 RCut阈值处理不同， [105],处的值 K_{cut} 由 NN-MLM 模型预测标签集的基数决定。对于 BR-MLM 和 LLS-MLM 方法，我们省略了基于标签基数的阈值处理，因为这会显著增加这些方法的训练复杂度。对于 BR-MLM 和 LLS-MLM 方法，局部RCut在训练过程中无需进行任何超参数调优。

表1：数据集的特征。表格顶行中， N 是训练集中实例的数量， N_{ts} 是测试集中实例的数量， M 是特征的数量， L 是标签的数量， Q 是测试集中新标签的数量，CARD 是标签基数，DENS 是标签密度， L_u 是唯一标签向量的数量， L_n 是测试集中新标签向量的数量。

数据集	N	N_{ts}	M	L	Q	CARD	DENS	L_u	L_n
Medical	333	645	1449	45	7	1.245	0.028	94	54
Emotions	391	202	72	6	0	1.869	0.311	27	4
Enron	1123	579	1001	53	1	3.378	0.064	753	225
Scene	1211	1196	294	6	0	1.074	0.179	15	1
Yeast	1500	917	103	14	0	4.237	0.303	198	38
Corel5k	4500	500	499	374	3	3.522	0.009	3175	252
Bibtex	4880	2515	1836	159	0	2.402	0.015	2856	836
Delicious	12920	3185	500	983	0	19.020	0.019	15806	3095
Tmc2007	21519	7077	500	22	0	2.220	0.101	1172	0
Mediamill	30993	12914	120	101	0	4.376	0.043	6555	2223

对于 ML-MLM，我们采用了第3.3节所述的基于标签基数的全局阈值化。对于 ML-MLM，可以直接利用现有的样本外距离预测进行基于基数的阈值处理。此外，排序损失和覆盖率指标上的排序性能表明，ML-MLM 的阈值化应经过调优，使其目标是在排序中比 BR-MLM 和 LLS-MLM 更深入地检索标签。我们通过实验比较了 ML-MLM 中基于基数的阈值处理和局部RCut阈值化所得到的预测标签集的标签基数，发现局部RCut阈值化的平均预测标签基数较低。因此，与基于基数的阈值处理相比，局部RCut是一种更保守的方法。

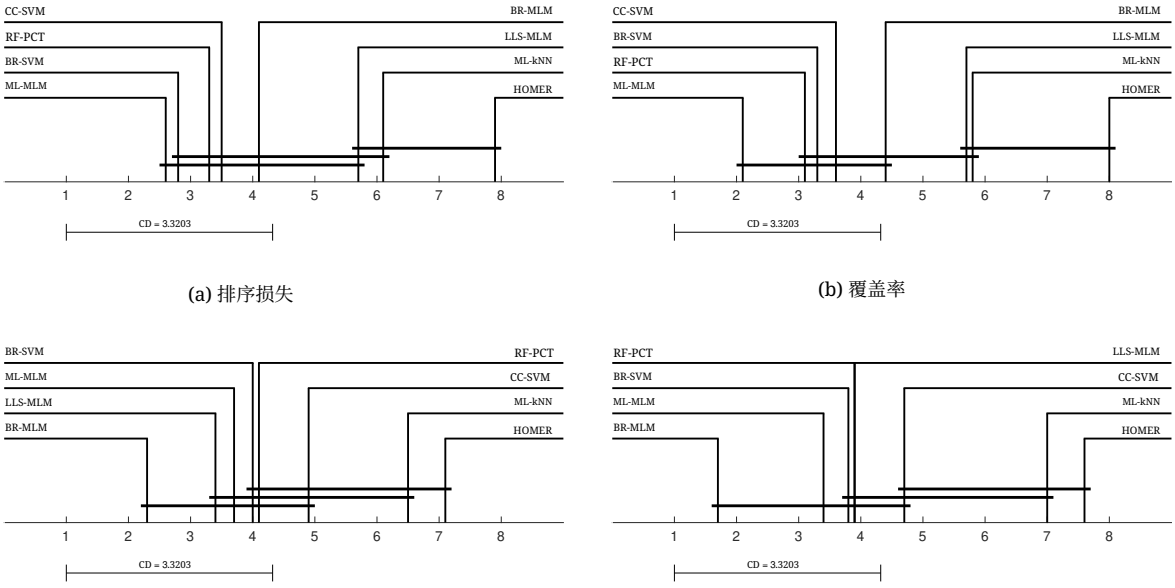
按照 [26, 3], 中的建议，我们将所提出的方法与二元相关性 [1, 107], 、分类器链 [40, 108], 、HOMER [109], 、RAKEL [1],、BPNN [45, 110] 和 RF-PCT [29]进行比较。值得注意的是，在两项研究中，RF-PCT 方法都被认为是表现最佳的方法之一 [26, 3] ，而在 [3] 中，AA 方法中的 BPNN 紧随其后。此外，我们还纳入了 ML-kNN[30] 方法进行比较，因为它与多标签分类方法分类体系 [1, 3]中提出的方法属于同一类别。

按照 [26, 3] 中关于选择基模型的建议，我们采用包括支持向量机和 J48 决策树学习器 [111] 在内的不同基模型来构建二元相关性和分类器链；下文分别称为 BR-SVM、BR-J48、CC-SVM 和 CC-J48。为扩展二元相关性的实验结果，我们还纳入了另外两种复杂基模型，即 AdaBoost（BR-Ada）和决策树随机森林（BR-RF）。遵循相同原则，Rakel 的实验结果分别采用支持向量机（Rakel-SVC）和随机梯度下降（Rakel-SGDC）得到。

我们使用了 [26] 中 BR-SVM、CC-SCM、ML-kNN、HOMER 和 RF-PCT 的结果。对于 BR-J48、CC-J48、BR-Ada、BR-RF、Rakel-SCV、Rakel-SGDC、ML-kNN (d)、ECC-J48、EBR-J48 和 BP-NN，我们在 Python 环境中使用 scikit-multilearn 库的 MEKA 封装器（http://scikit.ml/userguide.html）计算分类结果。对于后一组分类器，我们使用了 MEKA 推荐的默认参数。请注意，我们有来自 [26] 的 ML-kNN 结果，以及基于 MEKA 默认参数得到的 ML-kNN 结果。我们用“(d)”标记后者，以表示默认参数。
我们使用给定的数据集训练-测试划分，将我们的结果与 [26] 中给出的结果进行比较。在统计检验中，我们采用 Friedman 检验，随后进行 Nemenyi 事后检验（如果零假设被拒绝） [112]。统计显著性水平设为 0.05。

4.2 结果

本节展示用于比较的实验结果。我们分别分析基于排序的指标和基于二分的指标的结果。附录D和附录E分别针对基于排序的结果和基于二分的结果，提供了八个表格，详细列出了每种方法、数据集和评估指标的结果。这些表格的结果通过图2和图3中的Nemeneyi图进行了可视化。



(a) 排序损失

(b) 覆盖率

(c) 单一错误

(d) 平均精度

图 2：基于排序的指标结果。

4.2.1 基于排序的指标

对于全部四个基于排序的指标，Friedman零假设均以明显优势被拒绝。因此，我们对全部四个结果表（附录D中的表3至表6）进行了Nemenyi事后检验。这些检验的统计结果通过图2中的Nemenyi图呈现，其中横轴表示某种方法在给定指标下的平均排名。排名越低，表示性能越好。坐标轴上方的水平连接线表示，根据Nemenyi事后检验，连接线所覆盖的方法之间不存在具有统计显著性的差异。

从图2c和图2d可以看出，在排序损失和覆盖率指标方面，ML-MLM的平均排名最高，且与ML-kNN和HOMER之间的差异具有统计显著性。此外，在覆盖率方面，ML-MLM与LLS-MLM之间也存在具有统计显著性的差异。在单一错误和平均精度方面，BR-MLM的排名最高。在这两项指标上，BR-MLM和ML-MLM都属于第一组，因此彼此之间没有具有统计显著性的差异。此外，对于这两项基于排序的指标，ML-MLM与ML-kNN和HOMER之间的差异具有统计显著性。

根据表3—表6以及表1中的数据特征，可以得出一些观察结果。对于Tmc2007数据集，所有DRB方法在所有基于排序的指标上都取得了近乎完美的得分。表1显示，该数据集具有 $L_n = 0$，这意味着测试集中的所有标签向量都曾出现在训练集中。此外，考虑比率 L_n/N_{ts}最低的四个数据集：Tmc2007（0）、Scene（0.001）、Emotions（0.020）和Yeast（0.041）。对于这些数据集，在所有基于排序的指标上，ML-MLM的表现都优于最先进水平的RF-PCT方法。在 L_n/N_{ts} 比率较高的一端也可以观察到这一趋势。例如，在Delicious（0.972）和Enron（0.389）数据集上，RF-PCT的排序性能优于ML-MLM。

根据结果，ML-MLM 在覆盖率方面的性能最强。仅根据平均排名，ML-MLM 与 RF-PCT 之间的平均排名差异在覆盖率方面最大。在该指标上表现良好表明，如果正确设置 ML-MLM 的阈值化，至少在准确率指标的基于二分划分的比较中，它应当能够使 ML-MLM 占据优势。

总而言之，对于基于排序的指标，ML-MLM 与 RF-PCT 之间不存在具有统计显著性的差异。然而，在每项指标上，ML-MLM 的平均排名均优于 RF-PCT。

4.2.2 基于二分的指标

所有基于二分的结果也以明显的优势拒绝了Friedman零假设检验。因此，与第 4.2.1 节类似，对表7至表10进行了Nemenyi事后检验程序。这些检验的结果

再次通过图 3 中给出的临界图可视化结果进行总结。请注意，与第 4.2.1 节不同，此次比较纳入了 NN-MLM、BR-J48、CC-J48、BR-Ada、BR-RF、Rakel-SCV、Rakel-SGDC、ML-kNN (d)、ECC-J48、EBR-J48 和 BP-NN，因此共有 17 种方法，而此前的比较中有 8 种方法。因此，在本节的统计检验中，临界距离更大。

根据图 3，在准确率、微平均 F1 和宏平均 F1 这几项指标上，ML-MLM 的平均排名最佳。此外，在这三项指标上，ML-MLM 与 RF-PCT 之间的排名差异具有统计显著性。在汉明损失方面，BR-MLM 的表现最佳，且与 ML-kNN(d)、EBR-J48、Rakel-SGDC、ECC-J48 和 BPNN 相比具有统计显著性。尽管在汉明损失方面，ML-MLM 的平均排名不及 BR-MLM 和 RF-PCT，但这种差异不具有统计显著性。请注意，DRB 方法之间不存在具有统计显著性的差异。CC 的基于集成的构造（ECC-J48）排名低于原始分类器链。此外，在准确率、微平均 F1 和宏平均 F1 这些指标上，集成方法组中的 ECC-J48、EBR-J48 和 HOMER 的排名略低于基于 DBR 的方法。请注意，在这些指标上，集成方法与基于 DBR 的方法处于同一统计组。ML-kNN 和 BPNN 是基于二分划分的比较中表现最差的方法。因此，BPNN 的现成实现并不适用于小到中等规模的数据集。与一般的神经网络一样，BPNN 需要充足的数据，并且需要仔细微调参数。

正如预期，凭借较高的覆盖率表现，ML-MLM 的准确率优于其他方法。从统计角度来看，ML-MLM 与 ECC-J48、EBR-J48 和 HOMER 这一组方法之间的差异，在准确率、宏平均 F1 和微平均 F1 上相同。不过，若仅考察该组方法与 ML-MLM 的平均排名差异，可以发现准确率方面的差异更大。此外，从表 7 和表 9 可以看出，在十个数据集中，ML-MLM 有五个数据集的准确率最高，六个数据集的微平均 F1 最高。此外，逐标签二分指标——微平均 F1 和宏平均 F1——的结果表明，ML-MLM 输出的预测标签集准确且对所有标签的表现均衡。

ML-MLM 的一个小缺点是其汉明损失表现：其平均排名逊于所有其他 DRB 方法、BR-RF、BR-SVM 和 RF-PCT。由于 ML-MLM 的准确率更高而汉明损失更差，它在选择标签时可能比其他方法更为激进，从而增加假正例的数量。不过，鉴于其平均排名与表现更好的方法在统计上没有显著差异，且优于 Rakel-SVC、ML-kNN、HOMER、ML-kNN(d)、EBR-J48、Rakel-SGDC、ECC-J48 和 BPNN，我们认为这种权衡是合理的。值得注意的是，在许多数据集中，所有受测方法的汉明损失绝对值都很小。这主要是因为该指标按所有可能标签的数量进行了归一化。

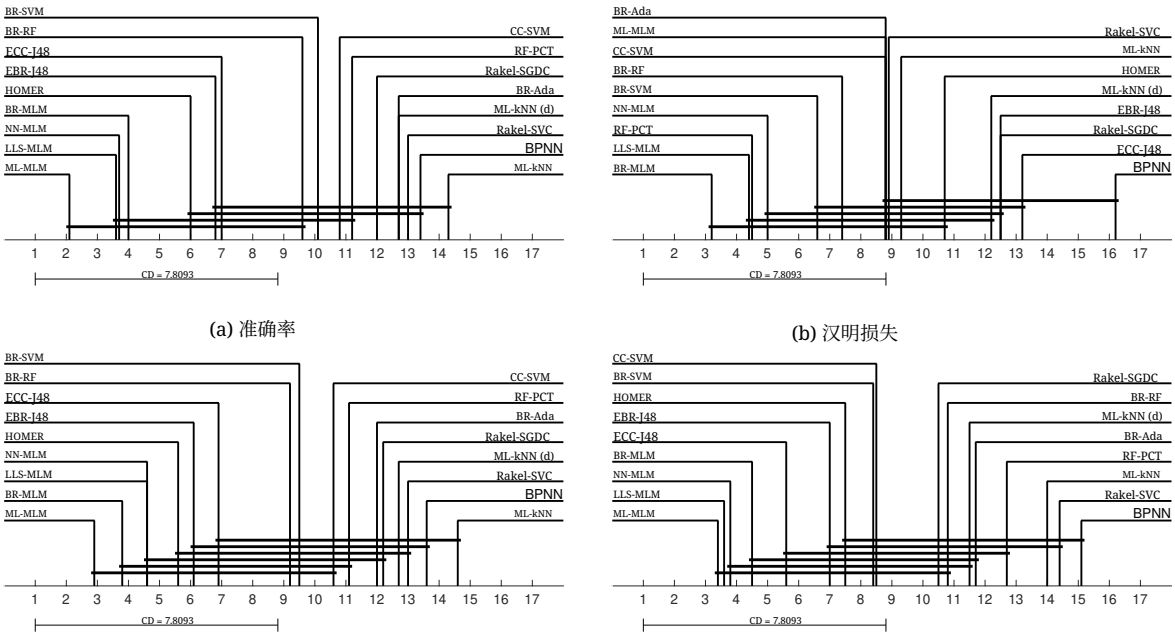
总体而言，结果表明，在四项基于二分的指标中，ML-MLM有三项在统计上优于最先进水平的方法RF-PCT，而在汉明损失方面与其统计上持平。此外，在所有基于二分的指标上，ML-MLM的排名均优于集成方法（ECC-J48、EBR-J48和HOMER），但这种差异未被判定为具有统计显著性。所有DRB方法在比较中表现良好，最简单的方法（NN-MLM）也表现出色：在三项指标上与RF-PCT统计上持平，在一项指标上优于RF-PCT。

4.3 评估ML-MLM的不确定性并对其进行解释

通过所选幂参数的取值和预测距离，可以理解ML-MLM如何进行预测。结果汇总于表2：对于每个数据集，测试集中与预测的最近邻对应的距离经过平方后取平均值。为便于解释标签差异，我们对表2中的预测距离进行了平方（详情见附录C）。图4通过箱线图展示了预测距离（未平方）的分布。

从表2可以看出，不同数据集所选取的幂参数值差异显著。Scene数据集的幂参数值最小（$P = 4$），而Medical数据集的最大（$P \approx 222.9$）。对于后者，ML-MLM预测的标签集合与NN-MLM高度相似，因为几乎所有权重都分配给了预测的最近邻。我们通过比较Medical数据集上NN-MLM与采用局部RCut阈值化的ML-MLM的基于二分划分的指标结果，验证了这一点。两者的比较结果完全相同。表7至表10中关于Medical数据集以及ML-MLM和NN-MLM方法的结果有所不同，这是因为ML-MLM采用了基于基数的阈值化（式8），而非局部RCut阈值化。对于Medical数据集，NN-MLM在四项指标中的三项上表现更好，这表明当幂参数值较大时，基于基数的阈值化可能不是最优选择。

从表2和数据集的特征（表1）中，我们可以注意到，对于类型差异很大的数据集，例如 Bibtex（$P = 128$）和 Tmc2007（$P = 59.7$），ML-MLM 可能会选择较大的幂参数值。然而，如果结合预测距离和幂参数值中的 Information，我们可以对 Tmc2007 作出如下解释：



(a) 准确率

(b) 汉明损失

(c) 微平均 F1

(d) 宏平均 F1

图 3：基于二分的指标结果。

由于最近的预测邻居已经包含了近乎最优预测所需的大量 Information，因此增加凸组合中其他参考点的贡献并不能使 ML-MLM 获益太多。由于 Bibtex 的预测距离和幂参数都较大，ML-MLM 主要是在进行外推，并不能通过组合多个参考点的标签信息而获得太多益处。

关于表2中的预测距离，以及 ML-MLM 的总体排序结果（表3—表6）和分类结果（表7—表10），我们可以总结如下。对于给定的实例，预测距离可直接用作不确定性估计，用以衡量排序或分类的可靠程度。距离越大，说明该实例可能离数据较远，模型正在进行外推。因此，预测可能不可靠。这样一来，我们便可根据预测距离，直接将预测划分为不同的不确定性类别
（例如，低： $distance < 1$；中： $1 < distance < \sqrt[-]{2}$；高： $distance > \sqrt[-]{2}$）。此外，对于给定的数据集，我们可以使用预测距离平方，以直观的方式表征机器学习问题的难度。

表2：ML-MLM 的选定幂参数 P取值及平均最小预测欧几里得距离平方。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
P	$2^{7.8}$	$2^{2.9}$	$2^{3.8}$	2^{2}	$2^{3.2}$	$2^{4.1}$	2^{7}	$2^{4.2}$	$2^{5.9}$	$2^{4.4}$
	≈ 222.9	≈ 7.5	≈ 13.9	≈ 4	≈ 9.2	≈ 17.1	≈ 128	≈ 18.4	≈ 59.7	≈ 21.1
距离	0.6 ± 0.4	0.8 ± 0.4	1.7 ± 1.1	0.3 ± 0.3	2.1 ± 0.9	3.7 ± 0.5	1.9 ± 0.9	16.6 ± 5.4	0.2 ± 0.2	2.4 ± 1.1

4.4 距离回归方法的时间复杂度

使用 Moore-Penrose 伪逆的距离回归步骤具有 $\mathcal{O}(NK^2)$ 时间复杂度。由于我们选择几乎所有不同的点作为参考点，对于具有 N 个不同输入实例的数据集，LLS-MLM、ML-MLM 和 NN-MLM 的时间复杂度为 $\mathcal{O}(N^3)$。在相同假设下，BR-MLM 的训练时间复杂度为$\mathcal{O}(LN^3)$，因为每个标签都需要一个不同的 Moore-Penrose 伪逆。因此，对于 DBR 方法，由于 L 依赖关系，BR-MLM 方法的训练时间复杂度最高。接下来，我们将基于输入实例互不相同的假设分析预测时间复杂度，即 $K = N$。就时间复杂度而言，这一假设对应最坏情况。

LLS 的线性系统可以通过 OLS 解法求解，其代价为 $\mathcal{O}(L^2N)$ [32]。因此，当标签数量较多时，这项代价会主导整体时间复杂度，在多标签分类问题中必须予以考虑。与之类似

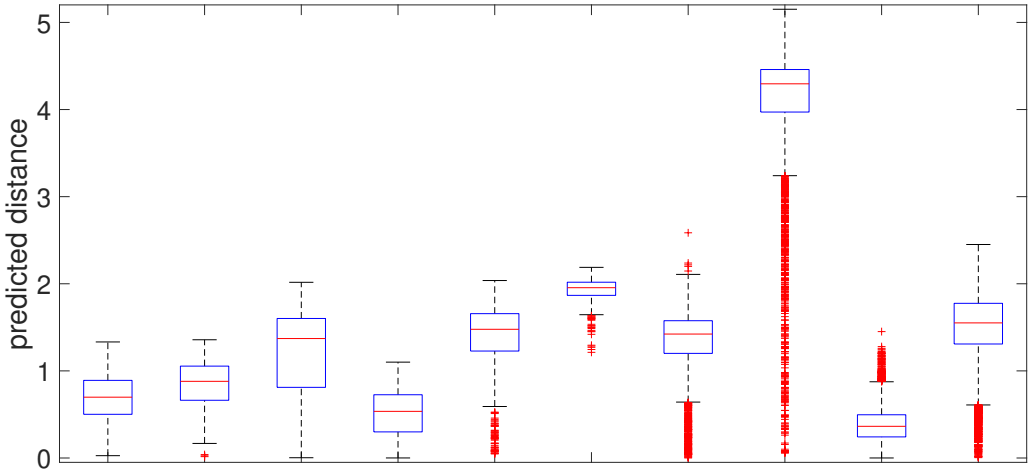


图 4：测试集的距离回归预测最小距离。预测距离越接近零，表明距离回归模型的插值程度越高。距离回归模型的整体插值能力在 Tmc2007 上最强，在 Delicious 上最弱。

于 ML-MLM，LLS 预测标签空间距离的代价为 $\mathcal{O}(N^2)$。因此，LLS 预测的计算代价为 $\mathcal{O}(L^2N + N^2)$。由此可见，在多标签分类场景中，就标签数量而言，ML-MLM 的计算代价低于 LLS。

对于 DBR 方法，NN-MLM 的预测速度最快，因为给定距离预测后，可以以 $\mathcal{O}(L_u)$ 的代价搜索预测集标签，其中 L_u 是训练集中不同标签向量的数量。然而，由于距离预测步骤，NN-MLM 的整体计算复杂度仍为 $\mathcal{O}(N^2)$。

与 NN-MLM 类似，ML-MLM 的预测时间复杂度由距离预测步骤决定。因此，ML-MLM 的时间复杂度也是 $\mathcal{O}(N^2)$。步骤 2（算法2）的朴素实现需要 $\mathcal{O}(KN)$ 的开销。然而，与 NN-MLM 类似，该开销可降至 $\mathcal{O}(C_uL_u)$，其中 L_u 是训练集中的唯一标签向量数量，C_u 是这些标签向量的标签基数。这种对标签基数的依赖源于对标签向量采用了稀疏结构。此外，对于标签基数较小的多标签分类问题，该复杂度可降至 $\mathcal{O}(L_u)$。

除了训练之外，BR-MLM 方法的预测计算成本也最高。BR-MLM 的预测时间复杂度为 $\mathcal{O}(L^2N)$，因为要使用 L 个模型分别估计每个标签的距离预测。

提醒一下，采用基于聚类的距离回归构造方法 [33]可以显著降低这些训练和预测的时间复杂度。此处，我们旨在比较在简化假设下各核心模型的成本。

5 讨论

所有 DRB方法在基于二分划分的指标比较中都表现良好。这一结果表明，所选的阈值方法很适合 DRB方法。本文提出的阈值方法——局部RCut——提供了一种直接的方式，可在预测时根据 NN-MLM 的预测基数自适应地设定阈值。需要注意的是，如果局部RCut对采用相同核心距离回归模型的 DRB 模型进行阈值化，就不会增加预测的复杂度。我们观察到，对于 DRB方法，局部RCut阈值化优于标准的固定 0.5 阈值化。因此，我们建议 DRB方法不要采用这种固定阈值化。这一观察结果与 [105] 中给出的结果和建议一致。为了提升较高幂参数值下 ML-MLM 的分类性能，可以考虑结合全局和局部 RCut 阈值化的混合方法。

根据平均排名进行比较时，所提出的 ML-MLM 方法总体上优于其他 DRB方法。然而，仅就覆盖率指标而言，ML-MLM 与其他 DRB方法之间存在统计学差异。在第3节中，我们指出，NN-MLM 方法的预测并不能保证为多标签分类问题求得多边定位问题的解。然而，LLS-MLM 和 NN-MLM 的分类结果高度相似（图 3 和表7至表10），这表明这种简单的最近邻启发式方法非常接近于像更复杂的 LLS方法那样求解多边定位问题。单从这些结果来看，似乎没有理由使用 LLS-MLM 而非 NN-MLM。不过，NN-MLM 的主要局限在于无法为预测标签提供排序或评分，因此其适用性有限。此外，NN-MLM 的预测仅限于训练集标签向量。而且，尽管 DRB方法通过截然不同的方法预测标签集合，但它们在比较中都表现良好。这表明，DRB 的构建有利于多标签分类问题，并且由此得到的距离预测可以通过多种方式加以利用，从而获得良好的解决方案。

综合考虑排序和分类性能以及训练/预测时间复杂度，DRB方法中最可行的选择是 ML-MLM。正如上一段所述，与 NN-MLM 和 BR-MLM 相比，使用 ML-MLM 有明显优势。与 LLS-MLM 相比，ML-MLM 在 RANKING LOSS 和 COVERATE 指标上的排序性能更优。此外，ML-MLM 的预测阶段时间复杂度与标签数量呈线性关系，而 LLS-MLM 的时间复杂度则呈二次关系。

6 结论

本文提出了一种简单的、基于距离回归的多标签学习方法，称为 ML-MLM。该方法的核心思想是通过距离回归和反距离加权来学习标签排序。除了是多标签分类（MLC）文献中最简单的方法之一之外，我们的实验评估还表明，对于小型到中型的多标签分类问题，该方法在多标签排序和分类性能方面达到了最先进水平。此外，在分类性能评估中，所提出的方法优于一种基于随机森林的方法；该方法被认为是最新多标签分类文献中的最先进方法之一。在排序性能评估中，所提出的方法与基于随机森林的方法表现相当。此外，我们还表明，所提出的方法优于 ML-kNN 方法；在多标签分类分类体系中，该方法被认为与我们的方法属于同一类别。

尽管所提出方法的训练时间复杂度与标签数量无关，但对于基本的乔列斯基分解方法而言，其关于观测值数量的训练时间复杂度是三次的。使用基本算法时，这是所提出方法在可扩展性方面的瓶颈。然而，本工作的主要目标是证明一种基于简单距离回归的方法能够在多标签分类问题中实现较高的准确率。因此，我们没有着手降低求解大规模普通最小二乘法问题所产生的三次训练时间复杂度。此外，尽管训练时间呈三次复杂度，但由于这类问题广受关注且研究历史悠久，已经存在针对它们的成熟实现。例如，可以使用随机近似和面向 GPU计算 的方法 [113, 114] 来提高此步骤的效率。

总体而言，结果显示并进一步证实了先前的观察：利用特征空间与多维输出空间之间的线性距离映射，有利于多输出学习问题。在本文的形式化框架中，所有不同点之间的联系都通过这些距离映射表示。这意味着，我们在特征层面利用不相似性简化学习过程，同时探索数据的整体不相似性结构以进行预测。换言之，我们假设各个特征同等重要，但仍保留这样的可能性：观测值之间的每一种不相似性都可能对预测有所贡献。

显然，如第 2 节所总结的那样，可以通过类似的技术扩展和修改来改进所提出的方法。特别是， [33] 中已经介绍的与 MLM 技术结合的集成技术，可用于针对训练数据的子集引入更具体的模型；这些子集可以根据特征（输入）或标签（输出），或它们之间的依赖关系进行选择（见第 2 节）。对于全局模型和局部模型，都可以使用特征选择进一步明确模型，使其能够封装多标签数据的相关行为。此外，由于 MLM 的核心构造是距离回归，而距离回归基于输入空间中的成对不相似性，因此还应进一步测试其他数据模态及其对应的不相似性度量。对于文本数据，在 [4, 5, 6, 7, 8, 9, 10], 中用于多标签分类任务时，可以应用许多相似性度量 [115]。图像数据也是如此，在 [11, 12, 13, 14], 中可以使用例如基于图像块的不相似性 [116]进行处理。

作者贡献

JH 开发了所提出的方法，编写了实现代码，设计并开展了实验，分析了结果，并撰写了论文的大部分内容。AH 负责开展并设计了有关 MEKA

封装器实现的实验，并参与撰写实验部分。AHS 构思了理论结果，撰写了证明，并参与撰写方法部分和初稿。TK 检索了文献，撰写了相关工作部分的大部分内容，并参与撰写初稿。CLCM、TK 和 JPPG 修订并审阅了初稿。JPPG 提供了 LLS 算法的参考实现。所有作者都参与撰写和编辑了最终稿。

资助

本研究得到 芬兰科学院通过欧洲高性能计算研究计划资助项目 351579 提供的支持。 e.

声明

作者声明，他们不存在利益冲突。

数据可用性

所用数据集均可公开获取。

代码可用性

所有代码均可在*Gitbub* https://github.com/jookriha/ml-mlm. 获取。

参考文献

[1] Grigorios Tsoumakas 和 Ioannis Katakis。多标签分类：概述。*International Journal ofData Warehousing and Mining (IJDWM)*，3(3):1–13，2007年。[2] Min-Ling Zhang 和 Zhi-Hua Zhou。多标签学习算法综述。*IEEE TransactionsonKnowledgeandDataEngineering*，26(8):1819–1837，2013年。[3] Jasmin Bogatinovski、Ljupˇco Todorovski、Sašo Džeroski 和 Dragi Kocev。多标签分类方法的全面比较研究。*Expert Systems with Applications*，203:117215，2022年。[4] Shuhua Monica Liu 和 Jiun-Hung Chen。一种基于多标签分类的情感分类方法。*ExpertSystems with Applications*，42(3):1083–1093，2015年。[5] Mingchu Jiang、Zhisong Pan 和 Na Li。使用 l21 范数最小化极限学习机进行多标签文本分类。*Neurocomputing*，261:4–10，2017年。[6] Alex M.G. Almeida、Ricardo Cerri、Emerson Cabrera Paraiso、Rafael Gomes Mantovani 和 Sylvio Barbon Junior。将多标签技术应用于短文本情绪识别。*Neurocomputing*，320:35–46，2018年。[7] Md Aslam Parwez、Muhammad Abulaish 等。使用卷积神经网络进行微博文本的多标签分类。*IEEE Access*，7:68678–68691，2019年。[8] Jingcheng Du、Qingyu Chen、Yifan Peng、Yang Xiang、Cui Tao 和 Zhiyong Lu。ML-Net：使用深度神经网络对生物医学文本进行多标签分类。*Journal of the American Medical InformaticsAssociation*，26(11):1279–1285，2019年。[9] Dong Zhang、Shu Zhao、Zhen Duan、Jie Chen、Yanping Zhang 和 Jie Tang。使用层次化透明表征的论文审稿人推荐多标签分类方法。*ACMTransactions onInformation Systems (TOIS)*，38(1):1–20，2020年。[10] Qingyu Chen、Alexis Allot、Robert Leaman、Rezarta Islamaj、Jingcheng Du、Li Fang、Kai Wang、Shuo Xu、Yuefu Zhang、Parsa Bagherzadeh 等。生物医学文献的多标签分类：BioCreative VII LitCovid 赛道中用于 COVID-19 文献主题标注的综述。*Database*，2022年，2022年。[11] Haomin Chen、Shun Miao、Daguang Xu、Gregory D Hager 和 Adam P Harrison。胸部X光图像的深度层次多标签分类。载于*Internationalconferenceon medical imaging with deeplearning*，第109–120页。PMLR，2019年。

[12] Haomin Chen、Shun Miao、Daguang Xu、Gregory D Hager 和 Adam P Harrison。应用于胸部X光异常分类体系的深度层次多标签分类。医学图像分析，66:101811，2020年。[13] Xiangji Pan、Kai Jin、Jing Cao、Zhifang Liu、Jian Wu、Kun You、Yifei Lu、Yufeng Xu、Zhaoan Su、Jiekai Jiang 等。基于深度学习，对糖尿病视网膜病变中的视网膜病变进行多标签分类，以自动分析眼底荧光血管造影。Graefe's Archive for Clinical and Experimental Ophthalmology，258(4):779–785，2020年。[14] Lei Bi、David Dagan Feng、Michael Fulham 和 Jinman Kim。通过超连接卷积神经网络对多模态皮肤病变进行多标签分类。Pattern Recognition，107:107502，2020年。[15] Konstantinos Pliakos、Celine Vens 和 Grigorios Tsoumakas。使用多标签分类和标签划分预测药物-靶标相互作用。IEEE/ACM计算生物学与生物信息学汇刊，18(4):1596–1607，2019年。[16] Hiba Chougrad、Hamid Zouaki 和 Omar Alheyane。用于乳腺癌早期诊断的多标签迁移学习。Neurocomputing，392:168–180，2020年。[17] Liang Zhou、Xiaoyuan Zheng、Di Yang、Ying Wang、Xuesong Bai 和 Xinhua Ye。多标签分类模型在糖尿病并发症诊断中的应用。BMC医学信息学与决策制定，21(1):1–10，2021年。[18] Israel Elujide、Stephen G Fashoto、Bunmi Fashoto、Elliot Mbunge、Sakinat O Folorunso 和 Jeremiah O Olamijuwon。深度学习和机器学习技术在精神病性障碍多标签分类性能上的应用。Informatics in Medicine Unlocked，23:100545，2021年。[19] Junxian Cai、Weiwei Sun、Jianfeng Guan 和 Ilsun You。用于心律失常多标签分类的 Multi-ECGnet。IEEE Access，8:110848–110858，2020年。[20] Ajay Kumar、Kumar Abhishek、Amit Kumar Singh、Pranav Nerurkar、Madhav Chandane、Sunil Bhirud、Dhiren Patel 和 Yann Busnel。遥感卫星图像的多标签分类。新兴电信技术汇刊，32(7):e3988，2021年。[21] Shuoyao Wang、Suzhi Bi 和 Ying-Jun Angela Zhang。智能电网中虚假数据注入攻击的定位检测：一种多标签分类方法。IEEE物联网期刊，7(9):8218–8227，2020年。[22] Zhanquan Sun、Chaoli Wang、Yangyang Zhao 和 Chao Yan。基于集成分类器的心电图信号多标签分类。IEEE Access，8:117986–117996，2020年。[23] Ding Li 和 Scott Dick。使用多标签分类方法进行非侵入式负荷监测。电气工程，103(1):607–619，2021年。[24] Michael R Maser、Alexander Y Cui、Serim Ryou、Travis J DeLano、Yisong Yue 和 Sarah E Reisman。用于预测交叉偶联反应条件的多标签分类模型。化学信息与建模杂志，61(1):156–166，2021年。[25] Kushagra Saini 和 Venkatnarayan Ramanathan。从分子结构预测气味：一种多标签分类方法。Scientific Reports，12(1):1–11，2022年。[26] Gjorgji Madjarov、Dragi Kocev、Dejan Gjorgjevikj 和 Sašo Džeroski。多标签学习方法的广泛实验比较。Pattern Recognition，45(9):3084–3104，2012年。[27] Amauri Holanda de Souza Junior、Francesco Corona、Guilherme A. Barreto、Yoan Miche 和 Amaury Lendasse。最小学习机：一种用于回归和分类的新型监督式距离方法。Neurocomputing，164:34–44，2015年。[28] Donald Shepard。一种用于不规则间距数据的二维插值函数。载于1968年ACM第23届全国会议论文集，第517–524页，1968年。[29] Dragi Kocev、Celine Vens、Jan Struyf 和 Sašo Džeroski。多目标决策树集成。载于欧洲机器学习会议，第624–631页。施普林格，2007年。[30] Min-Ling Zhang 和 Zhi-Hua Zhou。ML-kNN：一种用于多标签学习的惰性学习方法。Pattern Recognition，40(7):2038–2048，2007年。[31] Alberto Bemporad。通过反距离加权和径向基函数进行全局优化。多媒体工具与应用，77(2):571–595，2020年11月。[32] Joonas Hämäläinen、Alisson SC Alencar、Tommi Kärkkäinen、César LC Mattos、Amauri H Souza Júnior 和 Joao PP Gomes。最小学习机：理论结果与基于聚类的参考点选择。The Journal of Machine Learning Research，21，2020年。

[33] Joonas Hämäläinen、Paavo Nieminen 和 Tommi Kärkkäinen。通过多目标距离回归实现基于实例的多标签分类。载于《欧洲人工神经网络、计算智能与机器学习研讨会》。ESANN，2021年。[34] Diego P. P. Mesquita、João P. P. Gomes 和 Amauri H. Souza Junior。用于分类和回归的高效最小学习机集成。Neural Processing Letters，第 1–16 页，2017年。[35] Matthew R Boutell、Jiebo Luo、Xipeng Shen 和 Christopher M Brown。学习多标签场景分类。Pattern Recognition，37(9)：1757–1771，2004年。[36] Grigorios Tsoumakas、Ioannis Katakis 和 Ioannis Vlahavas。用于多标签分类的随机 k 标签集。IEEE Transactions on Knowledge and Data Engineering，23(7)：1079–1089，2010年。[37] Jesse Read、Peter Reutemann、Bernhard Pfahringer 和 Geoffrey Holmes。MEKA：Weka 的多标签/多目标扩展。The Journal of Machine Learning Research，17(1)：1–5，2016年。[38] Piotr Szymanski 和 Tomasz Kajdanowicz。Scikit-multilearn：用于执行多标签分类的基于 scikit 的 Python 环境。The Journal of Machine Learning Research，20(1)：209–230，2019年。[39] Marcel Wever、Alexander Tornede、Felix Mohr 和 Eyke Hüllermeier。用于多标签分类的自动化机器学习：综述与实证评估。IEEE Transactions on Pattern Analysis and Machine Intelligence，43(9)：3037–3054，2021年。[40] Jesse Read、Bernhard Pfahringer、Geoff Holmes 和 Eibe Frank。用于多标签分类的分类器链。Machine Learning，85(3)：333–359，2011年。[41] André Elisseeff 和 Jason Weston。用于多标签分类的核方法。Advances in Neural Information Processing Systems，14，2001年。[42] Isaac Triguero 和 Celine Vens。层次多标签分类技术的标注策略。Pattern Recognition，56：170–183，2016年。[43] Qingyao Wu、Mingkui Tan、Hengjie Song、Jian Chen 和 Michael K Ng。ML-FOREST：用于多标签分类的多标签树集成方法。IEEE Transactions on Knowledge and Data Engineering，28(10)：2665–2680，2016年。[44] Ignazio Pillai、Giorgio Fumera 和 Fabio Roli。带拒绝选项的多标签分类。Pattern Recognition，46(8)：2256–2266，2013年。[45] Min-Ling Zhang 和 Zhi-Hua Zhou。应用于功能基因组学和文本分类的多标签神经网络。IEEE transactions on Knowledge and Data Engineering，18(10)：1338–1351，2006年。[46] Grigorios Tsoumakas、Ioannis Katakis 和 Ioannis Vlahavas。标签数量众多领域中的有效且高效的多标签分类。载于 ECML/PKDD 2008 多维数据挖掘研讨会（MMD'08）论文集，第 21 卷，第 53–59 页，2008年。[47] Eleftherios Spyromitros-Xioufis、Grigorios Tsoumakas、William Groves 和 Ioannis Vlahavas。通过输入空间扩展实现多目标回归：将目标作为输入。Machine Learning，104(1)：55–98，2016年。[48] Xie Jun、Yu Lu、Zhu Lei 和 Duan Guolun。基于条件熵的用于多标签分类的分类器链。Neurocomputing，335：185–194，2019年。[49] Lior Rokach、Alon Schclar 和 Ehud Itach。多标签分类的集成方法。Expert Systems with Applications，41(16)：7507–7523，2014年。[50] Ran Wang、Sam Kwong、Xu Wang 和 Yuheng Jia。用于多标签分类的主动式 k 标签集成。Pattern Recognition，109：107583，2021年。[51] Dragi Kocev、Celine Vens、Jan Struyf 和 Sašo Džeroski。用于预测结构化输出的树集成。Pattern Recognition，46(3)：817–833，2013年。[52] Rafael B Pereira、Alexandre Plastino、Bianca Zadrozny 和 Luiz HC Merschmann。多标签分类特征选择方法的分类。Artificial Intelligence Review，49(1)：57–78，2018年。[53] Jianghong Ma、Haijun Zhang 和 Tommy WS Chow。具有标签特定特征与分类器的多标签分类：粗细粒度调优框架。IEEE Transactions on Cybernetics，51(2)：1028–1042，2019年。[54] Wei Weng、Yan-Nan Chen、Chin-Ling Chen、Shun-Xiang Wu 和 Jing-Hua Liu。用于多标签分类的非稀疏标签特定特征选择。Neurocomputing，377：85–94，2020年。[55] Jun Huang、Guorong Li、Qingming Huang 和 Xindong Wu。学习用于多标签分类的标签特定特征和类别依赖标签。IEEE Transactions on Knowledge and Data Engineering，28(12)：3309–3323，2016年。

[56] Vikas Kumar、Arun K Pujari、Vineet Padmanabhan 和 Venkateswara Rao Kagita。用于多标签分类的保持群组标签嵌入。Pattern Recognition，90：23–34，2019年。[57] Ming Huang、Fuzhen Zhuang、Xiao Zhang、Xiang Ao、Zhengyu Niu、Min-Ling Zhang 和 Qing He。用于多标签分类的监督式表示学习。Machine Learning，108(5)：747–763，2019年。[58] Wissam Siblini、Pascale Kuntz 和 Frank Meyer。多标签分类降维综述。IEEE Transactions on Knowledge and Data Engineering，33(3)：839–857，2019年。[59] Guoqiang Wu、Ruobing Zheng、Yingjie Tian 和 Dalian Liu。结合稳健低秩学习的联合排序支持向量机与二元关联，用于多标签分类。Neural Networks，122：24–39，2020年。[60] Xiaoyan Zhu、Jiaxuan Li、Jingtao Ren、Jiayin Wang 和 Guangtao Wang。用于多标签分类的动态集成学习。Information Sciences，623：94–111，2023年。[61] Yuelong Xia、Ke Chen 和 Yun Yang。采用加权分类器选择与堆叠集成的多标签分类。Information Sciences，557：421–442，2021年。[62] Joonas Hämäläinen 和 Tommi Kärkkäinen。结合基于距离的学习的多目标回归问题转换方法。ESANN，2020年。[63] Tommi Kärkkäinen。极限最小学习机：基于距离的基函数岭回归。Neurocomputing，342：33–48，2019年。[64] Vikas Chauhan 和 Aruna Tiwari。用于多标签分类的随机神经网络。Applied Soft Computing，115：108184，2022年。[65] Joakim Linja、Joonas Hämäläinen、Paavo Nieminen 和 Tommi Kärkkäinen。基于距离的回归特征选择：综述与单次包装法。Neurocomputing，2022年。[66] Jaesung Lee 和 Dae-Won Kim。使用多变量互信息进行多标签分类的特征选择。Pattern Recognition Letters，34(3)：349–357，2013年。[67] Wenbin Qian、Chuanzhen Xiong 和 Yinglong Wang。基于模糊相对可辨识性的多标签分类排序式特征选择。Applied Soft Computing，102：106995，2021年。[68] Zhi-Fen He、Ming Yang、Yang Gao、Hui-Dong Liu 和 Yilong Yin。缺失标签与特征选择下的联合多标签分类及标签相关性。Knowledge-Based Systems，163：145–158，2019年。[69] Lin Sun、Tianxiang Wang、Weiping Ding、Jiucheng Xu 和 Yaojin Lin。结合 Fisher 分数与多标签邻域粗糙集的多标签分类特征选择。Information Sciences，578：887–912，2021年。[70] Jianghong Ma 和 Tommy WS Chow。多标签分类中的基于主题的实例与特征选择。IEEE Transactions on Neural Networks and Learning Systems，33(1)：315–329，2022年。[71] Jianghong Ma、Bernard Chi Yuen Chiu 和 Tommy WS Chow。基于群组映射的多标签分类：一种结合局部特征选择与局部标签相关性的框架。IEEE Transactions on Cybernetics，52(6)：4596–4610，2022年。[72] Newton Spolaôr、Maria Carolina Monard、Grigorios Tsoumakas 和 Huei Diana Lee。多标签特征选择的系统综述及一种基于标签构造的新方法。Neurocomputing，180：3–15，2016年。[73] Shima Kashef、Hossein Nezamabadi-pour 和 Bahareh Nikpour。多标签特征选择：全面综述与指导性实验。Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery，8(2)：e1240，2018年。[74] Ze-Bang Yu 和 Min-Ling Zhang。采用标签专属特征生成的多标签分类：一种包装式方法。IEEE Transactions on Pattern Analysis and Machine Intelligence，44(9)：5199–5210，2022年。[75] Hamid Bayati、Mohammad Bagher Dowlatshahi 和 Amin Hashemi。Mssl：一种基于模因的多标签分类稀疏子空间学习算法。International Journal of Machine Learning and Cybernetics，13(11)：3607–3624，2022年。[76] Hongbin Dong、Jing Sun、Xiaohang Sun 和 Rui Ding。用于多标签分类的多目标特征选择。Knowledge-Based Systems，208：106456，2020年。[77] Amin Hashemi、Mohammad Bagher Dowlatshahi 和 Hossein Nezamabadi-pour。一种高效的基于帕累托的多标签分类特征选择算法。Information Sciences，581：428–447，2021年。[78] Wenbin Qian、Jintao Huang、Yinglong Wang 和 Yonghong Xie。结合粗糙集的多标签分类标签分布特征选择。International Journal of Approximate Reasoning，128：32–55，2021年。[79] Qiaoyu Tan、Yanming Yu、Guoxian Yu 和 Jun Wang。使用不完整……进行半监督多标签分类

Plete 标签信息。*Neurocom puting*，260:192–202，2017年。

[80] Tien Thanh Nguyen、Thi Thu Thuy Nguyen、Anh Vu Luong、Quoc Viet Hung Nguyen、Alan Wee-Chung Liew 和 Bela Stantic。基于标签相关性与数据流中一阶特征依赖的多标签分类。*PatternRecognition*，90:35–51，2019年。[81] Tien Thanh Nguyen、Manh Truong Dang、Anh Vu Luong、Alan Wee-Chung Liew、Tiancai Liang 和 John McCall。基于演化数据流增量聚类的多标签分类。*PatternRecognition*，95:96–113，2019年。[82] Felipe Kenji Nakano、Ricardo Cerri 和 Celine Vens。层次多标签分类的主动学习。*Data Mining and Knowledge Discovery*，34(5):1496–1530，2020年。[83] Marilyn Bello、Gonzalo Nápoles、Ricardo Sánchez、Rafael Bello 和 Koen Vanhoof。用于提取多标签分类问题中高层特征和标签的深度神经网络。*Neurocomputing*，413:259–270，2020年。[84] Dongjoo Yun、Jongbin Ryu 和 Jongwoo Lim。用于多标签分类的双重聚合特征金字塔网络。*Pattern Recognition Letters*，144:75–81，2021年。[85] Weiwei Liu、Haobo Wang、Xiaobo Shen 和 Ivor Tsang。多标签学习的新兴趋势。*IEEETransactions on Pattern Analysis and MachineIntelligence*，2021年。[86] K. Bhatia、K. Dahiya、H. Jain、P. Kar、A. Mittal、Y. Prabhu 和 M. Varma。极端分类资源库：多标签数据集与代码，2016年。[87] Kush Bhatia、Himanshu Jain、Purushottam Kar、Manik Varma 和 Prateek Jain。用于极端多标签分类的稀疏局部嵌入。*Advances in Neural Information Processing Systems*，28，2015年。[88] Wenjie Zhang、Junchi Yan、Xiangfeng Wang 和 Hongyuan Zha。深度极端多标签学习。载于*2018*年 *ACM* 多媒体检索国际会议论文集，第100–107页，2018年。[89] V. Roshan Joseph 和 Lulu Kang。基于回归的反距离加权及其在计算机实验中的应用。*Technometrics*，53(3):254–265，2011年8月。[90] Feng-Wen Chen 和 Chen-Wuing Liu。利用反距离加权（IDW）估算台湾中部的空间降雨分布。*Paddyand Water Environment*，10(3, SI):209–222，2012年9月。[91] Zhao-Yue Chen、Tian-Hao Zhang、Rong Zhang、Zhong-Min Zhu、Jun Yang、Ping-Yan Chen、Chun-Quan Ou 和 Yuming Guo。利用缺失值填充的卫星数据，通过极端梯度提升模型估算中国 PM2.5 浓度。*Atmospheric Environment*，202:180–189，2019年4月1日。[92] Nazife Cevik。基于动态反距离加权的局部人脸描述符。多媒体工具与应用，79(41-42):31087–31102，2020年11月。[93] Yucheng Shi、Weiguo He、Jian Zhao、Aoyu Hu、Jingna Pan、Haizheng Wang 和 Honglu Zhu。基于反距离加权的期望输出计算及其在分布式光伏电站异常检测中的应用。*Journal of Cleaner Production*，253，2020年4月20日。[94] Gjorgji Madjarov、Dejan Gjorgjevikj 和 Sašo Džeroski。多标签学习的两阶段架构。*PatternRecognition*，45(3):1019–1034，2012年。[95] Chunyu Zhang 和 Zhanshan Li。通过加权和标签熵引导的聚类集成，利用标签特定特征进行多标签学习。*Neurocomputing*，419:59–69，2021年。[96] Xiaoya Wei、Ziwei Yu、Changqing Zhang 和 Qinghua Hu。用于多标签分类的标签特定特征集成。载于*2018*年 *IEEE* 多媒体与博览会国际会议（*ICME*），第1–6页。IEEE，2018年。[97] Michael Rapp、Eneldo Loza Mencía、Johannes Fürnkranz、Vu-Linh Nguyen 和 Eyke Hüllermeier。学习梯度提升的多标签分类规则。载于数据库中的机器学习与知识发现：*2020*年欧洲会议 *ECML PKDD，2020*年*9*月*14–18*日，比利时根特，论文集，第三部分，第124–140页。施普林格，2020年。[98] Pawel Trajdos 和 Marek Kurzynski。用于多标签学习的动态分类器链。载于模式识别：第*41*届德国 *DAGM* 会议（*DAGM GCPR 2019*），*2019*年*9*月*10–13*日，德国多特蒙德，论文集*41*，第567–580页。施普林格，2019年。[99] Madson L. D. Dias、Lucas S. de Souza、Ajalmar R. da Rocha Neto 和 Amauri H. de Souza Junior。反向邻域：一种为最小学习机选择参考点的新方法。载于欧洲人工神经网络、计算智能与机器学习研讨会论文集——*ESANN 2018*，第201–206页，2018年。

[100] José A. V. Florêncio、Madson L. D. Dias、Ajalmar R. da Rocha Neto 和 Amauri H. de Souza Júnior。基于模糊 C 均值的最小学习机参考点选择方法。载于 Guilherme A. Barreto 和 Ricardo Coelho 编，《模糊信息处理》，第 398–407 页，Cham，2018年。施普林格国际出版公司。[101] Átilla N. Maia、Madson L. D. Dias、João P. P. Gomes 和 Ajalmar R. da Rocha Neto。最优选择的最小学习机。载于 Hujun Yin、David Camacho、Paulo Novais 和 Antonio J. Tallón-Ballesteros 编，《智能数据工程与自动化学习——*IDEAL 2018*》，第 670–678 页，Cham，2018年。施普林格国际出版公司。[102] David M. Allen。变量选择与数据增广之间的关系及一种预测方法。*Technometrics*，16(1):125–127，1974年。[103] Emil Eirola、Andrey Gritsenko、Anton Akusok、Kaj-Mikael Björk、Yoan Miche、Dušan Sovilj、Rui Nian、Bo He 和 Amaury Lendasse。用于多类分类的极限学习机：利用高斯混合模型改进预测。载于人工神经网络国际工作会议，第 153–164 页。施普林格，2015年。[104] Alisson SC Alencar、Ajalmar R Rocha Neto 和 João Paulo P Gomes。通过遗传算法为极限学习机提出一种新的剪枝方法。*Applied Soft Computing*，44:101–107，2016年。[105] Reem Alotaibi 和 Peter Flach。面向代价敏感分类的多标签阈值化。*Neurocomputing*，436:232–247，2021年。[106] Rafael B Pereira、Alexandre Plastino、Bianca Zadrozny 和 Luiz HC Merschmann。多标签分类性能度量的相关性分析。*Information Processing & Management*，54(3):359–369，2018年。[107] Min-Ling Zhang、Yu-Kun Li、Xu-Ying Liu 和 Xin Geng。多标签学习的二元关联：综述。*Frontiers of ComputerScience*，12(2):191–202，2018年4月。[108] Jesse Read、Bernhard Pfahringer、Geoff Holmes 和 Eibe Frank。分类器链：综述与展望。*Journal of Artificial IntelligenceResearch*，70:683–718，2021年。[109] Grigorios Tsoumakas、Ioannis Katakis 和 Ioannis Vlahavas。在标签数量庞大的领域中实现有效且高效的多标签分类。载于 *ECML/PKDD 2008* 多维数据挖掘研讨会（*MMD'08*）论文集，第 21 卷，第 53–59 页，2008年。[110] Jesse Read 和 Fernando Perez-Cruz。用于多标签分类的深度学习，2014年。[111] J. R. Quinlan。决策树归纳。1(1):81–106。[112] Janez Demšar。多个数据集上分类器的统计比较。*The Journal ofMachine LearningResearch*，7:1–30，2006年。[113] Joakim Linja、Joonas Hämäläinen、Paavo Nieminen 和 Tommi Kärkkäinen。随机化算法能否提高最小学习机的效率？*Machine Learningand Knowledge Extraction*，2(4):533–557，2020年。[114] Xiaoping Lai、Jiuwen Cao、Xiaofeng Huang、Tianlei Wang 和 Zhiping Lin。用于正则化极限学习机的最大分裂松弛 ADMM。*IEEE Transactionson Neural Networks and Learning Systems*，31(6):1899–1913，2019年。[115] Jiapeng Wang 和 Yihong Dong。文本相似度度量：综述。*Information*，11(9):421，2020年。[116] Alessia Amelio 和 Clara Pizzuti。基于图像块的图像差异度量。*Neurocomputing*，171:362–378，2016年。

附录 A 证明

为清楚起见，我们将命题1的证明分为两部分，分别对应编号的陈述。

ˆ第一部分：对于任意 $[\delta_k]_k$，NN-MLM 无法得到最优解。

ˆ证明。 为了说明对于任意 $[\delta_k]_k$ NN-MLM 并不总能得到最优解，只需给出一个反例。因此，考虑 $L = 2$, $\mathbf{y}_1 = [0,0]^\top, \mathbf{y}_2 = [0,1]^\top, \mathbf{y}_3 = [1,0]^\top, \mathbf{y}_4 = [1,1]^\top$, 以及 $\mathbf{t}_i = \mathbf{y}_i$，对所有 i 均成立。 ˆ ˆ ˆ ˆ 另外，定义 $\delta_1 = 1, \delta_2 = 10, \delta_3 = 2, \delta_4 = 2$。

在这种情况下，NN-MLM 给出的预测为 $\hat{\mathbf{y}} = \mathbf{y}_1 = [0,0]^\top$，其对应的损失为

$$\begin{aligned} J(\mathbf{y}_1) &= (-\hat{\delta}_1^2)^2 + (1-\hat{\delta}_2^2)^2 + (1-\hat{\delta}_3^2)^2 + (2-\hat{\delta}_4^2)^2 \\ &= 1 + 99^2 + 3^2 + 2^2 = 9815 \end{aligned}$$

现在，考虑值 $J(\mathbf{y}_3)$:

$$\begin{aligned} J(\mathbf{y}_3) &= (1-\hat{\delta}_1^2)^2 + (2-\hat{\delta}_2^2)^2 + (0-\hat{\delta}_3^2)^2 + (1-\hat{\delta}_4^2)^2 \\ &= 0 + 98^2 + 4^2 + 3^2 = 9629 \end{aligned}$$

因此，由于 $J(\mathbf{y}_3) < J(\mathbf{y}_1)$，$\hat{\mathbf{y}} = \mathbf{y}_1$ 不是公式 1 中多边定位问题的最优解。 □

第二部分：如果对于某个 $\mathbf{y}' \in \mathbb{Y}$，$\delta_k = \|\mathbf{y}' - \mathbf{t}_k\|$，那么 NN-MLM 会返回一个最优解

证明。回想一下，我们的损失函数为 $J(\mathbf{y}) = \sum_{k=1}^{K} \left(\|\mathbf{y} - \mathbf{t}_k\|^2 - \hat{\delta}_k^2 \right)^2$。我们可以将其改写为：$J$

$$J(\mathbf{y}) = \sum_{l=1}^{2^L} N_l \left(\|\mathbf{y} - \mathbf{y}_l\|^2 - \hat{\delta}_l^2 \right)^2$$

其中，$\mathbf{y}_l$ 是第 l 个多标签分配，而 $\hat{\delta}_l^2$ 是$\mathbf{y}$ 与关联于第 l 个多标签分配 $\mathbf{y}_l$的参考点之间距离的估计值。值 N_l 表示等于 $\mathbf{y}_l$的输出参考点数量。

我们关注这样的情形：对于某个 $\mathbf{y}' - \mathbf{y}_l\|^2$，有 $\hat{\delta}_l^2 = \|$，这给出

$$\operatorname*{argmin}_{\mathbf{y} \in \mathbb{Y}} \sum_{l=1}^{2^L} N_l \left(\|\mathbf{y} - \mathbf{y}_l\|^2 - \|\mathbf{y}' - \mathbf{y}_l\|^2 \right)^2 .$$

上述方程的一个极小值点是 $\mathbf{y} = \mathbf{y}'$，由此得到 $J(\mathbf{y}') = 0$。为了完成证明，我们只需从 $[\delta_k]_k$ 中恢复 $\mathbf{y}'$ 的值 $\hat{}$。我们通过计算满足 $\delta_{k^\star} = 0$ 的 $\mathbf{y}' = \mathbf{t}_{k^\star}$ 来实现这一点。值得注意的是，这对应于 NN-MLM 方法的预测。

□

附录 B 评估指标

在传统的单标签分类（SLC）中，比较不同机器学习方法的分类性能十分直接。在 SLC 中，机器学习模型提供给评估指标的输入只是一个标签集合，其标签基数与测试集相同。许多机器学习模型都有一个中间层，用于给出各个标签的预测分数，随后通过选择得分最高的标签来预测相关标签。通常，这些中间分数不会用于 SLC 评估指标。从具有此类中间层的分类器模型的角度来看，MLC 与 SLC 的区别仅在于预测阶段对中间值的处理方式。在 MLC 中，每个实例的相关标签数并没有一个全局固定值。MLC 指标可分为两大类：1）基于排序的指标；2）基于二分的指标。基于二分的指标还可进一步分为两个子类：2.1）基于样本的指标；2.2）基于标签的指标。

对于由实例 $\mathbf{X}_{ts} = \{\mathbf{x}_i^*\}_{i=1}^{N_{ts}}$ 组成的测试集以及真实标签集合 $\mathbf{Y}_{gt} = \{\mathbf{g}_i\}_{i=1}^{N_{ts}}$，其中 $\mathbf{X}_{ts} \subset \mathbb{R}^M$和 $\mathbf{Y}_{gt} \subset \{0,1\}^L$。对于分类器模型 h，我们得到一组多标签预测 $\mathbf{Y}_{pr} = \{\mathbf{y}_i^*\}_{i=1}^{N_{ts}}$，其中$\mathbf{y}_i^* = h(\mathbf{x}_i^*) \in \{0,1\}^L$。

汉明损失是一种基于 实例的评估指标，也是最常用的多标签分类指标之一。其定义为

$$\text{HAMMING LOSS} = \frac{1}{N_{ts} L} \sum_{i=1}^{N_{ts}} \mathbf{y}_i^* \Delta \mathbf{g}_i \tag{11}$$

其中，Δ 表示对称差，其定义为 $\sum_{j=1}^{L} \mathbf{a}_{(j)} \oplus \mathbf{b}(j)$，对于 $\mathbf{a},\mathbf{b} \in \{0,1\}^L$。在多标签分类问题中，准确率定义为

$$\text{ACCURACY} = \frac{1}{N_{ts}} \sum_{i=1}^{N_{ts}} \frac{\mathbf{y}_i^* \cdot \mathbf{g}_i}{\mathbf{y}_i^* \cdot \mathbf{y}_i^* + \mathbf{g}_i \cdot \mathbf{g}_i - \mathbf{y}_i^* \cdot \mathbf{g}_i} \tag{12}$$

微平均召回率和微平均精确率指标定义为

$$\text{MICRO RECALL} = \frac{\sum_{j=1}^{L} TP_j}{\sum_{j=1}^{L} TP_j + \sum_{j=1}^{L} FP_j}, \tag{13}$$

$$\text{MICRO PRECISION} = \frac{\sum_{j=1}^{L} TP_j}{\sum_{j=1}^{L} TP_j + \sum_{j=1}^{L} FN_j}, \tag{14}$$

其中，对于标签 j 和 N_{ts} 个实例，TP_j 表示真正例数，FP_j 表示假正例数，FN_j 表示假负例数。相应的宏平均指标定义为

$$\text{MACRO RECALL} = \frac{1}{L}\sum_{j=1}^{L} \frac{TP_j}{TP_j + FP_j}, \tag{15}$$

$$\text{MACRO PRECISION} = \frac{1}{L}\sum_{j=1}^{L} \frac{TP_j}{TP_j + FN_j}, \tag{16}$$

其中，TP_j、FP_j 和 FN_j 与微平均指标（公式 (13)–(14)）中的定义相同。除汉明损失外，其他常用的基于二分的多标签分类指标还包括微平均 F1 和宏平均 F1 [106, 26]，它们通过相应召回率和精确率的调和平均数来定义：

$$\text{MICRO F1} = \text{HMEAN}\,(\text{MICRO PRECISION}, \text{MICRO RECALL}), \tag{17}$$

$$\text{MACRO F1} = \text{HMEAN}\,(\text{MACRO PRECISION}, \text{MACRO RECALL}), \tag{18}$$

其中，$\text{HMEAN}(a, b) = \frac{2ab}{a+b}$ 适用于 $a, b \in \mathbb{R}_{+}$。

对于给定的分类器 h，我们将 h_0 记为一个为每个标签分配标签得分的函数。那么，对于分类器 h，预测标签得分集合 $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N_{test}}$，其中 $\mathbf{z}_i = h_0(\mathbf{x}_i) \in \mathbb{R}^L$ 且 $\mathbf{x}_i \in \mathbf{X}_{test}$。

排序损失是一种基于排序的多标签分类评估指标，用于衡量分类器的整体质量，且不受阈值选择的影响。排序损失指标定义为

$$\text{RANKING LOSS} = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}} \frac{|\{(j,k) \mid \mathbf{z}_{i(j)} < \mathbf{z}_{i(k)},\ j \in g_i^+,\ k \in g_i^0\}|}{|g_i^+||g_i^0|}, \tag{19}$$

其中，g_i^+ 和 g_i^0 分别是相关真实标签和不相关真实标签的集合。另一种常用的基于排序的指标——覆盖率，定义为

$$\text{COVERAGE} = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}} |\{j \mid \min_{k \in g_i^+} \mathbf{z}_{i(k)} \le \mathbf{z}_{i(j)}\}|, \tag{20}$$

其中，$g_i^+ = \{j \mid j = \{1, ..., L\}, g_{i(j)} = 1\}$ 表示实例 i的活跃真实标签。简言之，该指标表示从排序中找出所有实际标签所需的步数。

单一错误表示排序最高标签的误分类误差。其定义为

$$\text{ONE ERROR} = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}} 1 - \mathbf{g}_{i(j^*)}, \tag{21}$$

其中，$j^* = \text{argmax}_j\, \mathbf{z}^{i(j)}$。请注意，单一错误包含的信息与 $k = 1$ 处的精度（P@1）指标相同，因为单一错误 $= 1 - P$@1。

平均精度定义为

$$\text{AVERAGE PRECISION} = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}} \frac{1}{|g_i^+|} \sum_{j \in g_i^+} \frac{|\{m \mid \mathbf{z}_{i(j)} \le \mathbf{z}_{i(m)}, m \in g_i^+\}|}{|\{k \mid \mathbf{z}_{i(j)} \le \mathbf{z}_{i(k)}\}|}. \tag{22}$$

此指标计算排名高于标签 j 的相关标签的平均比例，除以覆盖率加一，并针对标签 j（覆盖率以上）。

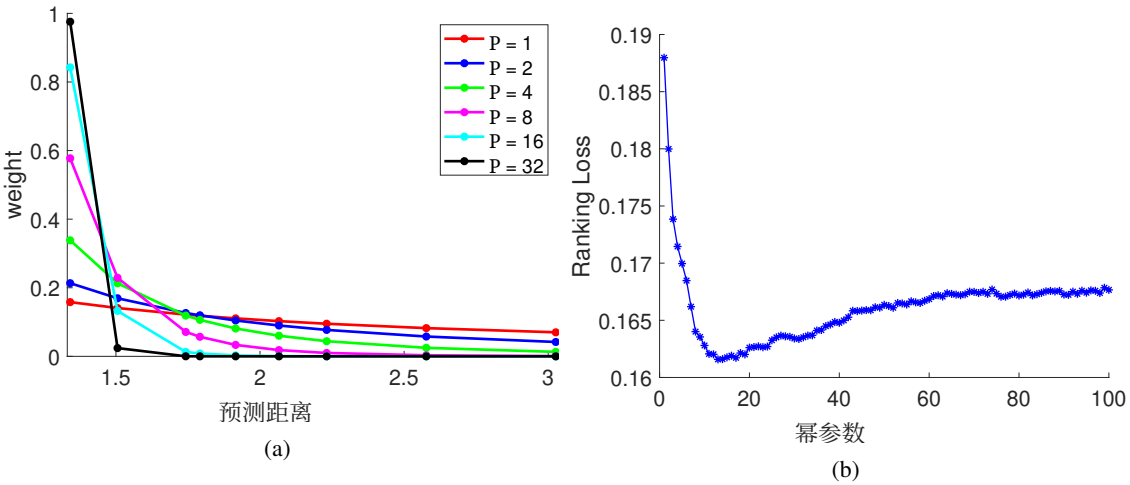


图 5：(a) 显示了在不同幂参数取值下，权重值随 MLM 的预测距离变化的情况。(b) 显示了酵母数据集的排序损失随幂参数 P 变化的情况。

附录 C 幂参数的影响

图 5 展示了幂参数对权重和整体排序性能的影响。图 5a 展示了一个包含八个预测距离的样本在 $P = 1, 2, 4, 8, 16, 32$ 下对应的权重。这组预测距离取自 MLM 模型对 Yeast 数据集的预测结果。需要注意的是，由于我们使用欧几里得距离作为距离度量，并采用标签的二进制表示，因此距离 $\sqrt{m}$ 对应于 m 标签差异。在该示例中，最小的预测距离接近 $\sqrt{2}$，对应于两个标签差异。需要注意的是，这反映了模型外推程度的估计。随着幂参数值的增大，较小预测距离所占的权重相对更大，这意味着预测更加依赖于由较小参考点集合给出的预测。图 5b 展示了所提出方法的排序损失（参见公式 (19)）随 P 变化的情况，实验采用 Yeast 数据集的测试集。对于较小的 P 值，排序损失逐渐降低，直至在 $P = 13$ 处达到最优值。此后，排序损失开始略微增加，这表明在标签向量的凸组合中，较小预测距离被赋予了过大的权重，从而导致标签的最优排序出现偏差。

附录D 基于排序的指标结果

表3：排序损失结果

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.021	0.246	0.084	**0.060**	**0.164**	0.117	0.068	0.114	0.003	0.061
CC-SVM	**0.019**	0.245	0.083	0.064	0.170	0.118	**0.067**	0.117	0.003	0.062
HOMER	0.090	0.297	0.183	0.119	0.205	0.352	0.255	0.379	0.028	0.177
RF-PCT	0.024	0.151	**0.079**	0.072	0.167	0.117	0.093	**0.106**	0.006	**0.047**
ML-kNN	0.045	0.283	0.093	0.093	0.172	0.130	0.217	0.129	0.031	0.055
BR-MLM	0.024	0.146	0.089	0.066	0.167	0.184	0.080	0.132	**0.000**	0.061
ML-MLM	0.030	**0.142**	0.081	0.065	0.166	**0.115**	0.078	0.118	**0.000**	0.051
LLS-MLM	0.029	0.155	0.111	0.068	0.174	0.194	0.089	0.148	**0.000**	0.082

表 4：覆盖率结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	1.610	2.307	12.530	**0.399**	6.330	104.800	**20.926**	530.126	1.311	20.481
CC-SVM	**1.471**	2.317	12.437	0.417	6.439	105.428	21.078	537.388	1.302	20.333
HOMER	5.324	2.634	24.190	0.739	7.285	250.800	65.626	933.956	2.369	47.046
RF-PCT	1.619	1.827	12.074	0.461	6.179	107.412	25.854	**504.999**	1.219	**16.926**
ML-kNN	2.844	2.490	13.181	0.569	6.414	113.046	56.266	589.898	2.155	18.719
BR-MLM	1.592	1.792	14.143	0.439	6.339	157.048	25.883	664.122	1.208	22.412
ML-MLM	2.026	**1.743**	**11.827**	0.426	**6.022**	**101.548**	24.094	510.747	**1.207**	17.956
LLS-MLM	2.008	1.847	16.057	0.450	6.610	166.308	29.010	715.828	**1.207**	27.611

表5：单一错误结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.135	0.386	0.237	**0.180**	0.236	0.660	0.346	0.354	0.029	0.188
CC-SVM	**0.123**	0.376	0.238	0.204	0.268	0.674	0.342	0.367	0.026	0.193
HOMER	0.216	0.411	0.314	0.216	0.248	0.652	0.466	0.509	0.050	0.219
RF-PCT	0.174	0.262	**0.221**	0.210	0.248	0.608	0.433	0.332	0.006	0.159
ML-kNN	0.279	0.406	0.280	0.242	**0.234**	0.706	0.576	0.416	0.190	0.182
BR-MLM	0.153	0.267	0.230	0.194	0.236	0.600	**0.341**	**0.309**	**0.002**	**0.113**
ML-MLM	0.146	**0.257**	0.295	0.195	**0.234**	0.626	0.357	0.422	**0.002**	0.121
LLS-MLM	0.158	0.262	0.275	0.191	0.236	**0.598**	0.345	0.356	**0.002**	0.129

表 6：平均精度结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.896	0.721	0.693	**0.893**	**0.768**	0.303	0.597	0.351	0.978	0.686
CC-SVM	**0.901**	0.724	0.695	0.881	0.755	0.293	0.599	0.343	0.981	0.672
HOMER	0.786	0.698	0.604	0.848	0.740	0.222	0.407	0.231	0.945	0.583
RF-PCT	0.868	0.812	0.698	0.874	0.757	**0.334**	0.525	0.395	0.996	0.737
ML-kNN	0.784	0.694	0.635	0.851	0.758	0.266	0.349	0.326	0.844	0.703
BR-MLM	0.884	0.820	**0.709**	0.883	0.766	0.327	**0.608**	**0.407**	**0.999**	**0.745**
ML-MLM	0.882	**0.827**	0.685	0.883	0.767	0.321	0.587	0.345	**0.999**	0.727
LLS-MLM	0.876	0.813	0.683	0.883	0.760	0.322	0.597	0.391	**0.999**	0.724

附录 E 基于二分的指标结果

表 7：准确率结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.206	0.361	0.446	0.689	0.520	0.030	0.348	0.136	0.891	0.403
CC-SVM	0.211	0.356	0.334	0.723	0.527	0.030	0.352	0.137	0.899	0.390
HOMER	0.713	0.471	0.478	0.717	0.559	0.179	0.330	0.207	0.888	0.413
RF-PCT	0.591	0.519	0.416	0.541	0.478	0.009	0.166	0.146	0.914	0.441
ML-kNN	0.528	0.319	0.319	0.629	0.492	0.014	0.129	0.102	0.574	0.421
BR-MLM	0.767	0.598	0.470	0.767	0.545	0.168	0.393	0.150	0.992	0.463
ML-MLM	0.762	**0.609**	0.473	0.764	**0.568**	**0.197**	**0.411**	0.223	**0.994**	0.466
LLS-MLM	0.767	0.586	0.465	**0.770**	0.543	0.173	0.400	0.149	0.993	0.465
NN-MLM	**0.775**	0.586	0.455	**0.770**	0.553	0.178	0.399	0.143	0.993	**0.467**
BPNN	0.543	0.283	0.344	0.200	0.518	0.134	0.100	0.149	0.327	0.349
BR-Ada	0.731	0.458	0.384	0.609	0.500	0.015	0.283	0.087	0.481	0.391
BR-RF	0.445	0.494	0.443	0.556	0.495	0.050	0.192	0.188	0.993	0.448
EBR-J48	0.655	0.519	0.479	0.648	0.511	0.139	0.343	**0.224**	0.828	0.443
ECC-J48	0.681	0.537	**0.480**	0.659	0.519	0.129	0.337	0.188	0.820	0.429
ML-kNN(d)	0.457	0.357	0.356	0.651	0.479	0.050	0.154	0.137	0.587	0.418
Rakel-SGDC	0.646	0.349	0.380	0.600	0.507	0.103	0.318	0.134	0.612	0.402
RakelSVC	0.414	0.254	0.412	0.684	0.507	0.022	0.203	0.117	0.725	0.399

表 8：汉明损失的结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.077	0.257	**0.045**	0.079	**0.190**	0.017	**0.012**	**0.018**	0.013	0.032
CC-SVM	0.077	0.256	0.064	0.082	0.193	0.017	**0.012**	**0.018**	0.013	0.032
HOMER	0.012	0.361	0.051	0.082	0.207	0.012	0.014	0.022	0.015	0.038
RF-PCT	0.014	0.189	0.046	0.094	0.197	**0.009**	0.013	**0.018**	0.011	**0.029**
ML-kNN	0.017	0.294	0.051	0.099	0.198	**0.009**	0.014	**0.018**	0.058	0.031
BR-MLM	**0.011**	0.194	0.047	0.078	0.193	0.010	0.013	**0.018**	**0.001**	**0.029**
ML-MLM	0.013	**0.186**	0.054	0.078	0.195	0.014	0.018	0.025	**0.001**	0.036
LLS-MLM	**0.011**	0.201	0.049	**0.077**	0.195	0.010	0.013	**0.018**	**0.001**	**0.029**
NN-MLM	**0.011**	0.202	0.050	**0.077**	0.193	0.010	0.013	**0.018**	**0.001**	**0.029**
BPNN	0.022	0.407	0.072	0.314	0.215	0.015	0.028	0.029	0.117	0.046
BR-Ada	**0.011**	0.238	0.048	0.099	0.205	**0.009**	0.013	**0.018**	0.071	0.032
BR-RF	0.017	0.202	0.046	0.092	0.195	0.010	0.013	**0.018**	**0.001**	**0.029**
EBR-J48	0.015	0.216	0.051	0.099	0.217	0.015	0.017	0.024	0.023	0.036
ECC-J48	0.014	0.236	0.056	0.109	0.233	0.023	0.018	**0.018**	0.025	0.041
ML-kNN(d)	0.018	0.304	0.053	0.090	0.209	0.010	0.015	0.019	0.056	0.033
Rakel-SGDC	0.014	0.400	0.057	0.143	0.204	0.012	0.014	**0.018**	0.053	0.032
Rakel-SVC	0.018	0.337	0.046	0.085	0.193	**0.009**	0.013	**0.018**	0.036	0.031

表 9：微平均 F1 的结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.343	0.509	0.564	0.761	0.652	0.059	0.457	0.234	0.932	0.533
CC-SVM	0.350	0.503	0.482	0.757	0.650	0.059	**0.462**	0.236	0.936	0.509
HOMER	0.773	0.588	0.591	0.764	0.673	0.275	0.429	0.339	0.927	0.553
RF-PCT	0.693	0.672	0.537	0.669	0.617	0.018	0.230	0.248	0.945	0.563
ML-kNN	0.634	0.457	0.466	0.661	0.625	0.030	0.206	0.175	0.682	0.545
BR-MLM	0.788	0.705	0.580	0.775	0.663	0.256	0.437	0.275	0.994	0.586
ML-MLM	0.765	**0.715**	0.571	**0.781**	**0.678**	**0.285**	0.408	0.348	**0.996**	**0.588**
LLS-MLM	0.784	0.695	0.570	0.778	0.659	0.259	0.435	0.272	0.994	0.583
NN-MLM	0.794	0.693	0.554	0.778	0.663	0.261	0.432	0.260	0.995	0.585
BPNN	0.603	0.386	0.479	0.229	0.645	0.217	0.165	0.265	0.423	0.495
BR-Ada	**0.796**	0.604	0.528	0.703	0.630	0.031	0.395	0.152	0.603	0.520
BR-RF	0.587	0.647	0.566	0.682	0.629	0.090	0.262	0.312	0.995	0.568
EBR-J48	0.739	0.657	**0.601**	0.722	0.642	0.224	0.452	**0.366**	0.888	0.583
ECC-J48	0.755	0.664	0.595	0.710	0.648	0.209	0.444	0.312	0.879	0.568
ML-kNN(d)	0.582	0.477	0.491	0.723	0.615	0.096	0.255	0.237	0.695	0.544
Rakel-SGDC	0.729	0.475	0.512	0.639	0.636	0.173	0.416	0.228	0.713	0.526
Rakel-SVC	0.564	0.319	0.534	0.742	0.638	0.045	0.267	0.202	0.806	0.519

表10：宏平均 F1 的结果。

数据集	Medical	Emotions	Enron	Scene	Yeast	Corel5k	Bibtex	Delicious	Tmc2007	Mediamill
BR-SVM	0.361	0.440	0.143	0.765	0.392	0.021	0.307	0.096	0.942	0.056
CC-SVM	0.371	0.420	0.153	0.762	0.390	0.021	0.316	0.100	0.947	0.052
HOMER	0.282	0.570	0.167	0.768	**0.447**	0.036	0.266	0.103	0.924	0.073
RF-PCT	0.207	0.650	0.122	0.658	0.322	0.004	0.055	0.083	0.857	0.112
ML-kNN	0.192	0.385	0.087	0.692	0.336	0.010	0.065	0.051	0.493	0.113
BR-MLM	0.297	0.694	0.216	0.780	0.422	0.038	0.265	0.159	0.993	0.180
ML-MLM	0.315	**0.703**	0.212	**0.789**	0.406	0.038	0.299	0.160	**0.994**	0.193
LLS-MLM	0.303	0.683	**0.218**	0.783	0.422	0.040	0.266	**0.162**	0.993	0.192
NN-MLM	0.307	0.681	0.213	0.783	0.423	0.042	0.259	0.160	0.993	**0.195**
BPNN	0.209	0.183	0.060	0.115	0.385	0.017	0.047	0.030	0.345	0.040
BR-Ada	**0.383**	0.591	0.152	0.707	0.346	0.007	0.247	0.054	0.451	0.052
BR-RF	0.129	0.628	0.180	0.676	0.347	0.007	0.080	0.158	0.993	0.111
EBR-J48	0.238	0.644	0.191	0.731	0.417	0.040	0.347	0.128	0.833	0.177
ECC-J48	0.258	0.657	0.204	0.722	0.436	**0.045**	**0.349**	0.158	0.828	0.187
MLkNN(d)	0.166	0.426	0.100	0.724	0.397	0.027	0.122	0.078	0.592	0.150
Rakel-SGDC	0.273	0.424	0.199	0.649	0.363	0.042	0.281	0.088	0.604	0.036
Rakel-SVC	0.095	0.107	0.115	0.745	0.360	0.006	0.067	0.051	0.694	0.035