智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于粒的标记增强标记分布学习

张远健, 赵天娜, 苗夺谦

引用本文:
张远健,赵天娜,苗夺谦. 基于粒的标记增强标记分布学习[J]. 智能系统学报, 2023, 18(2): 390–398.
ZHANG Yuanjian,ZHAO Tianna,MIAO Duoqian. Granule-based label enhancement in label distribution learning[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(2): 390–398.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202208015

DOI: 10.11992/tis.202208015
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20221008.1831.014.html

基于粒的标记增强标记分布学习

张远健[1], 赵天娜[2], 苗夺谦[2]
(1. 中国银联股份有限公司, 上海 201201; 2. 同济大学 电子与信息工程学院, 上海 201804)

摘 要: 标记分布学习能有效求解多标记学习任务, 然而分类器构造以获得大规模具有更强监督信息的标注为前提, 在许多应用中难以满足。一种替代的方案是以标记增强的方式从传统逻辑形式的标注中挖掘出隐含的数值型标记的重要程度。现有的标记增强方法大多假设增强后的标记需要在所有示例上保持原有逻辑标记的相关性, 不能有效保持局部标记相关性。基于粒计算理论, 提出了一种适用于标记分布学习的粒化标记增强学习方法。该方法通过 k 均值聚类构造具有局部相关性语义的信息粒, 并在粒的抽象层面上, 分别在图上依据逻辑标记的特性和属性空间的拓扑性质完成粒内示例的标记转化。最后, 将得到的标记分布在示例层面进行融合, 得到描述整个数据集标记重要程度的数值型标记。大量比较研究表明, 所提出的模型可以显著地提升多标记学习的性能。

关键词: 粒计算; 标记分布学习; 标记增强; 多标记; 不确定性; 局部相关性; 聚类; 拓扑
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2023)02−0390−09

中文引用格式: 张远健, 赵天娜, 苗夺谦. 基于粒的标记增强标记分布学习 [J]. 智能系统学报, 2023, 18(2): 390–398.
英文引用格式: ZHANG Yuanjian, ZHAO Tianna, MIAO Duoqian. Granule-based label enhancement in label distribution learning[J]. CAAI transactions on intelligent systems, 2023, 18(2): 390–398.

Granule-based label enhancement in label distribution learning

ZHANG Yuanjian[1], ZHAO Tianna[2], MIAO Duoqian[2]
(1. China UnionPay Co. Ltd, Shanghai 201201, China; 2. College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: Label distribution learning can effectively deal with multilabel learning tasks. However, the construction of a classifier is based on the premise of obtaining large-scale labels with strong supervision information, which is difficult to be satisfied in many applications. An alternative solution is to mine the importance of implicit numerical labels from the traditional logical form of annotation through label enhancement. Existing label enhancement methods mainly assume that the enhanced label must maintain the relevance of the original logical label in all instances, which fails to preserve local label correlation. This paper proposes a granular-based label enhancement distribution model applicable to label distribution learning, considering the methodology of granular computing. The method constructs information granules with local correlation semantics by employing k-means clustering and completes the labeling transformation of instances in granules on the graph according to the characteristics of logical labeling and the topological properties of attribute space at the abstract level of granules. Finally, the obtained label distribution is fused at the instance level, obtaining the numerical label describing the importance of the whole data set. Extensive studies have shown that the proposed model significantly improves the accuracy of multilabel learning.

Keywords: granular computing; label distribution learning; label enhancement; multi-label; uncertainty; local label correlation; clustering; topology

多标记分类[1-2]由于示例与标记的多重性而被广泛研究, 学习任务通常描述为判定示例与候选标记的相关性。受标记语义的约束, 传统多标

收稿日期: 2022−08−11. **网络出版日期:** 2022−10−09.
基金项目: 中国博士后科学基金资助项目 (2022M713491); 国家自然科学基金项目 (61976158).
通信作者: 张远健. E-mail: zhangyuanjian@unionpay.com.

记分类假设相关标记(不相关标记)对示例描述的相关性程度相同, 这显然不能满足人类复杂的认知需求[3]。例如, 在图像检索[4]时, 人们更多希望排名靠前的图片与检索关键词具有更强的关联性。这使得同样具有“天空”“森林”“水”和“云”等多个标记的两个自然场景的图片由于检索策略的不同而具有不同的排序; 在文本情感分析[5]中, 文章中蕴含的情绪通常是几种基础情感 (如快乐、悲伤、惊讶、愤怒、厌恶和恐惧) 的混合呈现, 若不加以区分情绪的组成, 则无法满足图书推荐等精准营销应用。上述实例表明, 仅以逻辑相关与否作为多标记分类的监督信息使得分类器在优化学习过程中损失了许多深层次语义信息。为了有效弥补分类器的学习能力缺陷, 标记分布学习[6]被提出。通过引入描述度这个度量, 标记分布学习将标记描述示例的能力从“是否”拓展成了“多少”。

然而, 获得大规模精准描述示例的数值标记是实施标记分布学习的前提, 而人工形式标注下, 数值标记的标注成本将显著高于逻辑标记的标注成本, 极大地制约标记分布学习在多标记学习场景的推广。针对上述问题, 一些学者从标记增强角度出发, 试图学习多标记上隐含的数值型标记分布。代表性模型有一阶段标记增强图模型[7](label enhanced multi-label learning, LEMLL)、模糊辨识关系的标记增强算法[8]、图拉普拉斯标记增强[9](graph Laplacian label enhancement, GLLE)。除了上述显式的标记增强方法外, 也有前人提出的模糊方法 FCM[10] 和 KM[11] 模型恢复标记重要程度, 以及基于图的标记传播的方法 LP[12] 和 ML[13] 模型。尽管上述研究对深入理解多标记数据有重要意义, 然而并没有深入地分析数据标记相关性在示例集上的分布, 使得分类准确性不够精细。

粒计算[14−18]是近年来涌现出的一类模拟人类认知复杂概念的方法论, 其主要观点是概念可以逐层粒化, 在粒化分析过程中处理概念的不确定性。在粒计算框架下, 表示学习可体现为一个由细到粗, 由具象到抽象的过程。从粒计算视角看表示学习, 是标记分布学习的新视角。已有一些基于粒的思想进行多标记数据的研究。本文作者[19−21]是国内较早系统地从粒的角度研究多标记学习的学者之一, 在全局标记相关性假设下构造的模型已经在多标记的监督学习、增量学习和主动学习等方面证明了粒计算的有效性。

在多标记数据的研究中, 标记相关性局部成立是一种更为广泛的情形, 是标记相关性在示例集上从“必然”向“或然”的拓展[22]。标记局部相关性在标记分布学习中已有一些研究, 如贾修一等[23−24]将标记局部相关性引入标记分布学习的预测模型中, 在优化目标函数的正则化项中加入标记局部相关性信息。在一定的聚类假设下, 示例自动地形成了信息粒, 刻画出潜在的主题。聚类/邻域粗糙集是利用特征空间信息形成粒的有效手段[25]。这些类簇形成一系列粒, 在粒的层面上进行局部处理。在每一个粒上, 通过利用示例的特征空间的拓扑性质和逻辑标记的信息恢复隐含的数值型标记分布。在粒化得到的局部标记分布基础上, 结合具体聚类策略, 可构造一个新的标记分布表示。

本文的主要创新点: 1)在优化多标记数据的标记分布时, 引入粒计算的思想, 通过粒化与融合策略, 能更全面地刻画示例标记分布, 减少由于具有相似局部相关性过少导致标记分布均匀化现象。2)根据局部标记相关性假设, 利用聚类策略构造了示例的粒结构, 增强了示例表示的鲁棒性。提出了一个基于粒的标记增强标记分布学习模型 (granule-graph Laplacian label enhancement, G-GLLE)。

为了验证模型的有效性, 我们与现有的标记增强算法在多个数据集下展开了性能对比。同时分析了在局部标记相关性假设下, 各模型所重构的数值标记关于示例的描述拟合程度以及粒化本身对多标记学习的贡献。实验结果均表明, 本文所构造的模型在 3 个评测上均能取得较好的学习效果。

1　相关工作

1.1　符号表示

首先, 本文先给出主要符号解释和定义[8]。示例记为x,第i个示例记为$\boldsymbol{x}_i$, 标记变量记为y, 第j个标记值记为y_j, $\boldsymbol{x}_i$的逻辑标记向量记为$\boldsymbol{l}_i=[l_{x_i}^{y_1}\ l_{x_i}^{y_2}\ \cdots\ l_{x_i}^{y_c}]^{\mathrm{T}}$, 其中$c$是可能的标记个数, $l_{x_i}^{y_1}\in\{0,1\}$。y对x的描述度记为d_x^y, $\boldsymbol{x}_i$的标记分布记为$\boldsymbol{d}_i=[d_{x_i}^{y_1}\ d_{x_i}^{y_2}\ \cdots\ d_{x_i}^{y_c}]^{\mathrm{T}}$, 其中$d_{x_i}^{y_c}\in[0,1]$。令$\mathcal{X}=\mathbf{R}^q$表示$q$维特征空间, 则标记增强形式化表示如下:

给定训练集$\boldsymbol{S}=\{(x_i,l_i)|1\leqslant i\leqslant n\}$, 其中$\boldsymbol{x}_i\in\mathcal{X}$, $\boldsymbol{l}_i\in\{0,1\}^c$, 标记增强从逻辑标记向量$\boldsymbol{l}_i$中恢复$\boldsymbol{x}_i$的标记分布$\boldsymbol{d}_i$, 此时, $\boldsymbol{S}$转化为标记分布学习的训练集。

1.2　经典标记增强算法

1.2.1　模糊标记增强

FCM[10]模型是基于模糊聚类 (FCM) 的标记增强算法, 采用模糊 C-均值聚类, 该方法试图通过迭代最小化一个目标函数来聚类特征向量。

KM[11] 模型是基于核函数的方法，根据逻辑标记将特征空间划分为两个集合，分别将两个集合样本点映射到高维特征空间，根据高维特征空间的拓扑性质计算每个示例到每个标记的隶属度，最后作 softmax 标准化得到标记分布。

基于模糊的标记增强方法对参数的鲁棒性强，不易受参数变化的影响，但对于不同类型数据的敏感性低，难以根据数据类型特点，形成独特的模型，缺少模糊隶属度与标签对应的数学理论依据。

1.2.2　基于图的标记增强

LP[12] 模型是基于标记传播的算法。它构建所有示例的图模型，通过迭代标记传播技术得到标记重要性信息，最后经 softmax 标准化得到标记分布。

ML[13] 模型是基于流形学习的算法。它根据示例的特征空间的拓扑性质构造图模型，在示例可以通过邻域中其他示例的线性组合构造的平滑假设下，构造有约束的二次规划，通过优化参数得到标记分布，最后作 softmax 的标准化。

GLLE[9] 模型是基于图拉普拉斯模型的标记增强算法，它依据平滑假设在根据示例的特征空间构造的图模型性质计算图拉普拉斯标记特征函数，利用特征空间映射的高维空间拓扑性质构造损失函数，通过优化目标函数的参数，得到标记分布，最后进行 softmax 标准化。

2　基于粒的标记增强模型

本节将详述如何根据聚类方法构造粒给出基于粒的标记增强模型。本文试图改进 GLLE[9] 的标记增强模型，提升预测效果。图 1 给出了基于粒的标记增强模型的整体框架。

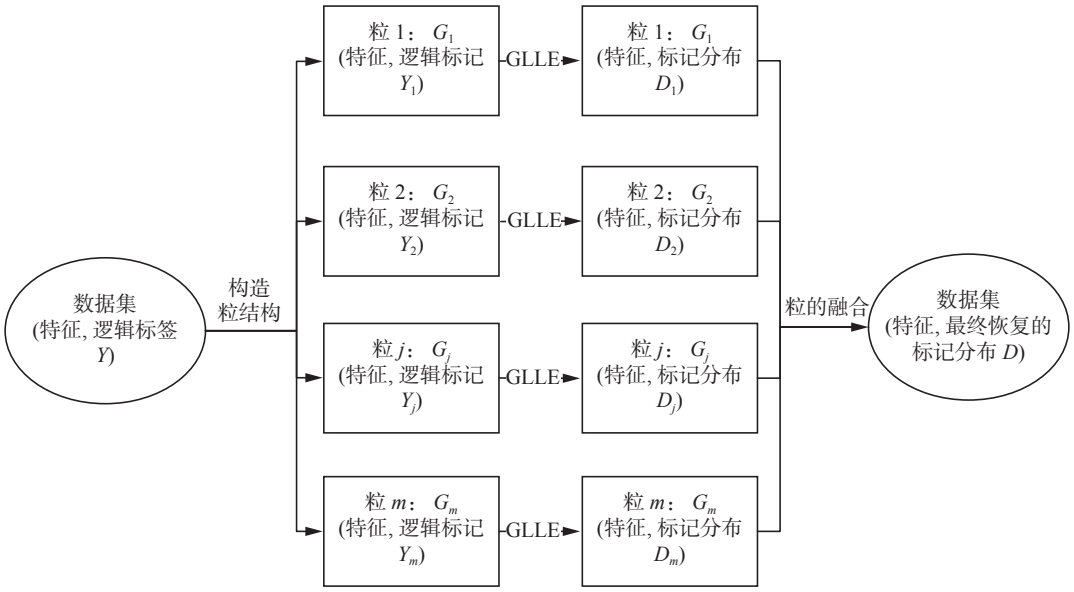


图 1　基于粒的标记增强标记模型 G-GLLE 的整体框架

Fig. 1　Framework of granule-based label enhancement for label distribution learning (G-GLLE)

不同的多标记数据的多个标记之间具有不同的局部相关性。若对所有数据采用同一标准整体处理标记相关性，不能精确描述不同示例的多标记之间的相关性，导致实验结果不佳。对于大规模数据集，若找到具有相似局部相关性的示例，对它们进行局部统一处理，挖掘该类簇中示例的隐含的数值型标记分布，再在粒的基础上进行标记分布的融合，将会大大提升工作效率，并提高分类精度。

因为标记相关性局部成立，所以，$(\boldsymbol{w},\boldsymbol{b})$序偶关系是多组的，我们用聚类方法找到哪些示例具有局部相关性，划分到同一粒中。局部相关性下，标记之间的相关关系并非对所有示例都成立，而一个$(\boldsymbol{w},\boldsymbol{b})$能表示的标记相关性是单一的，不能体现标记之间局部相关性。解决的策略为，寻找一组关于示例子集的表示，每组表示用一个图去刻画。策略上可使用类似于聚类的方法获得这样的表示，根据表示之间的示例分布，增加其他的正则项，以体现同一个示例在不同类簇间标记相关的一致性。在类簇的基础上，作带逻辑标记数据的数值型标记挖掘，得到的标记分布再作融合，体现粒的思想。

所以，本文提出的 G-GLLE 模型在训练阶段有两个问题需要解决。

1）如何构造粒？

2）如何在粒的基础上挖掘隐含的数值型标记信息？

幸运的是, 我们可以参考一些前人处理多标记数据的方法。Zhu 等[25] 认为“属性特征相似的示例, 拥有相似的标记描述”。依据流形假设, 本文进一步假设示例的某个标记只能属于一个局部标记相关假设, 并对数据集的属性空间进行聚类处理, 将属性特征相似的示例划分到相同类簇中, 构成不相交的示例粒。由上述操作过程可知, 同一粒中示例具有相似的特征, 即有相似的标记描述, 那么同一粒中示例的标记的相关性应该是一致的, 所以先从局部开始处理, 对同一粒 $G_i=[\mathbf{Idx}_i,\boldsymbol{C}_i]$中的示例进行建模, 结合该粒中示例的属性空间的拓扑性质和逻辑标记空间的特性, 在 G-GLLE 模型下求解, 该粒$G_i=[\mathbf{Idx}_i,\boldsymbol{C}_i]$的参数$(\boldsymbol{w}_i,\boldsymbol{b}_i),i=1,2,\cdots,m$。当得到每个粒中示例的数值型标记分布时, 在粒的基础上, 对所有粒进行融合, 得到整个数据集的粒的标记分布。

上述过程分为以下 3 个阶段。G-GLLE 模型 1）希望建立一个映射$\varphi:X\to\{G_1,G_2,\cdots,G_m\}$, 从原始的$q$维输入示例空间到$m$个新的$q$维空间$G_j,j=1,2,\cdots,m$; 2）然后在每个粒$G_j$上, 训练基于粒的$c$维逻辑标记$y$到$c$维数值型标记分布$D_j$的转化映射$f_k:G_j\to D_j$; 3）在整个示例空间, 将每个粒$G_j$上的示例标记分布$D_j$进行融合, $r:\{D_1,D_2,\cdots,D_m\}\to D$。

下面具体阐述这 3 个阶段。

阶段 1　基于簇的粒结构。

为了找到具有相似局部相关性的标记的示例, 需要对示例划分成不同的粒。根据流形假设, 以示例的属性空间的拓扑性质为依据, 度量示例构成粒。本文不研究如何构造粒结构, 只从一个新颖的角度探讨如何在粒结构上以标记分布方式的标注。为了便于理解, 选择符合文本假设的经典的构成粒的方式——聚类算法。粒结构形成过程中希望捕捉可识别的特征, 刻画具有相似局部标记相关性的示例。为了实现这个过程, 本文运用 k-means 聚类技术, 来挖掘数据隐藏的结构。形式化地, 将X划分到m个互不相交的簇G_i, 示例类别和簇中心分别用$\mathbf{Idx}_i$和$\boldsymbol{C}_i$表示, $i=1,2,\cdots,m$。

给定训练集S,特征矩阵$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$, 逻辑标记矩阵$\boldsymbol{L}=[\boldsymbol{l}_1\ \boldsymbol{l}_2\ \cdots\ \boldsymbol{l}_n]$, 本文目的是从逻辑标记矩阵$\boldsymbol{L}$恢复标记分布矩阵$\boldsymbol{D}=[\boldsymbol{d}_1\ \boldsymbol{d}_2\ \cdots\ \boldsymbol{d}_n]$。根据标记相关性局部假设, 我们希望将具有相似标记相关性的示例放在一起, 划分成粒, 分别局部处理, 在每个粒上单独处理将属于此粒中的示例, 从逻辑标记矩阵$\boldsymbol{L}$中恢复出标记分布矩阵$\boldsymbol{D}_i$。

具体地, 对于训练集S,运用 k-means 聚类算法, 根据特征矩阵$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$的拓扑信息, 将$\boldsymbol{S}$划分为$m$个互不相交的簇, 根据粒计算[13] 的思想, 这些簇构成了m个 q 维粒空间$\{G_1,G_2,\cdots,G_m\}$:

$$f^{k-\text{means}}:X\to\{G_1,G_2,\cdots,G_m\}\tag{1}$$

每个粒$G_i,i=1,2,\cdots,m$中的示例的特征空间拓扑性质相似。根据流形假设, 粒G_i中的示例具有相似的局部标记相关性。因此, 下一步我们分别在每个粒G_i上局部地从示例的逻辑标记中恢复隐藏的数值型标记分布, 这样可更充分地利用标记间的相关性。

阶段 2　基于粒的图拉普拉斯标记增强模型（G-GLLE）。

给定粒$G_h,h=1,2,\cdots,m$, 假设G_h中的示例为$[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_h]$。我们首先要恢复粒$G_h$中示例的逻辑标记矩阵$\boldsymbol{L}_h$中隐藏的标记分布矩阵$\boldsymbol{D}_h=[d_{h1}\ d_{h2}\ \cdots\ d_{hc}]$。G-GLLE 模型假设参数模型为

$$\boldsymbol{d}_i^h=\left(\boldsymbol{W}^h\right)^{\mathrm{T}}\varphi^h(\boldsymbol{x}_i)+\boldsymbol{b}^h\triangleq\hat{\boldsymbol{W}}^h\phi^h\tag{2}$$

式中: $\boldsymbol{W}^h=[\boldsymbol{w}^{h1}\ \boldsymbol{w}^{h2}\ \cdots\ \boldsymbol{w}^{hc}]$是一个权重矩阵; $\boldsymbol{b}^h\in\mathbf{R}^C$是偏置向量。$\varphi^h(x)$是把$x$转化到高维特征空间的非线性转化。为了方便描述, 令$\hat{\boldsymbol{W}}^h=[\boldsymbol{W}^h\ \boldsymbol{b}^h]$, $\phi_i^h=[\varphi^h(x_i);1]$。我们方法的目的是求解最优的参数$\boldsymbol{W}^{h*}$, 使得粒$G_h$中的示例$\boldsymbol{x}_i$产生合理的标记分布$\boldsymbol{d}_i^h$。因此学习目标转化为找到最优模型$\boldsymbol{W}^{h*}$, 可最小化下列目标函数:

$$\begin{aligned}\boldsymbol{W}^{h*}=\arg\min_{\hat{\boldsymbol{W}}^h}\sum_{i=1}^{n}\left\|\hat{\boldsymbol{W}}^h\phi_i^h-\boldsymbol{l}_i\right\|^2+\lambda\sum_{i,j}a_{i,j}\left\|\boldsymbol{d}_i-\boldsymbol{d}_j\right\|^2=\\ \mathrm{tr}\left((\hat{\boldsymbol{W}}^h\boldsymbol{\Phi}^h-\boldsymbol{L})^{\mathrm{T}}\left(\hat{\boldsymbol{W}}^h\boldsymbol{\Phi}^h-\boldsymbol{L}\right)\right)+\\ \lambda\mathrm{tr}\left(\hat{\boldsymbol{W}}^h\boldsymbol{\Phi}^h\boldsymbol{G}_L\left(\boldsymbol{\Phi}^h\right)^{\mathrm{T}}\left(\hat{\boldsymbol{W}}^h\right)^{\mathrm{T}}\right)\end{aligned}\tag{3}$$

式中: λ 是权重参数, $\boldsymbol{\Phi}^h=[\Phi_1^h\ \Phi_2^h\ \cdots\ \Phi_n^h]$, $\boldsymbol{G}_L=\hat{\boldsymbol{A}}-\boldsymbol{A}$是图拉普拉斯矩阵, $\hat{\boldsymbol{A}}$是对角矩阵,

$$a_{ij}=\begin{cases}\exp\left(-\dfrac{\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2}{2\sigma^2}\right), & \boldsymbol{x}_j\in N(i)\\ 0, & \boldsymbol{x}_j\notin N(i)\end{cases}\tag{4}$$

其中: $N(i)$是$\boldsymbol{x}_i$的 k 近邻的集合, 宽度系数$\sigma=1$, $a_{ii}=\sum_{j=1}^{n}a_{ij}$。

阶段 2 的优化求解过程主要参考徐宁等[9] 提出的标记增强模型 (GLLE), 求解最优的参数$\boldsymbol{W}^{h*}$, 使得粒G_h中的示例$\boldsymbol{x}_i$产生合理的标记分布$\boldsymbol{d}_i^h$。

阶段 3　基于粒的标记分布融合。

上述两阶段已经求出粒G_h对应的最优参数$\boldsymbol{W}^{h*}$以及标记分布$D^h,h=1,2,\cdots,m$。由于本文采用 k-means 聚类, 得到的粒是关于原数据集的一个划分, 任意两个粒之间不相交, 不存在共有的示例,

使得每个示例对应的标记分布唯一，所以在基于粒的标记分布融合过程中，为了求出整体训练集 S 中所有示例的标记分布，只需要在粒的层面将每个粒 G_h 对应的标记分布 $D^h, h=1,2,\cdots,m$ 拼接在一起即可。因此，恢复出最终的标记分布为 $D=\{D^1, D^2,\cdots,D^h,\cdots,D^m\}$。

3　实验及结果

3.1　数据集

实验在 6 个真实数据集 (http://cse.seu.edu.cn/PersonalPage/xgeng/LDL/index.htm) 上进行，涉及人脸表情图像，自然场景图象等领域。数据集具体描述如表 1 所示。

表 1　基准标记分布数据集特性

Table 1　Characteristics on benchmarks

数据集	示例数	特征数	标记数
SBU_3DFE(3DFE)	2 500	243	6
Natural Scene(Natural)	2000	294	9
SJAFFE	213	243	6
Yeast-spoem(SPOEM)	2 465	24	2
Yeast-spo5(SPO5)	2 465	24	3
Yeast-spo(SPO)	2 465	24	6

3.2　评价指标

为了估计恢复后的标记分布的效果，本文采用 5 个指标度量恢复标记分布和真实标记分布的平均距离和相似性。5 个评价指标分别为切比雪夫距离 (Chebyshev, 简记为 Cheb)、克拉克距离 (Clark)、KL 散度 (KL Divergence, 简记为 KL)、余弦距离 (Cosine) 以及交叉相似性 (Intersection, 简记为 Intersec)。假设真实的标记分布是 $\boldsymbol{d}=(d_1,d_2,\cdots,d_c)$，预测的标记分布是 $\hat{\boldsymbol{d}}=(\hat{d}_1,\hat{d}_2,\cdots,\hat{d}_c)$，5 个指标的定义如表 2 所示。↑表示指标值越大，预测效果越好；↓表示指标值越小，预测效果越好。

3.3　实验结果

3.3.1　标记分布恢复能力对比

为了评价 G-GLLE 在标记增强方面的能力，我们设计了如下实验比较不同标记增强算法在标记分布恢复方面的能力。首先生成 6 个标记分布数据集对应的逻辑标记。在多标记分类指标上评测，对于训练集中的示例，采用如下方式完成从数值标记到逻辑标记的二值化：若标记分布值相对大，且和超过 0.5 的标记，则标为 1；其余标记标为 0。

表 2　评价指标

Table 2　Evaluation indexes

指标名称	计算公式
(Cheb)↓	$\mathrm{Dis}_1(\boldsymbol{d},\hat{\boldsymbol{d}})=\max_j\lvert d_j-\hat{d}_j\rvert$
Clark↓	$\mathrm{Dis}_2(\boldsymbol{d},\hat{\boldsymbol{d}})=\sqrt{\sum_{j=1}^{c}\frac{(d_j-\hat{d}_j)^2}{(d_j+\hat{d}_j)^2}}$
(KL)↓	$\mathrm{Dis}_3(\boldsymbol{d},\hat{\boldsymbol{d}})=\sum_{j=1}^{c}d_j\ln\frac{d_j}{\hat{d}_j}$
Cosine↑	$\mathrm{Dis}_4(\boldsymbol{d},\hat{\boldsymbol{d}})=\frac{\sum_{j=1}^{c}d_j\hat{d}_j}{\sqrt{\sum_{j=1}^{c}d_j^2}\sqrt{\sum_{j=1}^{c}\hat{d}_j^2}}$
(Intersec)↑	$\mathrm{Dis}_5(\boldsymbol{d},\hat{\boldsymbol{d}})=\sum_{j=1}^{c}\min(d_j,\hat{d}_j)$

基于所获得的逻辑标记，实验选取了 6 个对比的标记增强学习代表性算法 (GLLE[9]、LEMLL[7]、FCM[10]、KM[11]、LP[12]、ML[13]) 和本文提出的算法 G-GLLE。比较上述算法在标记增强后获得的数值标记与真实的数值标记之间的相似性。

对 G-GLLE 和 GLLE 模型，参数 λ 设置为 0.01 并且邻域数 k 设为数据集标记个数加 1(记为 $c+1$)。GLLE 的核函数是高斯核函数。LP 的参数设为 0.5。ML 的邻域数 k 设为 $c+1$。FCM 的参数 β 设为 2。KM 的核函数设为线性核函数。LEMLL 中 k 设为 10, ε 设为 0.1, α、β 和 γ 从集合 { 1/ 64, 1/ 16, 1 /4, 1, 4, 16, 64} 中选择。

十折交叉验证下，所有算法恢复的标记分布在 5 个度量下的平均性能如表 3~8 所示，其中无法得到指标结果的值标注为 NaN。

表 3　数据集 3DFE 的实验结果

Table 3　Experimental results of 3DFE

评价标准	G-GLLE	GLLE	FCM	KM	LP	ML	LEMLL
Cheb	**0.0939**	0.123 1	0.139 2	0.0961	0.106 7	0.203 5	**0.095 5**
Clark	0.3029	0.380 2	0.407 9	0.305 7	**0.295 7**	1.116 6	**0.2923**
KL	0.043 9	0.065 8	0.080 6	0.0431	**0.042 9**	0.279 9	**0.0399**
Cosine	0.9526	0.930 2	0.912 5	**0.955 0**	0.9526	0.861 6	**0.9577**
Intersec	0.8819	0.853 0	0.840 8	0.881 2	**0.883 6**	0.691 9	**0.8872**

表 4　数据集 Natural 的实验结果

Table 4　Experimental results of Natural

评价标准	G-GLLE	GLLE	FCM	KM	LP	ML	LEMLL
Cheb	**0.3135**	0.3346	0.3613	0.3147	NaN	**0.2571**	0.3159
Clark	**2.4506**	2.4607	2.4787	**2.4508**	NaN	2.4948	2.4888
KL	**2.7613**	3.2342	3.5842	2.8549	NaN	**1.1377**	3.0110
Cosine	**0.7744**	0.6812	0.5969	**0.7612**	NaN	**0.8661**	0.7348
Intersec	0.4930	0.4241	0.3751	0.4790	**1.0000**	**0.6564**	0.4554

表 5　数据集 SJAFFE 的实验结果

Table 5　Experimental results of SJAFFE

评价标准	G-GLLE	GLLE	FCM	KM	LP	ML	LEMLL
Cheb	**0.0754**	**0.0743**	0.1228	0.1002	0.0886	0.1826	0.0730
Clark	0.3011	**0.2951**	0.4363	0.3464	0.3040	1.0203	**0.2753**
KL	0.0343	**0.0332**	0.0770	0.0468	0.0363	0.2430	**0.0296**
Cosine	0.9674	**0.9683**	0.9219	0.9541	0.9640	0.8760	**0.9709**
Intersec	0.8940	0.8960	0.8449	0.8804	**0.8962**	0.7122	**0.9023**

表 6　数据集 SPOEM 的实验结果

Table 6　Experimental results of SPOEM

评价标准	G-GLLE	GLLE	FCM	KM	LP	ML	LEMLL
Cheb	**0.0636**	0.0835	0.2082	0.0763	0.0886	0.3035	**0.0660**
Clark	**0.0974**	0.1246	0.3184	0.1190	0.1401	0.6227	**0.1019**
KL	**0.0157**	0.0252	0.1110	0.0197	0.0242	0.2432	**0.0154**
Cosine	**0.9881**	0.9799	0.9075	0.9847	0.9809	0.8733	**0.9880**
Intersec	**0.9364**	0.9165	0.7918	0.9237	0.9114	0.6965	**0.9340**

表 7　数据集 SPO5 的实验结果

Table 7　Experimental results of SPO5

评价标准	G-GLLE	GLLE	FCM	KM	LP	ML	LEMLL
Cheb	0.0781	0.0964	0.2210	**0.0716**	**0.0668**	0.2756	0.0768
Clark	0.1559	0.1911	0.3917	**0.1556**	0.1379	0.7700	**0.1552**
KL	0.0213	0.0320	0.1170	**0.0180**	**0.0164**	0.2568	0.0195
Cosine	0.9816	0.9721	0.8905	**0.9859**	**0.9866**	0.8718	0.9832
Intersec	0.9219	0.9036	0.7790	**0.9284**	**0.9332**	0.7244	0.9232

表 8　数据集 SPO 的实验结果

Table 8　Experimental results of SPO

评价标准	G-GLLE	GLLE	FCM	KM	LP	ML	LEMLL
Cheb	**0.0496**	0.0588	0.0622	**0.0496**	**0.0456**	0.3108	0.0484
Clark	0.2162	0.2532	0.2715	0.2341	**0.1887**	1.3901	**0.2130**
KL	0.0190	0.0266	0.0305	0.0204	**0.0156**	0.5314	**0.0170**
Cosine	0.9831	0.9762	0.9727	0.9820	**0.9861**	0.7616	**0.9848**
Intersec	0.9277	0.9137	0.9058	0.9202	**0.9385**	0.5808	**0.9294**

通过表 3~8 可以看出，G-GLLE 算法有良好的恢复标记分布的能力，恢复标记分布标记的效果至少与现有的恢复标记分布算法的效果相当。本文对每个数据集的每个度量指标排在第 1、第 2、第 3 的算法结果分别以同时加粗和下划线、仅加粗、仅下划线的形式高亮显示，发现 G-GLLE 算法的恢复标记分布的效果排名前 3 名的概率在 28/36，而且与第 1 名和第 2 名的在各指标上的效果相差不大。G-GLLE 在图像数据集 3DFE、Natural、SPO、SPOEM 4 个数据集上排名基本上都排在前 3 名。

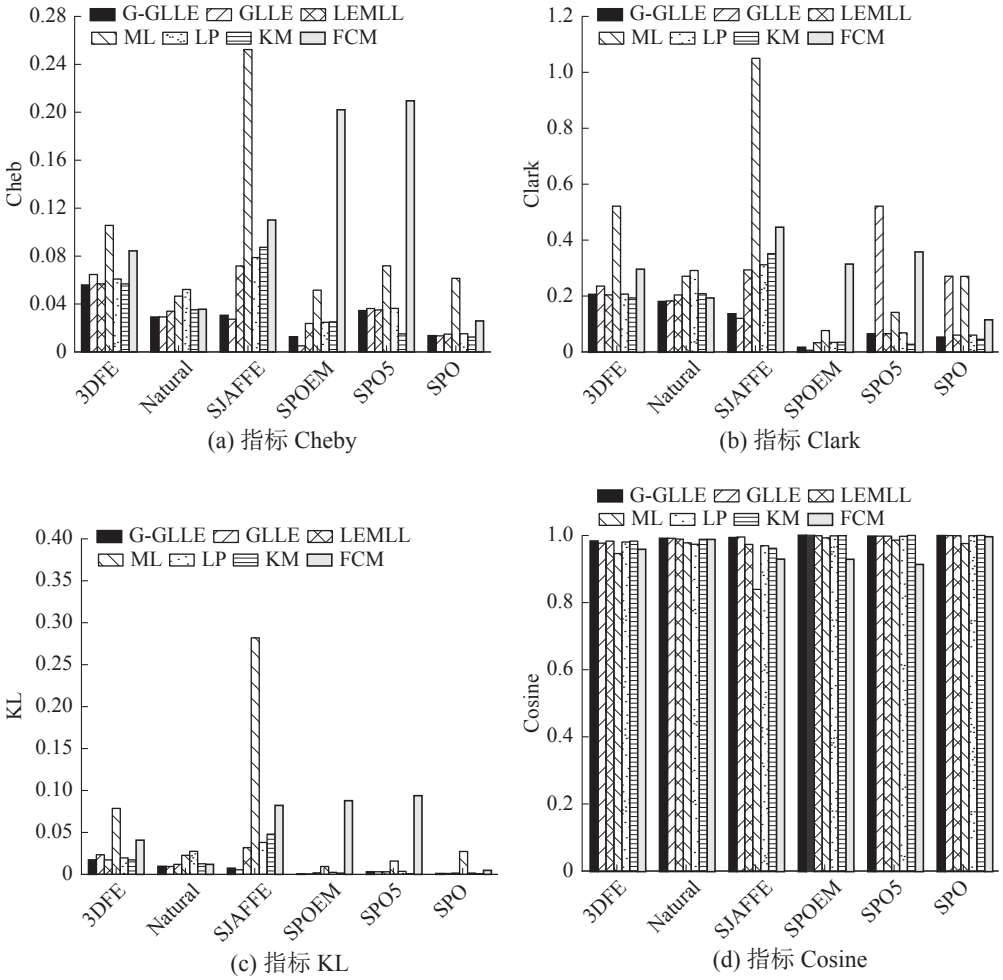
3.3.2　标记分布学习能力对比

为了进一步测试对逻辑标记数据集在标记增强预处理后作标记分布学习的有效性，我们从基于所选数据集的逻辑分布出发，经 7 个标记增强算法（G-GLLE、GLLE[9]、LEMLL[7]、ML[13]、LP[12]、KM[11]、FCM[10]）得到数值标记分别作为训练集，在标记分布学习基准算法 SA-BFGS[26] 上训练标记分布学习模型，并用所训练的 LDL 模型测试数据集标记分布学习效果。

预测实验的对比结果如图 2 所示。由 5 个对比图可以直观地看出，在所比较的 6 个数据集分别对应的 6 种数值标记恢复监督信息下，G-GLLE 算法恢复的标记分布对于提升 LDL 算法的运算效果整体上效果更好。特别是在 2 个图像数据集 3DFE 和 Natural 上优势更明显。

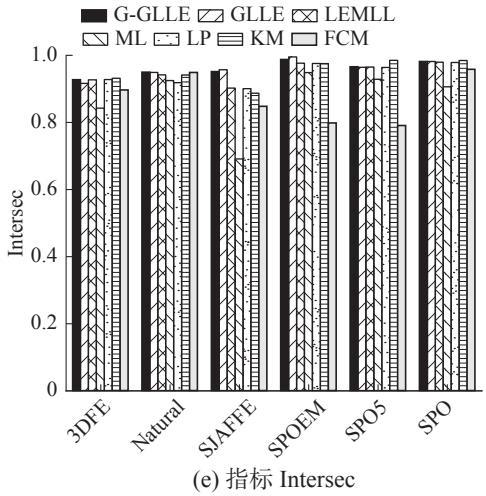
进一步对比 G-GLLE 和 GLLE 时发现，对于 3DFE、Natural Scene 和 Movie 数据集，从粒的角度挖掘数据特征的算法 G-GLLE 效果好。这可能是由于 3DFE、Natural Scene 和 Movie 数据集的标记局部相关性的不一致性明显，所以由标记局部相关性为依据划分知识粒的效果好，能较大可能性地划分出标记局部一致性的相关性的示例构成的粒。在各个一致性粒的基础上，粒间示例的标记局部相关差异性强，粒内示例的标记局部相关一致性强，此时对于同一粒上的示例求解关于示例的标记分布表示知识比较一致，可求得相似的参数，且符合平滑假设。

而在 SJAFFE 数据集上，2 个算法在 5 个指标上的效果相差不大，并且从示例挖掘标记特性的 GLLE 算法效果略好些。这可能是由于标记局部相关性的不一致性差异不大，各标记的重要性相似，使得通过诸如 *k*-means 聚类方式，较难得到有效的标记分布表示知识粒，当强制将示例划分为几类时，标记局部语义信息可能不完整，导致数据集的标记分布学习效果不佳。



(a) 指标 Cheby　(b) 指标 Clark

(c) 指标 KL　(d) 指标 Cosine



(e) 指标 Intersec

图 2　各算法在 5 个指标上的标记分布学习能力对比

Fig. 2　Comparisons on capability of label distribution learning on five indexes

4　结束语

现有的标记增强学习仅降低了大规模标记分布标注的必要性, 而没有进一步研究示例层次的局部相关性。本文提出的算法 G-GLLE 从粒的角度处理数据, 利用了标记分布的局部性信息, 提供了恢复多标记信息的新视角。已有实验结果证明本文方法能有效地挖掘标记局部相关性。未来可以在粒的构造上作进一步的研究, 设计不同的粒化策略以进一步优化标记分布的局部性形态, 提升标记分布学习的有效性。

参考文献:

[1] ZHANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2014, 26(8): 1819–1837.

[2] GIBAJA E, VENTURA S. A tutorial on multilabel learning[J]. ACM computing surveys, 2015, 47(3): 52.

[3] 耿新, 徐宁. 标记分布学习与标记增强 [J]. 中国科学:信息科学, 2018, 48(5): 521–530.
GENG Xin, XU Ning. Label distribution learning and label enhancement[J]. Scientia sinica (informationis), 2018, 48(5): 521–530.

[4] YAO Yiyang, WANG Luo, ZHANG Luming, et al. Learning latent stable patterns for image understanding with weak and noisy labels[J]. IEEE transactions on cybernetics, 2019, 49(12): 4243–4252.

[5] 曾雪强, 华鑫, 刘平生, 等. 基于情感轮和情感词典的文本情感分布标记增强方法 [J]. 计算机学报, 2021, 44(6): 1080–1094.
ZENG Xueqiang, HUA Xin, LIU Pingsheng, et al. Emotion wheel and lexicon based text emotion distribution label enhancement method[J]. Chinese journal of computers, 2021, 44(6): 1080–1094.

[6] GENG Xin. Label distribution learning[J]. IEEE transactions on knowledge and data engineering, 2016, 28(7): 1734–1748.

[7] SHAO Ruifeng, XU Ning, GENG Xin. Multi-label learning with label enhancement[C]//IEEE International Conference on Data Mining. Singapore: IEEE, 2018: 437-446.

[8] 熊传镇, 钱文彬, 王映龙. 基于标记增强和模糊辨识度的标记分布特征选择 [J]. 数据采集与处理, 2021, 36(3): 529-543.
XIONG Chuanzhen, QIAN Wenbin, WANG Yinglong. Label enhancement and fuzzy discernibility based label distribution feature selection[J]. Journal of data acquisition and processing, 2021, 36(3): 529-543.

[9] XU Ning, LIU Yunpeng, GENG Xin. Label enhancement for label distribution learning[J]. IEEE transactions on knowledge and data engineering, 2021, 33(4): 1632–1643.

[10] EL GAYAR N, SCHWENKER F, PALM G. A study of the robustness of KNN classifiers trained using soft labels[M]//Artificial Neural Networks in Pattern Recognition. Berlin: Springer Berlin, 2006: 67−80.

[11] JIANG Xiufeng, YI Zhang, LYU Jiancheng. Fuzzy SVM with a new fuzzy membership function[J]. Neural computing and applications, 2006, 15(3/4): 268–276.

[12] LI Yukun, ZHANG Minling, GENG Xin. Leveraging implicit relative labeling-importance information for effective multi-label learning[J]. 2015 IEEE international conference on data mining, 2015: 251−260.

[13] HOU Peng, GENG Xin, ZHANG Minling. Multi-label

manifold learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2016, 30(1): 1680–1686.

[14] YAO Jing tao, VASILAKOS A V, PEDRYCZ W. Granular computing: perspectives and challenges[J]. IEEE transactions on cybernetics, 2013, 43(6): 1977–1989.

[15] 刘清, 邱桃荣, 刘澜. 基于非标准分析的粒计算研究 [J]. 计算机学报, 2015, 38(8): 1618−1627.
LIU Qing, QIU Taorong, LIU Lan. The research of granular computing based on nonstandard analysis[J]. Chinese journal of computers, 2015, 38(8): 1618−1627.

[16] PEDRYCZ W. Granular computing for data analytics: a manifesto of human-centric computing[J]. IEEE/CAA journal of automatica sinica, 2018, 5(6): 1025–1034.

[17] 徐健锋, 苗夺谦, 张远健. 分段延迟代价敏感三支决策 [J]. 软件学报, 2022, 33(10): 3754–3775.
XU Jianfeng, MIAO Duoqian, ZHANG Yuanjian. Piecewise delay cost-sensitive three-way decisions[J]. Journal of software, 2022, 33(10): 3754–3775.

[18] 苗夺谦, 高阳, 吴伟志, 等. 粒计算与知识发现白皮书 [R]. 北京: 中国人工智能学会, 2022.
MIAO Duoqian, GAO Yang, WU Weizhi, et al. White paper for granular computing and knowledge discovery [R]. Beijing: Chinese Association for Artificial Intelligence, 2022.

[19] ZHANG Yuanjian, MIAO Duoqian, ZHANG Zhifei, et al. A three-way selective ensemble model for multi-label classification[J]. International journal of approximate reasoning, 2018, 103: 394–413.

[20] ZHANG Yuanjian, MIAO Duoqian, PEDRYCZ W, et al. Granular structure-based incremental updating for multi-label classification[J]. Knowledge-based systems, 2020, 189: 105066.

[21] ZHANG Yuanjian, ZHAO Tianna, MIAO Duoqian, et al. Granular multilabel batch active learning with pairwise label correlation[J]. IEEE transactions on systems, man, and cybernetics:systems, 2022, 52(5): 3079–3091.

[22] HUANG Shengjun, ZHOU Zhihua. Multi-label learning by exploiting label correlations locally[C]//AAAI'12: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. New York: ACM, 2012: 949−955.

[23] ZHENG Xiang, JIA Xiuyi, LI Weiwei. Label distribution learning by exploiting sample correlations locally[J]. Proceedings of the AAAI conference on artificial intelligence, 2018, 32(1): 3310–3317.

[24] JIA Xiuyi, LI Zechao, ZHENG Xiang, et al. Label distribution learning with label correlations on local samples[J]. IEEE transactions on knowledge and data engineering, 2021, 33(4): 1619–1631.

[25] ZHANG Jujie, FANG Min, LI Xiao. Clustered intrinsic label correlations for multi-label classification[J]. Expert systems with applications, 2017, 81: 134–146.

[26] GENG Xin, YIN Chao, ZHOU Zhihua. Facial age estimation by learning from label distributions[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(10): 2401–2412.

作者简介：

张远健, 博士, 中国银联股份有限公司博士后, 中国计算机学会会员, 主要研究方向为多标记分类、粒计算、联邦学习, 主持中国博士后面上基金 1 项。发表学术论文 10 余篇。

赵天娜, 博士研究生, 中国人工智能学会会员, 主要研究方向为标记分布学习、粒计算、不确定性。发表学术论文 7 篇。

苗夺谦, 教授, 博士, 国际粗糙集学会理事长、中国人工智能学会会士、中国计算机学会杰出会员, 主要研究方向为粒计算、不确定性、大数据分析。荣获中国人工智能学会吴文俊人工智能自然科学二等奖 1 项; 主持国家自然科学基金面上项目 7 项, 出版教材和学术著作 10 余部。发表学术论文 180 余篇, ESI 高被引 8 篇。