中国科学: 信息科学 2018 年 第 48 卷 第 5 期: 521–530
SCIENTIA SINICA Informationis
人工智能优青专刊 · 论文

标记分布学习与标记增强

耿新[1,2*], 徐宁[1,2]

1. 东南大学计算机科学与工程学院, 南京 211189
2. 计算机网络和信息集成教育部重点实验室 (东南大学), 南京 211189
* 通信作者. E-mail: xgeng@seu.edu.cn

收稿日期: 2018–02–06; 接受日期: 2018–04–11; 网络出版日期: 2018–05–11

国家重点研发计划项目 (批准号: 2017YFB1002801)、国家自然科学基金优秀青年科学基金项目 (批准号: 61622203)、江苏省杰出青年基金项目 (批准号: BK20140022) 资助, 并受到软件新技术与产业化协同创新中心和无线通信技术协同创新中心支持

摘要 本文主要介绍了标记分布学习和标记增强的相关概念及算法. 标记分布学习是一种新型机器学习范式, 传统的单标记和多标记学习都可以看做是该范式的特例. 标记分布学习将不同标记对示例的重要程度用标记分布来显式刻画, 已经在多个应用领域中取得很好的效果. 然而, 现有的多数数据集中却仅具有简单的逻辑标记而非完整的标记分布, 因此无法直接应用标记分布学习. 为解决这一问题, 可以通过挖掘训练集中蕴含的标记重要性信息, 恢复出每个示例的标记分布. 我们将原始逻辑标记提升为标记分布的过程定义为标记增强. 本文给出了标记分布学习和标记增强的形式化定义, 介绍了典型标记分布学习和标记增强算法, 并对这些算法进行了分析讨论.

关键词 标记分布, 标记分布学习, 标记增强, 多标记学习, 标记多义性

1 引言

标记多义性问题是机器学习领域的热门研究方向. 在现有的机器学习范式中, 主要存在两种数据标注方式: (1) 一个示例分配一个标记; (2) 一个示例分配多个标记. 单标记学习 (single-label learning, SLL) 假设训练集中所有的示例均用第 1 种方式标注, 而多标记学习 (multi-label learning, MLL) [1] 允许训练示例用第 2 种方式标注. 因此, 多标记学习可以处理一个示例属于多个类别的多义性情况. 无论是单标记学习还是多标记学习, 都旨在回答一个本质的问题, 即 “哪些标记可以描述该示例?”. 然而, 它们都没有直接回答另一个更深一层的问题 “每个标记如何描述该示例?”, 即每个标记对该示例的相对重要程度如何?

对于真实世界中的许多问题, 不同标记的重要程度往往是不同的, 例如, 一幅自然场景图像 [2] 被标注了 “天空”、“水”、“森林” 和 “云” 等多个标记, 而这些标记具体描述该图像的程度却是不同的; 在人脸情感分析 [3] 中, 人的面部表情常常是多种基础情感 (如快乐、悲伤、惊讶、愤怒、厌恶和恐惧)

引用格式: 耿新, 徐宁. 标记分布学习与标记增强. 中国科学: 信息科学, 2018, 48: 521–530, doi: 10.1360/N112018-00029
Geng X, Xu N. Label distribution learning and label enhancement (in Chinese). Sci Sin Inform, 2018, 48: 521–530, doi: 10.1360/N112018-00029

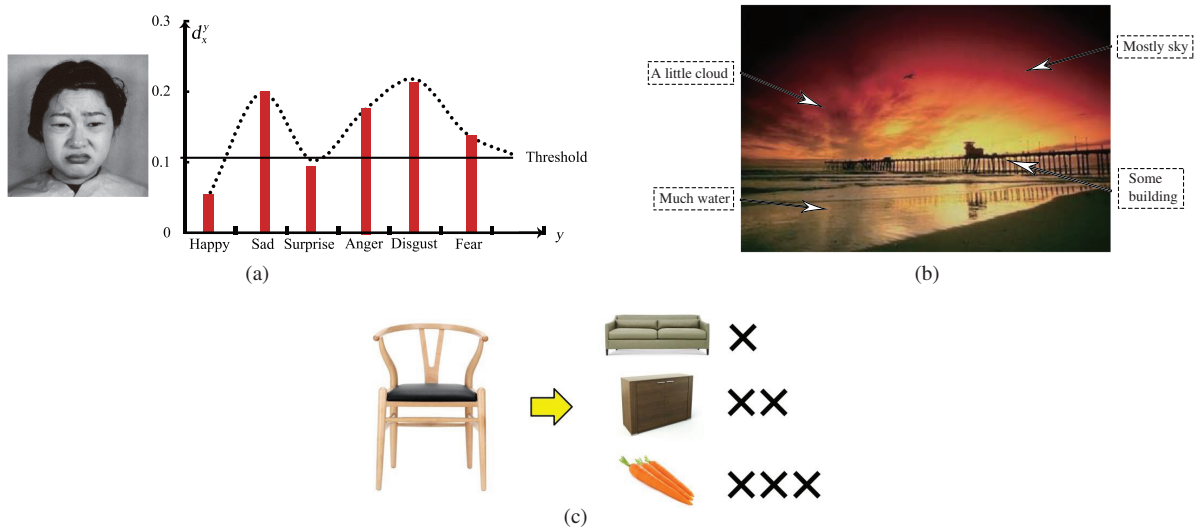


图 1　(网络版彩图) 标记分布的普适性

Figure 1　(Color online) Pervasiveness of label distribution. (a) Relevant/irrelevant labels; (b) relevant labels; (c) irrelevant labels

混合的结果, 而这些基础情感在一个具体的表情中常常表达出不同的强度, 从而呈现出纷繁复杂的情感. 类似的例子还有很多, 因为一旦一个示例与多个标记同时相关, 那么这些标记一般情况下对该示例来说不会恰好都一样重要, 而更可能会有主次先后之分. 对于类似上述例子的应用, 一种很自然的方法是对于一个示例 $\boldsymbol{x}$, 将一个实数 $d_{\boldsymbol{x}}^y$ 赋予每个可能的标记 y, 表示 y 描述 $\boldsymbol{x}$ 的程度. 不失一般性, 假设 $d_{\boldsymbol{x}}^y \in [0,1]$, 并进一步假设标记集合为完备集, 即用集合中的所有标记一定可以完整地描述一个示例, 所以 $\sum_y d_{\boldsymbol{x}}^y = 1$. 满足以上两个条件的 $d_{\boldsymbol{x}}^y$ 称为 y 对 $\boldsymbol{x}$ 的描述度. 对一个示例, 所有标记的描述度构成一种类似概率分布的数据结构, 所以被称为标记分布, 而在以标记分布标注的数据集上学习的过程就称为标记分布学习 (label distribution learning, LDL) [4].

事实上, 标记分布在许多监督学习问题中具有一定的普适性, 这是因为标记与示例的相关或不相关常常是相对的, 这具体体现在如下 3 个层面.

(1) 相关与不相关的划分是相对的. 将标记集合分为相关和不相关两个子集往往是对实际问题的一种简化, 而事实上两者之间的 "边界" 可能并没有那么明确. 例如, 在图 1(a) 中, 一个人脸面部表情表达了几种不同强度的基础情感的混合情感. 为了区分相关标记和不相关标记, 需要人为设定一个阈值 y_0, 强度高于 y_0 的情感就被认为是相关标记, 而相反则被认为是不相关标记. 显然这一划分过程依赖于 y_0 的选择, 并非一种绝对客观的划分. 而这一划分过程也会导致基础感情强度信息的丢失.

(2) 相关标记的 "相关性" 是相对的. 当一个示例与多个标记相关时, 这些标记对于这个示例的重要程度一般情况下不会恰好完全一样. 例如, 在图 1(b) 中, 对一副自然场景图像来说, "天空"、"水"、"建筑" 和 "云" 都是相关标记, 但是它们各自对该图像的描述程度却是不同的, 即 "相关性" 不同.

(3) 不相关标记的 "不相关性" 也是相对的. 对一个示例来说, 众多不相关标记的 "不相关性" 可能差异明显. 例如, 在图 1(c) 中, 对于一把椅子来说, 沙发虽然是个不相关标记, 但是由于功能和结构的相似, 它们的 "不相关性" 较小. 相对地, 柜子与椅子的 "不相关性" 就要大一些. 而胡萝卜相比沙发和柜子, 其与椅子的 "不相关性" 是最大的.

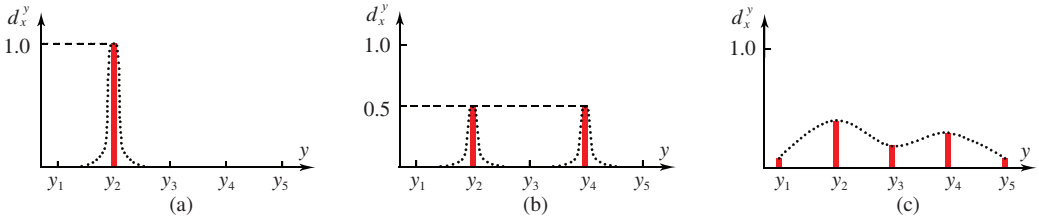


图 2 (网络版彩图) 单标记、多标记与一般情况下的标记分布

Figure 2 (Color online) Example label distributions for (a) single-label, (b) multi-label, and (c) the general case

如果用逻辑值 “1” 和 “0” 分别表示相关和不相关的话, 这 3 个层面的相对性分别对应为: “1” 和 “0” 之间的区分并不绝对, “1” 和 “1” 之间本质上是有差异的, “0” 和 “0” 之间本质上也是有差异的. 正是由于相关标记与不相关标记的这种相对性, 使得很多监督学习问题本质上更适合用标记分布来描述.

然而, 实践中直接获得每个标记的描述度在许多应用中并不现实, 这是因为, 一方面, 逐个考量标记对示例的描述度代价高昂, 另一方面, 每个标记的描述度也往往没有客观的量化标准. 目前普遍数据标注方法是, 对于一个示例 $\boldsymbol{x}$, 将 $l_{\boldsymbol{x}}^y \in \{0,1\}$ 赋予每个可能的标记 y, 如果 $l_{\boldsymbol{x}}^y = 1$ 则表示 y 是 $\boldsymbol{x}$ 的相关标记, 如果 $l_{\boldsymbol{x}}^y = 0$ 则表示 y 是 $\boldsymbol{x}$ 的不相关标记. $l_{\boldsymbol{x}}^y$ 表达了是与否的逻辑关系, 所以对一个示例来讲, 所有标记的逻辑值 $l_{\boldsymbol{x}}^y$ 构成的逻辑向量被称为逻辑标记. 多数现有数据使用逻辑标记作为示例的监督信息, 其实是对问题的简化. 尽管如此, 这些数据中的监督信息本质上却是遵循某种标记分布的. 这种标记分布虽然没有显式给出, 却有可能通过对数据集的分析将其恢复出来, 这一过程就称为标记增强. 标记增强是指将训练样本中的原始逻辑标记转化为标记分布的过程. 与基于嵌入 (embedding) 的多标记分类方法 [5] 相似, 标记增强也同样依赖于对隐含在训练样本中的标记相关信息的挖掘. 假设 $\mathcal{Y}$ 表示样本的原始逻辑标记空间, $\mathcal{D}$ 表示经过标记增强后的标记分布空间, 那么, 标记增强将原始的标记空间 $\mathcal{Y} = \{0,1\}^c$ 拓展为 $\mathcal{D} = [0,1]^c$, 其中 c 表示标记的个数. 事实上, $\mathcal{D}$ 构成 c 维欧式空间中的一个超立方体, 而 $\mathcal{Y}$ 仅位于该超立方体的顶点. 标记增强利用隐含于数据中的标记间相关性, 可以有效加强示例的监督信息, 进而通过标记分布学习获得更好的预测效果.

本文剩余部分组织如下: 第 2 节首先给出标记分布学习的定义, 然后介绍几种具有代表性的标记分布学习算法. 接着, 在第 3 节给出标记增强的定义, 并介绍标记增强的几种典型算法. 最后, 在第 4 节讨论了本文介绍的各种标记分布学习算法和标记增强算法的优点与缺点, 随后给出本文结论, 并指出未来的研究方向.

2 标记分布学习

2.1 概念定义

首先, 本文主要的符号表示如下. 示例用 $\boldsymbol{x}$ 表示, 第 i 个示例用 $\boldsymbol{x}_i$ 表示; 标记用 y 表示, 第 j 个标记用 y_j 表示. y_j 对于 $\boldsymbol{x}_i$ 的描述度用 $d_{\boldsymbol{x}_i}^y$ 表示, 其满足 $d_{\boldsymbol{x}_i}^y \in [0,1]$ 并且 $\sum_y d_{\boldsymbol{x}_i}^y = 1$. $\boldsymbol{x}_i$ 的标记分布用 $\boldsymbol{d}_i = [d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \ldots, d_{\boldsymbol{x}_i}^{y_c}]$ 表示, 所以 $\boldsymbol{d}_i \in [0,1]^c$, 其中 c 是可能的标记数目.

传统的单标记和多标记标注在这一定义下都可以看作标记分布的特例. 图 2 给出了单标记、多标记, 以及一般情况下标记分布的例子. 具体来说, 对于单标记标注, 如图 2(a) 所示例子, 这时只有一个相关标记 y_2, 因此 $d_{\boldsymbol{x}}^{y_2} = 1$, 而其他所有标记的描述度均为 0. 对于多标记标注, 如图 2(b) 所示例子,

两个相关标记 y_2 和 y_4 在没有额外信息的情况下只能假设其重要程度相等, 所以 $d_{\boldsymbol{x}}^{y_2} = d_{\boldsymbol{x}}^{y_4} = 0.5$, 而其他所有标记的描述度均为 0. 最后, 图 2(c) 给出了一般情况下标记分布的一个例子, 其仅需满足条件 $d_{\boldsymbol{x}}^{y} \in [0,1]$ 并且 $\sum_y d_{\boldsymbol{x}}^{y} = 1$. 通过这些例子可以看出, 单标记和多标记标注都可以看作标记分布的特例, 标记分布比传统示例标注方式更加通用, 因此可以为机器学习提供更多的灵活性. 值得注意的是, 标记分布学习与多输出回归具有一定联系. 具体地, 如果多输出回归满足标记分布的限定条件, 即 $d_{\boldsymbol{x}_i}^{y} \in [0,1]$ 且 $\sum_y d_{\boldsymbol{x}_i}^{y} = 1$, 那么多输出回归就成为了标记分布学习. 因此, 标记分布学习可以看作多输出回归的特例.

因为标记分布与概率分布满足相同的约束条件, 因此标记分布学习可以借用很多统计学的理论和方法. 首先, 描述度可以用条件概率的形式来表示, 即 $d_{\boldsymbol{x}}^{y} = P(y|\boldsymbol{x})$. 那么标记分布学习可以描述如下:

假设 $\mathcal{X} = \mathbb{R}^q$ 表示示例的特征空间, $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ 表示标记空间. 给定一个训练集 $S = \{(\boldsymbol{x}_1, \boldsymbol{d}_1), (\boldsymbol{x}_2, \boldsymbol{d}_2), \ldots, (\boldsymbol{x}_n, \boldsymbol{d}_n)\}$, 标记分布学习的目标是从 S 中学习得到一个条件概率质量函数 $p(y|\boldsymbol{x})$, 其中 $\boldsymbol{x} \in \mathcal{X}$ 且 $y \in \mathcal{Y}$.

假设 $p(y|\boldsymbol{x})$ 的参数模型表示为 $p(y|\boldsymbol{x};\boldsymbol{\theta})$, 其中 $\boldsymbol{\theta}$ 是参数向量. 给定训练集 S, 标记分布学习的目标是找到一个 $\boldsymbol{\theta}$, 使得给定示例 $\boldsymbol{x}_i$, $p(y|\boldsymbol{x};\boldsymbol{\theta})$ 能生成与 $\boldsymbol{x}_i$ 的真实标记分布 $\boldsymbol{d}_i$ 尽可能相似的标记分布. 如果使用 Kullback-Leibler 散度来度量两个分布之间距离的话, 那么最佳的参数 $\boldsymbol{\theta}$ 为:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_i \sum_j \left(d_{\boldsymbol{x}_i}^{y_j} \ln \frac{d_{\boldsymbol{x}_i}^{y_j}}{p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta})} \right) = \arg\max_{\boldsymbol{\theta}} \sum_i \sum_j d_{\boldsymbol{x}_i}^{y_j} \ln p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta}). \tag{1}$$

有了式 (1) 中的优化目标, 首先回望一下传统的单标记和多标记学习这两个特例在这一优化目标下会得到什么结果. 对于单标记学习, $d_{\boldsymbol{x}}^{y} = \mathrm{Kr}(y_j, y(\boldsymbol{x}_i))$, 这里 $\mathrm{Kr}(\cdot,\cdot)$ 是 Kronecker delta 函数, $y(\boldsymbol{x}_i)$ 是 $\boldsymbol{x}_i$ 的单标记. 这时, 式 (1) 可以简化为

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_i \ln p(y(\boldsymbol{x}_i)|\boldsymbol{x}_i;\boldsymbol{\theta}), \tag{2}$$

这实际上是 $\boldsymbol{\theta}$ 的极大似然估计 (maximum likelihood, ML), 而后面使用 $p(y|\boldsymbol{x};\boldsymbol{\theta})$ 进行分类等价于最大后验决策 (maximum a posteriori, MAP). 对于多标记学习, 每个示例 $\boldsymbol{x}_i$ 使用一个标记集合 Y_i 来标注, 因此,

$$d_{\boldsymbol{x}_i}^{y_j} = \begin{cases} \dfrac{1}{|Y_i|}, & y_j \in Y_i, \\ 0, & y_j \notin Y_i. \end{cases}$$

此时, 式 (1) 变化为

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_i \frac{1}{|Y_i|} \sum_{y \in Y_i} \ln p(y|\boldsymbol{x}_i;\boldsymbol{\theta}), \tag{3}$$

而式 (3) 可以看作是使用相关标记集合的势的倒数进行加权的极大似然估计. 实际上, 这等价于首先采用一种基于熵的标记分配方法 (entropy-based label assignment, ELA) [1] 将多标记数据转化为加权的单标记数据, 然后再用极大似然估计来估计参数 $\boldsymbol{\theta}$. 通过以上分析可以看出, 在适当的约束条件下, 标记分布学习算法可以转化为常见的单标记或多标记学习算法. 因此, 标记分布学习可以看作一个更加通用的学习框架, 包含了作为特例的单标记和多标记学习. 此外, 标记分布学习是一个较为灵活的学习框架, 例如文献 [6] 使用了深度神经网络, 使得标记分布学习达到了更好的效果.

2.2 标记分布学习算法

可以依循 3 种策略为标记分布学习设计算法. 第 1 种策略是问题转化, 即将标记分布学习问题转化为传统的单标记或多标记学习问题. 根据该策略我们提出两种算法, 分别是 PT-Bayes 和 PT-SVM, 其中 "PT" 表示 "问题转化 (problem transformation)". 第 2 种策略是算法改造, 即将传统单标记或多标记学习算法改造为能够处理标记分布数据的学习算法. 根据该策略我们也提出两种算法, 分别是 AA-kNN 和 AA-BP, 其中 "AA" 表示 "算法改造 (algorithm adaptation)". 第 3 种策略是根据标记分布学习本身固有的特性而设计的专用算法, 本小节同样介绍两种专用算法, 分别是 SA-IIS 和 SA-BFGS, 其中 "SA" 表示 "专用算法 (specialized algorithm)".

2.2.1 "问题转化" 算法

将标记分布学习问题转化为单标记学习问题的一个直接方法, 就是将训练样本转化成加权的单标记样本. 具体地, 将每个训练样本 $(\boldsymbol{x}_i, \boldsymbol{d}_i)$ 转化成 c 个单标记样本 $(\boldsymbol{x}_i, y_j)$, 每个样本的权值为 $d_{\boldsymbol{x}_i}^{y_j}$. 然后依据每个样本的权值, 对训练集进行重采样. 经过重采样的训练集转化成一个含有 $c \times n$ 个样本的标准单标记训练集, 任何单标记学习算法都能应用于这个训练集上. 为了预测示例 $\boldsymbol{x}$ 的标记分布, 分类器必须能够输出每个标记 y_j 的描述度, 即 $d_{\boldsymbol{x}}^{y_j} = P(y_j|\boldsymbol{x})$. 这里, 可以采用两个经典算法, 即 Bayes 分类器和 SVM, 分别记为 PT-Bayes 和 PT-SVM. 具体地, Bayes 分类器假设每个类服从 Gauss 先验分布, 由此计算出的后验概率即为对应标记的描述度. 对于 SVM, 每个标记的概率估计通过一种逐对耦合多类方法 [7] 得到, 每个二值向量机的概率都是用改进的 Platt 后验概率 [8] 计算得到.

2.2.2 "算法改造" 算法

某些传统算法可以自然地扩展为能够处理标记分布的算法, 这里提出两种改造算法. 第 1 个从 k-NN 改造而来, 即 AA-kNN. 给定一个新的示例 $\boldsymbol{x}$, 首先在训练集中找出 $\boldsymbol{x}$ 的 k 近邻. 接着, 将 k 个近邻的标记分布的均值作为对 $\boldsymbol{x}$ 的标记分布预测. 第 2 种算法是由反向传播 (backpropagation, BP) 神经网络算法改造而来, 即 AA-BP. 假设 3 层前馈神经网络有 q ($\boldsymbol{x}$ 的维度) 个输入单元, c (标记的个数) 个输出单元, 每个输出单元输出标记 y_j 的描述度. 于是 BP 算法的目标就是最小化这个真实标记分布和神经网络输出的标记分布之间误差的平方和.

2.2.3 专用算法

与问题转化和算法改造这两种间接策略相比, 专用算法与标记分布问题更加匹配, 比如直接求解式 (1) 中的优化问题. 这里介绍两种专用算法, 即 SA-IIS 和 SA-BFGS, 其中关键一步就是解决如式 (1) 中的优化问题.

假定 $p(y|\boldsymbol{x};\boldsymbol{\theta})$ 为最大熵模型 [9], 即

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z}\exp\left(\sum_k \theta_{y,k} g_k(\boldsymbol{x})\right), \tag{4}$$

其中 $Z = \sum_y \exp(\sum_k \theta_{y,k} g_k(\boldsymbol{x}))$ 为归一化因子, $\theta_{y,k}$ 是 $\boldsymbol{\theta}$ 中的元素, $g_k(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的第 k 个分量. 将式 (4) 代入式 (1) 中, 得到 $\boldsymbol{\theta}$ 的目标函数:

$$T(\theta) = \sum_{i,j} d_{\boldsymbol{x}_i}^{y_j} \sum_k \theta_{y_j,k} g(\boldsymbol{x}_i) - \sum_i \ln \sum_j \exp\left(\sum_k \theta_{y_j,k} g_k(\boldsymbol{x}_i)\right). \tag{5}$$

SA-IIS 使用了一种类似改进迭代尺度算法 (improved iterative scaling, IIS) [10] 的方法对式 (5) 进行优化. 而 SA-BFGS 基于拟 Newton 法 BFGS [11], 进一步改善了 SA-IIS 算法, 将目标函数的优化与一阶梯度函数相关联, 比标准 Newton 线性搜索方法的效率更高.

3 标记增强

3.1 概念定义

$\boldsymbol{x}_i$ 的逻辑标记用 $\boldsymbol{l}_i = [l_{\boldsymbol{x}_i}^{y_1}, l_{\boldsymbol{x}_i}^{y_2}, \ldots, l_{\boldsymbol{x}_i}^{y_c}]$ 表示, 其中 $l_{\boldsymbol{x}_i}^{y_j} \in \{0,1\}$ 表示 y_j 是否是 $\boldsymbol{x}_i$ 的相关标记, c 是可能的标记数目, 则 $\boldsymbol{l}_i \in \{0,1\}^c$. 假设 $\mathcal{X} = \mathbb{R}^q$ 表示示例的特征空间, $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ 表示标记空间, 则面向标记分布学习的标记增强定义如下:

给定训练集 $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{l}_i) | 1 \leqslant i \leqslant n\}$, 标记增强即将每个示例 $\boldsymbol{x}_i$ 的逻辑标记 $\boldsymbol{l}_i$ 转化为相应的标记分布 $\boldsymbol{d}_i$, 从而得到标记分布训练集 $\mathcal{E} = \{(\boldsymbol{x}_i, \boldsymbol{d}_i) | 1 \leqslant i \leqslant n\}$ 的过程.

3.2 标记增强算法

3.2.1 基于模糊方法的标记增强

基于模糊方法的标记增强 [12~15] 利用模糊数学的思想, 通过模糊聚类、模糊运算和核隶属度等方法, 挖掘出标记间相关信息, 将逻辑标记转化为标记分布. 值得注意的是, 这类方法提出的目的一般是为了将模糊性引入原本刚性的逻辑标记, 而并未明确其可以将逻辑标记增强为标记分布. 但是, 很多模糊标记增强方法实际上可以基于模糊隶属度轻松生成标记分布. 本小节介绍两种基于模糊方法的标记增强算法, 分别是基于模糊聚类的标记增强算法和基于核隶属度的标记增强算法.

基于模糊聚类的标记增强 [12] 通过模糊 C – 均值聚类 (fuzzy C-means algorithm, FCM) [16] 和模糊运算, 将训练集中每个示例的逻辑标记转化为相应的标记分布, 从而得到标记分布训练集. FCM 是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法, 该算法把 n 个样本分为 p 个模糊聚类, 并求每个聚类的中心, 使得所有训练样本到聚类中心的加权 (权值由样本点对相应聚类的隶属度决定) 距离之和最小. 假设 FCM 将训练集 $\mathcal{S}$ 分成 p 个聚类, $\boldsymbol{\mu}_k$ 表示第 k 个聚类的中心, 则可用如下公式计算示例 $\boldsymbol{x}_i$ 对于每个聚类的隶属度 $\boldsymbol{m}_{\boldsymbol{x}_i} = [m_{\boldsymbol{x}_i}^1, m_{\boldsymbol{x}_i}^2, \ldots, m_{\boldsymbol{x}_i}^p]$:

$$m_{\boldsymbol{x}_i}^k = \frac{1}{\sum_{j=1}^{p} \left(\frac{\mathrm{Dist}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)}{\mathrm{Dist}(\boldsymbol{x}_i, \boldsymbol{\mu}_j)} \right)^{\frac{1}{\beta - 1}}}, \tag{6}$$

其中, Dist 是任意的距离度量, β 是模糊因子, 且满足 $\beta > 1$. 得到 $\boldsymbol{m}_{\boldsymbol{x}_i}$ 后, 进一步构建一个关联矩阵 $\boldsymbol{A}$. 首先初始化一个 $c \times p$ 的零矩阵 $\boldsymbol{A}$, 然后用如下公式更新 $\boldsymbol{A}$ 的第 j 行 $\boldsymbol{A}_j$:

$$\boldsymbol{A}_j = \boldsymbol{A}_j + \boldsymbol{m}_{\boldsymbol{x}_i}, \quad \text{if } \ l_{\boldsymbol{x}_i}^{y_j} = 1, \tag{7}$$

即 $\boldsymbol{A}_j$ 为所有属于第 j 个类的样本的隶属度向量之和. 经过行归一化后得到的矩阵 $\boldsymbol{A}$ 可以被当作一个 “模糊关系” 矩阵, 即 $\boldsymbol{A}$ 中的元素 a_{jk} 表示了第 j 个类别 (标记) 与第 k 个聚类的关联强度. 根据模糊逻辑推理机制 [16], 将关联矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{x}_i$ 对聚类的隶属度 $\boldsymbol{m}_{\boldsymbol{x}_i}$ 进行模糊合成运算 $\boldsymbol{v}_i = \boldsymbol{A} \circ \boldsymbol{m}_{\boldsymbol{x}_i}$, 从而将 $\boldsymbol{x}_i$ 对聚类的隶属度转化为对类别的隶属度. 最后, 对隶属度向量 $\boldsymbol{v}_i$ 进行归一化, 使向量中元素的和为 1, 即得到标记分布 $\boldsymbol{d}_i$. 基于模糊聚类的标记增强算法利用模糊聚类过程中产生的示例对每个

聚类的隶属度, 通过类别和聚类的关联矩阵, 将示例对聚类的隶属度转化为对类别的隶属度, 从而生成标记分布. 在这一过程中, 模糊聚类反映了示例空间的拓扑关系, 而通过关联矩阵, 将这种关系转化到标记空间, 从而有可能使得简单的逻辑标记产生更丰富的语义, 转变为标记分布.

基于核隶属度的标记增强方法源于一种模糊支持向量机中核隶属度的生成过程 [13], 通过一个非线性映射函数将示例 $\boldsymbol{x}_i$ 映射到高维空间, 利用核函数计算该高维空间中正类的中心、半径和各示例 $\boldsymbol{x}_i$ 到正类中心的距离, 进而通过隶属度函数计算示例 $\boldsymbol{x}_i$ 的标记分布. 具体的, 对于训练集 $\mathcal{S}$ 和某个标记 y_j, 根据 y_j 的逻辑值, 将 $\mathcal{S}$ 分为两个集合, 其中 $\boldsymbol{x}_i$ 的逻辑标记 $l_{\boldsymbol{x}_i}^{y_j}=1$ 的集合用 $C_+^{y_j}$ 表示. 那么, 正类集合在特征空间的中心为 $\boldsymbol{\Psi}_+^{y_j}=\frac{1}{n_+}\sum_{\boldsymbol{x}_i\in C_+^{y_j}}\varphi(\boldsymbol{x}_i)$, 这里 n_+ 表示该集合中示例的数量, $\varphi(\boldsymbol{x}_i)$ 是一个非线性映射函数, 由核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)$ 确定. 该集合的半径定义为 $r_+=\max\|\boldsymbol{\Psi}_+^{y_j}-\varphi(\boldsymbol{x}_i)\|$, 集合中的 $\boldsymbol{x}_i$ 到中心的距离是 $d_{i+}=\|\varphi(\boldsymbol{x}_i)-\Psi_+^{y_j}\|$. 那么, $\boldsymbol{x}_i$ 对于标记 y_j 的隶属度为

$$m_{\boldsymbol{x}_i}^{y_j}=\begin{cases}1-\sqrt{\dfrac{d_{i+}^2}{(r_+^2+\delta)}}, & \text{if } l_{\boldsymbol{x}_i}^{y_j}=1,\\ 0, & \text{if } l_{\boldsymbol{x}_i}^{y_j}=0,\end{cases}\tag{8}$$

其中 $\delta>0$, 涉及 $\varphi(\boldsymbol{x}_i)$ 的计算均可以由核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 间接计算. 最后, 将 $m_{\boldsymbol{x}_i}^{y_j}$ 归一化, 即可得到 $\boldsymbol{x}_i$ 的标记分布 $\boldsymbol{d}_i$. 基于核隶属度的标记增强算法利用核技巧在高维空间中计算示例对每个类别的隶属度, 从而能够挖掘训练数据中类别标记间较为复杂的非线性关系.

3.2.2 基于图的标记增强

基于图的标记增强算法用图模型表示示例间的拓扑结构, 通过引入一些模型假设, 建立示例间相关性与标记间相关性之间的关系, 进而将示例的逻辑标记增强为标记分布. 本小节介绍两种基于图模型的标记增强算法, 分别是基于标记传播的标记增强算法和基于流形的标记增强算法.

基于标记传播的标记增强 [17] 将半监督学习 [18] 中的标记传播技术应用于标记增强中. 该方法首先根据示例间相似度构建一个图, 然后根据图中的拓扑关系在示例间传播标记. 由于标记的传播会受到路径上权值的影响, 会自然形成不同标记的描述度差异. 当标记传播收敛时, 每个示例的原有逻辑标记即可增强为标记分布. 具体的, 假设多标记训练集 $\mathcal{S}$ 中, $G=\langle V,E\rangle$ 表示以 $\mathcal{S}$ 中的示例为顶点的全连通图, 其中 V 表示顶点的集合, E 表示顶点两两之间的边的集合, $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的边上的权值为它们之间的相似度:

$$a_{ij}=\begin{cases}\exp\left(-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2}\right), & \text{if } i\neq j,\\ 0, & \text{if } i=j,\end{cases}\tag{9}$$

所有边的权值构成相似度矩阵 $\boldsymbol{A}=[a_{ij}]_{n\times n}$. 标记传播矩阵 $\boldsymbol{P}$ 由相似度矩阵计算 $\boldsymbol{P}=\hat{\boldsymbol{A}}^{-\frac{1}{2}}\boldsymbol{A}\hat{\boldsymbol{A}}^{-\frac{1}{2}}$, 这里 $\hat{\boldsymbol{A}}$ 是一个对角矩阵, 其中 $\hat{a}_{ii}=\sum_{j=1}^n a_{ij}$. 假设所有标记对所有示例的描述度构成一个描述度矩阵 $\boldsymbol{F}$, 该算法使用迭代方法不断更新 $\boldsymbol{F}$. $\boldsymbol{F}$ 的初始值 $\boldsymbol{F}^0=\boldsymbol{\Phi}=[\phi_{ij}]_{n\times c}$ 由示例 $\boldsymbol{x}_i$ 的逻辑标记构成, 即 $\forall_{i=1}^n\forall_{j=1}^c:\phi_{ij}=l_{\boldsymbol{x}_i}^{y_j}$. 在此基础上, 使用标记传播对描述度矩阵进行更新,

$$\boldsymbol{F}^{(t)}=\alpha\boldsymbol{P}\boldsymbol{F}^{(t-1)}+(1-\alpha)\boldsymbol{\Phi},\tag{10}$$

其中, α 是平衡参数, 控制了初始的逻辑标记和标记传播对最终描述度的影响程度. 经过迭代, 最终 $\boldsymbol{F}$ 收敛到 $\boldsymbol{F}^*=(1-\alpha)(I-\alpha\boldsymbol{P})^{-1}\boldsymbol{\Phi}$. 经过归一化处理 $\forall_{i=1}^n\forall_{j=1}^c:d_{\boldsymbol{x}_i}^{y_j}=\frac{f_{ij}^*}{\sum_{k=1}^c f_{ik}^*}$, 即得到示例 $\boldsymbol{x}_i$ 的标记

分布 $\boldsymbol{d}_i$. 基于标记传播的标记增强算法通过图模型表示示例间的拓扑结构, 构造了基于示例间相关性的标记传播矩阵, 利用传播过程中路径权值的不同使得不同标记的描述度自然产生差异, 从而反映出蕴含在训练数据中的标记间关系.

基于流形的标记增强算法 [19] 假设数据在特征空间和标记空间均分布在某种流形上, 并利用平滑假设将两个空间的流形联系起来, 从而可以利用特征空间流形的拓扑关系指导标记空间流形的构建, 在此基础上将示例的逻辑标记增强为标记分布. 具体的, 该算法用图 $\boldsymbol{G}=\langle V,E,\boldsymbol{W}\rangle$ 表示多标记训练集 $\mathcal{S}$ 的特征空间的拓扑结构, 其中 V 是由示例构成的顶点集合, E 是边的集合, $\boldsymbol{W}$ 是图的边权重矩阵. 首先, 在特征空间中, 假设示例分布的流形满足局部线性, 即任意示例 $\boldsymbol{x}_i$ 可以由它的 k – 近邻的线性组合重构, 重构权值矩阵 $\boldsymbol{W}$ 可通过最小化下式得到:

$$\Omega(\boldsymbol{W})=\sum_{i=1}^{n}\left\|\boldsymbol{x}_i-\sum_{j\neq i}w_{ij}\boldsymbol{x}_j\right\|^2, \tag{11}$$

其中, $\sum_{j=1}^{n}w_{ij}=1$. 如果 $\boldsymbol{x}_j$ 不是 $\boldsymbol{x}_i$ 的 k – 近邻, 那么 $w_{ij}=0$. 通过平滑假设 [20], 即特征相似的示例的标记也很可能相似, 可将特征空间的拓扑结构迁移到标记空间中, 即共享同样的局部线性重构权值矩阵 $\boldsymbol{W}$. 这样, 标记空间的标记分布可由最小化下式得到:

$$\Psi(\hat{\boldsymbol{d}})=\sum_{i=1}^{n}\left\|\hat{\boldsymbol{d}}_i-\sum_{j\neq i}w_{ij}\hat{\boldsymbol{d}}_j\right\|^2,\quad \text{s.t.}\quad d_{\boldsymbol{x}_i}^{y_l}l_{\boldsymbol{x}_i}^{y_l}>\lambda,\ \forall 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant c, \tag{12}$$

其中, $\lambda>0$ 是个预先设定的参数. 值得指出的是, 为了方便构建上述约束条件, 文献 [19] 中定义的逻辑标记 $\boldsymbol{l}_i\in\{-1,1\}^c$, 而不是其他方法中常用的 $\boldsymbol{l}_i\in\{0,1\}^c$, 但两者本质上并没有区别. 这样, 约束条件 $d_{\boldsymbol{x}_i}^{y_l}l_{\boldsymbol{x}_i}^{y_l}>\lambda$ 可以确保 $d_{\boldsymbol{x}_i}$ 与 $l_{\boldsymbol{x}_i}^{y_l}$ 同号. 通过求解上述二次规划问题确定 $\hat{\boldsymbol{d}}_i$ 后, 经过归一化即可得到标记分布 $\boldsymbol{d}_i$, 进而得到标记分布训练集. 基于流形的方法通过重构特征空间和标记空间的流形, 利用平滑假设, 将特征空间的拓扑关系迁移到标记空间中, 建立示例间相关性与标记间相关性之间的关系, 从而将逻辑标记增强为标记分布.

4 讨论与结论

本文简要介绍了 3 类标记分布学习算法和两类可用于实现标记增强的方法, 每一类方法分别介绍了几种实现算法. 在标记分布学习中, SA-IIS 和 SA-BFGS 的优化目标是直接最小化真实标记分布和预测标记分布之间的距离, 因此, 通常情况下这两种特别设计的 LDL 算法比其他 4 种从传统算法转化来的 LDL 算法的效果更好. 特别地, SA-BFGS 使用了更加高效的优化过程, 因此通常比 SA-IIS 表现更好. 在从传统算法转化来的 LDL 算法中, AA-BP 表现一般弱于 AA-kNN. 这是因为 BP 神经网络中大量的参数使得 AA-BP 容易过拟合. PT-Bayes 中的 Gauss 分布假设常常不适用于复杂的真实数据集, 因此表现弱于 PT-SVM. AA-kNN 的表现通常优于 PT-SVM, 因为 AA-kNN 的预测过程保持了标记分布的整体结构, 而 PT-SVM 使用了权值重采样, 破坏了原始标记分布的整体结构.

在标记增强算法中, 有些是专门为了将逻辑标记增强为标记分布而提出的, 如基于图模型的标记增强方法, 有些则是源于其他领域的工作, 但可以借鉴到本文语境中实现标记增强, 如基于模糊方法的标记增强. 这些不同的增强方法通过从样本中学习来获得额外的监督信息, 从而将训练样本原有的简单逻辑标记转化为信息量更为丰富的标记分布. 不同类型的标记增强算法具有其固有的优点和缺点.

基于模糊方法的标记增强利用模糊隶属度, 将标记间相关信息与逻辑标记融合, 不需要建立精确的数学模型, 有较强的鲁棒性, 数据和参数等对算法影响较小. 但是, 这类方法缺乏挖掘标记空间和特征空间关系的明确机制, 往往难以生成匹配特定数据特点的标记分布. 基于图模型的标记增强充分利用了特征空间的拓扑关系来指导标记空间相关性信息挖掘, 有良好的数学基础, 有利于形成适合数据本身特点的标记分布. 然而这类方法模型较为复杂, 数据和参数对算法的影响较大.

总而言之, 标记分布学习是一种比传统单标记和多标记学习更为泛化的学习范式, 能够处理标记的不同重要程度 (描述度) , 对许多实际应用问题具有本质上的普适性. 标记分布学习需要专门的算法设计以及专门的评价指标. 而目前很多实际问题中应用标记分布学习的主要困难是缺乏用标记分布标注的数据, 为此, 标记增强应运而生. 标记增强利用特征向量间拓扑关系或标记间相关性, 将简单逻辑标记增强为标记分布, 可以极大地拓展标记分布学习的应用范围. 简单来说, 标记增强可以看做为标记分布学习服务的一种数据预处理技术.

未来在标记分布学习和标记增强方面的研究至少可以包括如下 3 方面的内容:

(1) 研究标记分布学习与传统学习范式, 如单标记和多标记学习之间的关系. 例如通过标记增强, 使用标记分布学习解决多标记学习问题.

(2) 提出能够充分利用标记间相关性的标记增强算法. 标记间相关性既可以体现在标记空间, 也可能从示例空间迁移而来, 这方面不论是理论层面还是应用层面都还有很大的研究空间.

(3) 研究标记分布学习算法与标记增强算法之间的关系. 提出密切配合特定标记分布学习算法的标记增强算法, 使得增强后得到的标记分布更有利于特定标记分布学习算法的学习过程.

参考文献

1 Tsoumakas G, Katakis I. Multi-label classification: an overview. Int J Data Warehous Min, 2007, 3: 1–13

2 Geng X, Luo L. Multilabel ranking with inconsistent rankers. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 3742–3747

3 Zhou Y, Xue H, Geng X. Emotion distribution recognition from facial expressions. In: Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, 2015. 1247–1250

4 Geng X. Label distribution learning. IEEE Trans Knowl Data Eng, 2016, 28: 1734–1748

5 Zhou W J, Yu Y, Zhang M L. Binary linear compression for multi-label classification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, 2017. 3546–3552

6 Gao B B, Xing C, Xie C W, et al. Deep label distribution learning with label ambiguity. IEEE Trans Image Process, 2017, 26: 2825–2838

7 Wu T F, Lin C J, Weng R C. Probability estimates for multiclass classification by pairwise coupling. J Mach Learn Res, 2004, 5: 975–1005

8 Lin H T, Lin C J, Weng R C. A note on Platt's probabilistic outputs for support vector machines. Mach Learn, 2007, 68: 267–276

9 Berger A L, Pietra S D, Pietra V J D. A maximum entropy approach to natural language processing. Comput Linguist, 1996, 22: 39–71

10 Pietra S D, Pietra V D, Lafferty J D. Inducing features of random fields. IEEE Trans Pattern Anal Machine Intel, 1997, 19: 380–393

11 Nocedal J, Wright S. Numerical Optimization. 2nd ed. New York: Springer, 2006

12 Gayar N E, Schwenker F, Palm G. A study of the robustness of KNN classifiers trained using soft labels. In: Proceedings of the 2nd Conference Artificial Neural Networks in Pattern Recognition, Berlin, 2006. 67–80

13 Jiang X F, Yi Z, Lv J C. Fuzzy SVM with a new fuzzy membership function. Neural Comput Appl, 2006, 15: 268–276

14 Lin X T, Chen X W. Mr.KNN: soft relevance for multi-label classification. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, New York, 2010. 349–358

15 Jiang J Y, Tsai S C, Lee S J. FSKNN: multi-label text categorization based on fuzzy similarity and k nearest neighbors. Expert Syst Appl, 2012, 39: 2813–2821
16 Klir J G, Yuan B. Fuzzy Sets and Fuzzy Logic: Theory and Applications. Upper Saddle River: Prentice Hall, 1995
17 Li Y K, Zhang M L, Geng X. Leveraging implicit relative labeling-importance information for effective multi-label learning. In: Proceedings of IEEE International Conference on Data Mining, Piscataway, 2015. 251–260
18 Zhu X J, Goldberg A B. Introduction to Semi-Supervised Learning. Boca Raton: Morgan and Claypool Publishers, 2009
19 Hou P, Geng X, Zhang M L. Multi-label manifold learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Menlo Park, 2016. 1680–1686
20 Zhu X J. Semi-supervised learning with graphs. Dissertation for Ph.D. Degree. Pittsburgh: Carnegie Mellon University, 2005

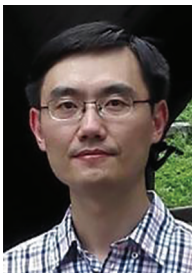
Label distribution learning and label enhancement

Xin GENG[1,2*] & Ning XU[1,2]

1. *School of Computer Science and Engineering, Southeast University, Nanjing* 211189*, China*;
2. *Key Laboratory of Computer Network and Information Integration, Ministry of Education, Nanjing* 211189*, China*
* Corresponding author. E-mail: xgeng@seu.edu.cn

Abstract This paper introduces the concepts and algorithms for label distribution learning (LDL) and label enhancement. LDL is a general machine learning paradigm with traditional single-label learning and multi-label learning as its special cases. A label distribution covers a certain number of labels, representing the degree to which each label describes the instance. Thus, LDL has been successfully applied to many real-world problems. Unfortunately, many existing datasets only have simple logical labels rather than label distributions. One way to solve the problem is to transform the logical labels into label distributions by mining the latent label importance from the training examples. Such a process of transforming logical labels into label distributions is defined as label enhancement. This paper provides formal definitions of label distribution learning and label enhancement. Subsequently, six representative LDL algorithms and four typical LE algorithms are briefly introduced and comparatively analyzed.

Keywords label distribution, label distribution learning, label enhancement, multi-label learning, learning with ambiguity



Xin GENG received his B.Sc. and M.Sc. degrees in computer science from Nanjing University, China, in 2001 and 2004, respectively, and his Ph.D. degree from Deakin University, Australia in 2008. He joined the School of Computer Science and Engineering at Southeast University, China, in 2008, and is currently a professor and vice dean (research) of the school. His research interests include pattern recognition, machine learning, and computer vision.

Ning XU was born in 1988. He is currently pursuing his Ph.D. in School of Computer Science and Engineering at Southeast University, China. His main research interests include pattern recognition and machine learning.