计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):125−131 [doi: 10.15888/j.cnki.csa.007530]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于标记特定特征和相关性的 ML-KNN 改进算法①

李　永, 许　鹏

(北京工业大学 软件学院, 北京 100124)
通讯作者: 许　鹏, E-mail: xupeng0727@foxmail.com

摘　要: 目前大部分已经存在的多标记学习算法在模型训练过程中所采用的共同策略是基于相同的标记属性特征集合预测所有标记类别. 但这种思路并未对每个标记所独有的标记特征进行考虑. 在标记空间中, 这种标记特定的属性特征对于区分其它类别标记和描述自身特性是非常有帮助的信息. 针对这一问题, 本文提出了基于标记特定特征和相关性的 ML-KNN 改进算法 MLF-KNN. 不同于之前的多标记算法直接在原始训练数据集上进行操作, 而是首先对训练数据集进行预处理, 为每一种标记类别构造其特征属性, 在得到的标记属性空间上进一步构造 L_1-范数并进行优化从而引入标记之间的相关性, 最后使用改进后的 ML-KNN 算法进行预测分类. 实验结果表明, 在公开数据集 image 和 yeast 上, 本文提出的算法 MLF-KNN 分类性能优于 ML-KNN, 同时与其它另外 3 种多标记学习算法相比也表现出一定的优越性.

关键词: 多标记学习; 标记特定特征; 标记相关性; 多标记 K 近邻; L_1-范数

引用格式: 李永,许鹏.基于标记特定特征和相关性的 ML-KNN 改进算法.计算机系统应用,2021,30(2):125–131. http://www.c-s-a.org.cn/1003-3254/7530.html

Improved ML-KNN Algorithm Based on Label Specific Features and Correlations

LI Yong, XU Peng

(School of Software, Beijing University of Technology, Beijing 100124, China)

Abstract: The common strategy adopted by most existing multi-label learning algorithms in model training is to predict all the label categories based on the same label feature set. However, this idea does not take into account the label-specific features of each label, which are very helpful for distinguishing other categories of labels and describing itself in the label space. For this reason, an improved ML-KNN algorithm based on label-specific features, i.e., MLF-KNN, is proposed in this study. Different from the previous multi-label algorithms which directly operate on the original training data set, the algorithm proposed in this study first builds features for each category of label by preprocessing the training data set. Then, it further constructs and optimizes L_1-norm in the obtained label space, thus introducing the correlation between labels. Finally, the improved algorithm is applied for prediction and classification. The experimental results show that the improved algorithm has achieved certain advantages compared with the ML-KNN algorithm and other three multi-label learning algorithms on the public image and yeast data sets.

Key words: multi-label learning; label specific features; label correlation; ML-KNN; L_1-norm

传统监督学习认为一个对象只具有一个标记类别, 属于“单示例, 单标记”类型. 在上述单一语义的情境中, 监督学习已经取得了巨大的发展成果. 然而在真实世界中, 一个对象往往同时具有多种语义信息, 属于“单

① 收稿时间: 2019-12-26; 修改时间: 2020-01-20; 采用时间: 2020-02-11; csa 在线出版时间: 2021-01-27

示例, 多标记”类型. 例如在一篇新闻稿中可同时包含改革和经济两个主题, 一张风景图中天空和云朵往往会伴随出现, 在这种情况下很难用单一语义标记去描述对象信息. 为此, 多标记学习框架应运而生, 用于解决真实世界中一个对象同时具有多个语义标记的问题. 因其可以良好的反映真实世界中包含的多语义信息, 目前在文本分类[1], 图像标注[2], 生物基因分析[3] 等领域得到了广泛应用. 同时众多学者也提出了一些多标记学习算法, 并取得了一定的成功. 但是目前大部分已有的多标记学习算法所采用的共同策略是基于同一特征属性空间预测所有标记类别, 忽略了每个标记独有的特征信息, 因此这种思路存在改进优化的空间. 其中 ML-KNN[4] 作为一种使用简单, 分类性能优异的多标记学习算法, 其数学形式相对简单, 模型易于优化, 在实际应用和学术科研中得到了广泛应用. 但是 ML-KNN 算法在模型训练过程中并没有考虑标记之间的相关性, 同时也忽略了标记特定特征信息. 因此, 在模型训练过程中考虑标记相关性和引入标记特定特征信息, 可以进一步提高算法的分类性能, 基于此提出了本文算法. 本文的整体组织结构如下: 第 1 部分介绍相关工作, 第 2 部分描述本算法的实现, 第 3 部分给出实验以及实验结果, 最后进行了总结.

1 相关工作

在传统“单示例, 单标记”的单语义环境中, 传统监督学习已经取得了巨大的发展[5–8]. 然而在真实世界中, 语义信息往往是丰富多彩的, 传统监督学习已经不能对同时从属于多个标记类别下的单个示例进行很好的语义表达. 相比于传统监督学习, 多标记学习可以更好的反映真实世界中包含的语义信息. 不同于传统监督学习所假设的“单示例, 单标记”情形, 多标记学习所研究的任务场景属于“单示例, 多标记”类型. 我们假设 $\mathcal{X}=\mathbb{R}^d$ 代表 d 维特征空间, $\mathcal{Y}=\{l_1,l_2,\cdots,l_q\}$ 代表包含 q 个标记类别的标记空间. 多标记学习的任务是在训练集 $D=\{(x_i,Y_i)|1\le i\le m\}$ 中训练一个分类器 $h:\mathcal{X}\to 2^{\mathcal{Y}}$, 预测未知示例 $x\in\mathcal{X}$ 所从属的标记集合 $h(x)\subseteq\mathcal{Y}$, 其中 $x_i\in\mathcal{X}$ 为特征空间中的一个示例, 是一个 d 维特征向量, $Y_i\in\mathcal{Y}$ 为示例 x_i 所从属的标记集合, 是标记空间 $\mathcal{Y}$ 中的一个子集.

目前多标记学习研究所面临的主要挑战是标记空间爆炸性增长的问题[9]. 即类别标记集合数随着标记种类的增加而呈指数级增长, 假设样本中包含有 q 个类别标记信息, 则标记输出空间规模即可达到 2^q 级别的大小, 为每个标记子集单独训练一个分类器是不切实际的. 为了解决这个标记空间爆炸的问题, 目前关于多标记学习的研究大都集中在通过挖掘标记之间的相关性来降低指数级别增长的标记空间. 根据标记相关性的利用策略, 可以将多标记学习算法分为 3 类: (1) 一阶策略, 完全忽略标记之间的相关性, 只是将一个多分类问题转换为多个独立的二分类问题, 这类方法虽然实现简单但是缺少良好的泛化性能. (2) 二阶策略, 考虑标记之间的成对关联, 例如相关标记与无关标记之间的排序关系, 两两标记之间的交互关系等构造多标记学习框架. 这类算法具有一定的泛化性能, 但是无法很好的解决标记之间的关系超过二阶时的问题. (3) 高阶策略, 考虑单个标记与其它全部标记之间的相关性, 这类算法可以很好的反映真实世界的标记相关性, 但同时模型复杂度往往较高. 目前众多学者也提出了一些多标记学习算法, 例如由 Boutell 等提出的一阶策略算法 BR (Binary Relevance)[10], 由 Fürnkranz 等提出的二阶策略算法 Calibrated Label Ranking[11], 由 Read 等提出的高阶策略算法 Classifer Chains[12] 等. 上述算法的共同点是将多分类问题分解为多个二分类问题进行解决, 属于问题转换型. 由 Clare 等提出的一阶策略算法 ML-KNN, ML-DT[13] 和 Elisseeff A 提出的二阶策略算法 Rank-SVM[14] 等算法是采用当前的机器学习算法直接处理多分类问题, 属于算法适应型. 但无论是问题转换型算法还是算法适应型算法, 上述提到的算法在预测标记时都是假设所有标记共享同一特征空间, 并未对每种类别标记独有的特征信息进行考虑. 即多标记学习框架得到的 q 个实值函数 $\{f_1,f_2,\cdots,f_q\}$ 都是基于相同的属性特征空间训练而来. 但是这种思路可能并不是最优的, 例如在图像识别领域, 在判断“天空”和“沙漠”时, 颜色特征是需要优先考虑的, 而在判断“蓝天”和“大海”时, 考虑纹理特征相关属性会大大提高分类器的性能. 由此可见, 每个标记都可能具有与其最大相关性的属性特征, 考虑每个标记独有的属性特征对于提高算法分类性能具有一定的帮助, 这些属性特征是对该标记最有区别度的特征, 称为标记特定特征信息.

其中, ML-KNN 作为一种使用简单, 分类性能高效的算法, 在实际应用中得到了广泛应用. 但是 ML-KNN

算法属于一阶策略, 并未对标记之间相关性进行考虑, 同时也没有考虑标记特定特征信息. 虽然该算法取得了巨大的成果, 但是并未对标记相关性和标记特定特征信息加以利用, 存在优化改进的空间. 由 Zhang 等提出了 LIFT 算法[15] 首次提出从引入标记特定特征信息这一角度出发的研究思路, 针对每一个标记信息提取其特征信息, 从而构建标记特征空间, 之后基于标记特征空间进行模型训练. 该算法基于多种公开数据集证明了其思路的有效性, 为多标记学习指明了一个新的研究方向. 但是该算法并未考虑标记之间的相关性, 属于一阶策略算法. 基于此, 我们通过融入标记相关性和引入标记特定特征信息对 ML-KNN 算法进行改进, 进一步提高算法分类性能.

2 ML-KNN 算法以及改进算法

2.1 ML-KNN 算法

ML-KNN 算法是在 k 近邻算法的基础上, 综合贝叶斯理论提出的能够处理多标记分类问题的 KNN 算法. 在此引入一些符号用于对该算法进行说明. 在多标记训练数据集 $D=\{(x_i,Y_i)|1\le i\le m\}$ 中, $x_i\in\mathcal{X}$, $Y_i\in\mathcal{Y}$, Y_x 为示例 x_i 所对应的 q 维标记向量. 如果示例 x 具有类别标记 $l(1\le l\le q)$, 则定义 $Y_x(l)=1$, 否则 $Y_x(l)=0$. 另外假设 $N(x)$ 代表示例 x 在训练集 D 中的 k 个近邻集合, 对这 k 个近邻集合中拥有标记 l 的样本数量用 $C_x(l)$ 表示, 其中:

$$C_x(l)=\sum_{a\in N(x)}Y_a(l),\ l\in\mathcal{Y} \tag{1}$$

其中, C_x 代表了示例 x 所对应的 k 个近邻集合中所包含的标记信息. ML-KNN 算法为懒惰算法, 当有新的测试示例 t 需要进行预测分类时, ML-KNN 首先识别示例 t 在训练数据集 D 中的 k 个近邻集合 $N(t)$, 在这里设定 H_1^l 代表示例 t 拥有标记 l, 相反, 当示例 t 没有标记 l 时用 H_0^l 表示. 另外引入 E_j^l 表示在 $N(t)$ 中有 j 个示例拥有标记 l, 其中 $j\in\{0,1,\cdots,k\}$, 基于向量 C_t, 示例 t 所对应的类别向量 Y_t 可以使用如下最大后验概率来计算.

$$Y_t(l)=\arg\max_{b\in\{0,1\}}\frac{P\left(H_b^l\right)P\left(E_{C_t(l)}^l|H_b^l\right)}{P\left(E_{C_t(l)}^l\right)} \tag{2}$$

该式所代表的含义为已知在测试示例 t 的 k 个近邻集合中有 $C_t(l)$ 个样本与标记 l 相关, 示例 t 与标记 l 是否相关取决于 $N(t)$ 中是否与标记 l 相关的最大概率. 根据贝叶斯规则, 上式可以进一步变换为:

$$\begin{aligned}Y_t(l)&=\arg\max_{b\in\{0,1\}}\frac{P\left(H_b^l\right)P\left(E_{C_t(l)}^l|H_b^l\right)}{P\left(E_{C_t(l)}^l\right)}\\&=\arg\max_{b\in\{0,1\}}P\left(H_b^l\right)P\left(E_{C_t(l)}^l|H_b^l\right)\end{aligned} \tag{3}$$

由上述公式可知, 计算 $Y_t(l)$ 需要得到先验概率 $P\left(H_b^l\right)$ 和条件概率 $P(E_{C_t(l)}^l|H_b^l)$, 这两个值都是可以从训练数据样本计算得到的.

2.2 基于 ML-KNN 算法的改进算法

LIFT 算法是由张敏灵等学者在研究多标记学习算法时提出的一种全新的算法. 该算法不同于之前研究点集中于挖掘标记之间相关性上, 而忽略了标记特定特征信息. LIFT 算法从挖掘标记特征这一角度出发, 针对每一个标记类别, 通过挖掘其特征信息构建标记特征空间, 在模型训练过程中引入标记特征, 并且经过大量实验表明 LIFT 算法分类性能在公开数据集上优于其它多标记学习算法, 同时也证明了在模型训练过程中引入标记特定特征信息提高算法分类性能这一思路的有效性, 为后续多标记学习研究提供了一种新的思路.

2.2.1 LIFT 算法基本模型

LIFT 算法通过对训练样本中的每个类别标记进行聚类操作, 分析每个标记的特征信息, 将原始样本集合根据特征标记划分为与当前标记呈正相关的样本和负相关的样本. 具体来说, 对于标记 $l_j\in\mathcal{Y}$, 根据式 (4) 和式 (5) 分别计算其正相关的示例集合 P_j 和负相关示例集合 N_j.

$$P_j=\{x_i|(x_i,Y_i)\in D,l_j\in Y_i\} \tag{4}$$

$$N_j=\{x_i|(x_i,Y_i)\in D,l_j\notin Y_i\} \tag{5}$$

其中, Y_i 表示实例 x_i 所对应的标记向量, D 为训练数据集, P_j 代表包含标记 l_j 的示例集合, N_j 代表不包含标记 l_j 的示例集合. 之后采用 K-means 聚类算法分别对两个示例集合进行聚类操作, 得到在 P_j 上的 m_j^+ 个聚类中心 $\left\{p_1^j,p_2^j,\cdots,p_{m_j^+}^j\right\}$ 和在 N_j 上的 m_j^- 个聚类中心 $\left\{n_1^j,n_2^j,\cdots,n_{m_j^-}^j\right\}$. 为了解决聚类中心不平衡的问题, LIFT 算法设定 $m_j=m_j^+=m_j^-$, 其中 $m_j=r\cdot\min\left(|P_j|,|N_j|\right)$. 这两个聚类中心集合分别代表着正负相关示例集合的内在特征结构, 可作为构建标记特定特征空间的基础. 为了构建标记 l_j 的特定特征空间 Z_j, 采用如下映射 ϕ_j: $\mathcal{X}\to Z_j$, ϕ_j 表达如下:

$$\phi_j(x)=\left[d\left(x,p_1^j\right),\cdots,d\left(x,p_{m_j^+}^j\right),d\left(x,n_1^j\right),\cdots,d\left(x,n_{m_j^-}^j\right)\right] \tag{6}$$

其中, $d(\cdot,\cdot)$代表两个向量之间的欧式距离. 根据映射ϕ_j, 可以将原始训练数据集 D 转换为标记l_j对应的二分类训练集 Z. 之后在新的训练集 Z 上进行模型的训练, 可以构建出一个基于标记特定特征推导出的二分类器簇$\left\{g_1,g_2,\cdots,g_q\right\}$. 一般地, 对于标记$l_j\in\mathcal{Y}$, 其对应的二分类训练数据集$D_j^*$可由映射$\phi_j$表示为:

$$D_j^*=\left\{\left(\phi_j(x_i),Y_i(j)\right)\mid(x_i,Y_i)\in D\right\} \tag{7}$$

其中, 如果标记$l_j\in Y_i$, $Y_i(j)=1$, 否则$Y_i(j)=0$. 基于训练数据集D_j^*, 可推导出一个分类器$g_j^*:Z_j\to R$. 对于未知示例 u, 其对应的标记集合可以形式化的表示为$Y=\left\{l_k\mid g_k^*(\phi_k(u))>t,1\le k\le q\right\}$, 其中 t 代表一个阈值函数, 一般设置为常数 0.

2.2.2 MLF-KNN 算法

ML-KNN 算法属于一阶策略算法, 在模型训练过程中没有考虑标记之间的相关性信息, 同时在预测标记时是基于相同的属性特征集合, 忽略了每个标记所独有的属性特征信息, 因此 ML-KNN 算法存在优化改进的空间. LIFT 算法经过大量实验表明该算法的分类性能与其它多标记学习算法相比具有一定的竞争力, 证明了在模型训练过程中引入标记特定特征信息可以提高算法分类性能这一思路的可行性和有效性. 为此, 我们借鉴该思路, 对 ML-KNN 算法进行改进, 并且融入标记相关性信息, 提出基于标记特定特征新的多标记分类新算法 MLF-KNN (Multi-Label Feature-K Nearest Neighbor). 本算法首先对多标记训练样本集合进行预处理, 通过对每个类别标记进行聚类分析构建基本标记特征, 之后通过稀疏正则化的方式协同增强与其它类别标记的信息从而增强对每个标记特征信息的表述, 进而引入当前标记与其它标记之间的相关性. 基于得到的标记特定特征, 使用改进后的 ML-KNN 算法进行分类. 不失一般性, 引入一些符号进行本算法的说明. 在模型训练之前, 首先需要构建每个标记所对应的正负相关示例集合. 对于训练样本中的每个标记$l_k\in\mathcal{Y}(1\le k\le q)$, MLF-KNN 算法根据式 (4) 和式 (5) 将原始示例集合分为m_k个正相关特征集合P_k和m_k个负相关特征集合N_k, 其中$m_k=r\cdot\min(|l_k\in Y_k|,|l_k\notin Y_k|)$. 之后在集合$P_k$和$N_k$中采用 K-means 聚类算法构建聚类中心, 用于构建基本标记特征空间. 在正相关示例集合P_k中聚类中心可以通过$\left\{p_1^k,p_2^k,\cdots,p_{m_k}^k\right\}$表示, 负相关特征集合$N_k$中聚类中心可以通过$\left\{n_1^k,n_2^k,\cdots,n_{m_k}^k\right\}$表示. 对于示例$x\in\mathcal{X}$, 可通过计算示例 x 与聚类中心之间的距离构建其基本标记特征, 最终生成的基本标记特征空间是一个大小为$2m_k$的矩阵$\phi_k:\mathcal{X}\to\mathbb{R}^{2m_k}$.

$$\phi_k(x)=\left[d\left(x,p_1^k\right),\cdots,d\left(x,p_{m_k}^k\right),d\left(x,n_1^k\right),\cdots,d\left(x,n_{m_k}^k\right)\right]^{\mathrm{T}} \tag{8}$$

其中, $d(\cdot,\cdot)$不在采用 LIFT 算法中的欧式距离进行度量, 而是采用标准欧式距离进行计算, 相比于欧式距离, 标准欧式距离对于两个向量中的维度不一致的情况进行了考虑, 对于维度不一致的示例具有更好的包容性, 根据式 (7), 生成标记l_k的基本特定特征空间. 此时基本标记特定特征空间只是针对标记l_k所构建, 并未考虑l_k与其它标记之间的相关性. 假设$y_k=[y_{k1},y_{k2},\cdots,y_{km}]^{\mathrm{T}}$代表标记向量$l_k$, 类似的, 如果$l_k\in Y_i$, 则$y_{ki}=1$, 否则$y_{ki}=0$. 进一步, 设定:

$$\varphi_k(x)=\left[\phi_1(x),\cdots,\phi_{k-1}(x),\phi_{k+1}(x),\cdots,\phi_q(x)\right]^{\mathrm{T}} \tag{9}$$

其中, $\varphi_k(x)$表示除标记l_k以外的其它所有类别标记的特定特征信息, 是一个维度为d_k的特征向量, 其中$d_k=\sum_{k'\ne k}2m_{k'}$. 将映射$\phi_k$应用于全部训练样本上, 从而可以构建出关于标记$l_k$并且维度为$m\times d_k$的标记特定特征矩阵$X_k$, 其中$X_k=[\varphi_k(x_1),\varphi_k(x_2),\cdots,\varphi_k(x_m)]^{\mathrm{T}}$. 为了解决 ML-KNN 算法并未考虑标记之间相关性的问题, 我们需要对标记之间相关性进行挖掘并将相关性信息引入标记特定特征空间中. 具体来说, 需要通过引入标记l_k与其它类别标记之间的相关性对之前生成的基本标记特征空间X_k进行增强. 在本方法中将标记之间相关性问题转换为最小二乘法优化问题, 通过引入 L_1 正则化项对其进行优化求解从而引入标记l_k与其它类别标记之间的相关性, 优化最小二乘法问题如下:

$$\min_{\beta_k\in\mathbb{R}^{d_k}}\|y_k-X_k\cdot\beta_k\|_2^2+\lambda\cdot\|\beta_k\|_1 \tag{10}$$

其中, $\beta_k=\left[\beta_{k1},\beta_{k2},\cdots,\beta_{kd_k}\right]^{\mathrm{T}}$为权值向量, 代表对每一个标记的影响程度, λ代表正则系数. 使用稀疏回归方法解决上述最小二乘法问题后, 即可得到其它类别标记所独有的特征信息, 进而通过与权值向量β_k结合对矩阵ϕ_k进行增强, 通过增强ϕ_k即可引入l_k与其它类别标记信息之间的相关性. 使用集合$\mathcal{I}_k$存储在φ_k中权值大

于阈值γ的索引.

$$\mathcal{I}_k=\{a\,|\,|\beta_{ka}|\geq\gamma,1\leq a\leq d_k\} \tag{11}$$

根据之前生成的基本标记特征空间ϕ_k中的数值取值介于 0 到 1 之间, 在此我们将阈值γ设定为常数值 0.5. 对于标记l_k, 通过融合标记相关性信息后对基本标记特征进行增强, 即可生成最终标记特定特征ψ_k: $\mathcal{X}\to Z_k$, 其中$\psi_k(x)$表示形式如下:

$$\psi_k(x)=\left[\phi_k(x),\prod\nolimits_{\mathcal{I}_k}(\varphi_k(x))\right] \tag{12}$$

其中, $\prod_{\mathcal{I}}(u)$代表在索引集$\mathcal{I}$上对 u 的投影操作. 综上, 关于标记 l_k 的二分类训练集合 D_k 能够表达为:

$$D_k=\{(\psi_k(x_i),y_{ki})\,|\,1\leq i\leq m\} \tag{13}$$

针对训练样本中的每一个标记类别分别构造其特定特征空间, 在标记空间集合中应用改进后的 ML-KNN 算法进行模型训练, 其中在计算两个标记之间距离时 MLF-KNN 算法不同于 ML-KNN 采用欧式距离直接计算, 而是采用如下 r 阶 Minkowski 距离进行计算[16].

$$dist(x,y)=\left(\sum_{l=1}^{d}\left\|x^l-y^l\right\|^r\right)^{1/r} \tag{14}$$

其中, x^l 代表示例 x 的第 l 维, ||·||表示取该实数值的绝对值. 在计算两个样本之间的距离时采用 Minkowski 距离度量方法的主要出发点是考虑到不同数据集内数据可能具有不同分布从而需要采取不同的距离计算方法, 采用 Minkowski 方法可以提高本算法对不同数据集的包容性. 当阶数取值为 1 时, 可以转换为曼哈顿距离. 当阶数取值为 2 时, 可以转换为欧氏距离. 当阶数继续增大到无穷大时可转换为切比雪夫距离. 在本实验中所采用的数据集中的数据取值规范, 分布较为独立, 因此设定 r 值为 2 转换为欧式距离进行试验. 当数据集中的数据分布具有关联性或者局限性时, 可以改变 r 值的取值以适配不同的数据分布. MLF-KNN 算法描述如算法 1 所示.

算法 1. MLF-KNN 算法

步骤 1. 构造基本标记特征集.
For l_k in $\mathcal{Y}=\{l_1,l_2,\cdots,l_q\}$ do
利用式 (4) 和式 (5) 构建标记正相关样本集P_k和负相关样本集N_k;
End for
For l_k in $\mathcal{Y}=\{l_1,l_2,\cdots,l_q\}$ do
通过式 (8) 构建针对l_k的标记特征映射ϕ_k;
End for

步骤 2. 增强基本标记特征集, 构建标记特征.
For l_k in $\mathcal{Y}=\{l_1,l_2,\cdots,l_q\}$ do
通过对式 (10) 进行 L_1-范数正则化计算权重向量β_k;
通过式 (11), 式 (12), 构建最终标记特征;
End for
步骤 3. 构建二分类训练集$\mathcal{D}_k$.
For l_k in $\mathcal{Y}=\{l_1,l_2,\cdots,l_q\}$ do
根据式 (13) 构建标记l_k的二分类训练集$\mathcal{D}_k$;
End for
步骤 4. 计算先验概率.
For l_k in $\mathcal{Y}=\{l_1,l_2,\cdots,l_q\}$ do
计算标记l_k的先验概率:
$P(H_b^l)(b\in\{0,1\},l\in\mathcal{Y})$;
End for
步骤 5. 计算示例在标记特定特征空间中的 k 近邻集合.
For x_t^k in $\mathcal{D}_k$ do
通过式 (14) 计算 k 近邻集合$N(x)$, 从而根据式 (1) 计算$C_x(l)$
End for
步骤 6. 计算各类标记条件概率$P(E^l_{C_x(l)}|H^l_1)$和条件概率$P(E^l_{C_x(l)}|H^l_0)$.
步骤 7. 对待测样例 t 进行分类, 通过式 (8) 构建新的样本表达t_k, 分别在$\mathcal{D}_k$计算$N(t)$和$C_t(l)$, 利用式 (3) 计算其对应的标记向量Y_t.

3 实验

为了检验 MLF-KNN 算法的分类效果, 将本算法与 ML-KNN, LIFT, Rank-SVM 等 3 个多标记学习算法在公开酵母数据集 yeast 和图像数据集 image 上进行比较. 其中 ML-KNN 算法基于传统 k 近邻技术处理多标记问题, 基于已有样本的先验概率, 通过在训练样本中寻找距离最近的 k 个实例从而对未知示例进行标记预测. 但是该方法没有考虑标记之间的关联信息, 属于算法适应型一阶策略算法. LIFT 算法通过构造每一个标记独有的特征信息, 基于标记独有的示例集合进行模型训练, 是从标记特征信息研究的新型算法, 但是同样没有引入标记之间的相关信息, 也属于一阶策略算法. Rank-SVM 算法使用最大化间隔思想处理多标记问题, 该算法的核心是通过一组线性分类器对 Ranking Loss 指标进行优化, 并通过引入“核技巧”处理非线性分类问题[9], 属于算法适应型二阶策略算法. 实验所采用的数据集 yeast 和 image 在多标记学习领域是两个公开的数据集, 分别在生物领域和图像领域具有一定的代表性, 两个数据集详细信息如表 1 所示.

表 1 数据集详细信息

数据集	训练样本数	属性个数	标记个数	平均标记个数	所属领域
yeast	2417	103	14	1.236	生物图像
image	2000	294	5	4.237	

由于每个示例同时类属于多个标记, 因此传统的单标记评价指标例如精度 (accuracy)、查准率 (precision) 和查全率 (recall) 等指标不再适用. 本文评价指标基于 Haming Loss, One-error, Coverage, Ranking Loss, Average Precision[17] 等 5 种在多标记学习领域广泛使用的评价指标. 其中, 对于 Average Precision 指标信息, 数值越大代表分类性能越好, 在表中使用↑表示, 其余评价指标数值越小则代表其分类性能越好, 在表中使用↓表示. 表 2 为各个算法在数据集 yeast 上的测试结果, 表 3 为在数据集 image 上的测试结果. 本文提出的算法基于 ML-KNN 改进而来, 图 1 为 MLF-KNN 算法与 ML-KNN 算法在数据集 yeast 上的随 k 值变化的 Coverage 评价指标对比结果图

表 2　本文算法与其它算法在数据集 yeast 上的实验结果对比

评价指标	MLF-KNN	ML-KNN	LIFT	Rank-SVM
Haming Loss↓	**0.189**	0.198	0.197	0.207
One-error↓	0.285	**0.234**	0.285	0.243
Coverage↓	**0.965**	7.414	0.974	7.087
Ranking Loss↓	0.172	0.173	**0.169**	0.195
Average Precision↑	0.308	0.730	0.470	**0.749**

表 3　本文算法与其它算法在数据集 image 上的实验果对比

评价指标	MLF-KNN	ML-KNN	LIFT	Rank-SVM
Haming Loss↓	**0.152**	0.165	0.159	0.253
One-error↓	0.820	0.329	**0.273**	0.491
Coverage↓	**0.441**	1.980	0.768	1.382
Ranking Loss↓	0.173	0.175	**0.130**	0.277
Average Precision↑	0.505	0.800	**0.819**	0.682

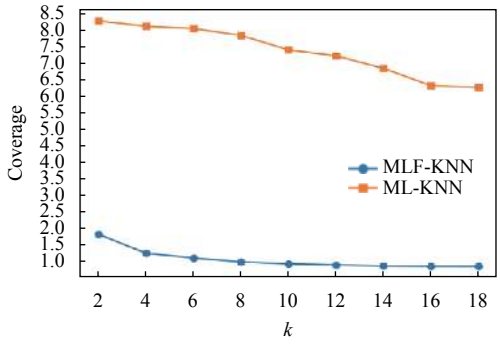

图 1　MLF-KNN 与 ML-KNN 算法在数据集 yeast 数据集随 k 值变化的 Coverage 变化曲线图

由图 1 可以看出, 本文提出的算法相比于 ML-KNN 算法在 Coverage 评价指标上表现性能十分优异, 当 k 值继续增大时, Coverage 值可以接近最优解 1.0, 代表 MLF-KNN 算法对于测试样本的预测所有相关性标记的平均查找深度小, 分类性能有所提高. 由表 2 和表 3 的对比实验结果图可以看出, 本文所提出的算法 MLF-KNN 在数据集 yeast 当中表现优异. 反映在具体评价指标上, MLF-KNN 算法在 One-error, Ranking Loss 和 Average Precision 指标上表现不是最优, 虽然 Ranking Loss 指标上的表现低于 LIFT 算法, 但是相比 ML-KNN 算法而言该评价指标有所提高. 在数据集 image 中, 本算法同样在评价指标 Haming Loss 和 Coverage 上表现最优. 实验结果表明对 ML-KNN 算法引入标记特定特征和标记相关性进行改进可以进一步提高算法分类性能, 尤其在 Coverage 上表现更为明显, 同时也证明了从标记特定特征对算法进行改进创新这一思路的有效性.

4　结论与展望

ML-KNN 算法在模型训练过程中没有考虑到标记之间的相关性, 同时在预测不同标记时基于同一特征空间, 忽略了每个标记所特定的标记特征. LIFT 算法从利用每个标记所独有的特征出发, 为多标记学习研究指明了一个新的方向. 基于此, 我们对 ML-KNN 算法进行改进, 在构造标记特征空间的同时对其进行增强, 引入与其它标记之间的相关性, 提高了算法分类性能. 在之后的工作中, 可以进一步对构建标记特征空间的方法进行创新, 另外本算法并未对标记不平衡的问题进行考虑, 当正负标记样本数量差异过大时会制约算法的分类性能, 所以下一步工作可以将当前多标记学习领域对正负标记不平衡问题的研究[18] 引入到本算法当中.

参考文献

1 Nam J, Kim J, Mencía EL, *et al*. Large-scale multi-label text classification —Revisiting neural networks. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Nancy, France, 2014: 437–452.

2 Wang H, Huang H, Ding C. Image annotation using multi-label correlated green's function. Proceedings of the IEEE 12th International Conference on Computer Vision. Kyoto, Japan. 2009. 2029–2034.

3 Goncalves EC, Plastino A, Freitas AA. A genetic algorithm for optimizing the label ordering in multi-label classifier chains. Proceedings of the IEEE 25th International Conference on Tools with Artificial Intelligence. Herndon,

VA, USA. 2013. 469–476.
4 Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7): 2038–2048. [doi: 10.1016/j.patcog.2006.12.019]
5 Tsoumakas G, Spyromitros-Xioufis E, Vilcek J, *et al*. Mulan: A java library for multi-label learning. Journal of Machine Learning Research, 2011, 12: 2411–2414.
6 广凯, 潘金贵. 一种基于向量夹角的 k 近邻多标记文本分类算法. 计算机科学, 2008, 35(4): 205–206. [doi: 10.3969/j.issn.1002-137X.2008.04.061]
7 陈琳琳, 陈德刚. 一种基于核对齐的分类器链的多标记学习算法. 南京大学学报 (自然科学版), 2018, 54(4): 725–732.
8 Sun YY, Zhang Y, Zhou ZH. Multi-label learning with weak label. Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, GA, USA. 2010. 593–598.
9 Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819–1837. [doi: 10.1109/TKDE.2013.39]
10 Boutell MR, Luo JB, Shen XP, *et al*. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757–1771. [doi: 10.1016/j.patcog.2004.03.009]
11 Fürnkranz J, Hüllermeier E, Mencía EL, *et al*. Multilabel classification via calibrated label ranking. Machine Learning, 2008, 73(2): 133–153. [doi: 10.1007/s10994-008-5064-8]
12 Read J, Pfahringer B, Holmes G, *et al*. Classifier chains for multi-label classification. Machine Learning, 2011, 85(3): 333. [doi: 10.1007/s10994-011-5256-5]
13 Clare A, King RD. Knowledge discovery in multi-label phenotype data. Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery. Freiburg, Germany. 2001. 42–53.
14 Elisseeff A, Weston J. A kernel method for multi-labelled classification. Proceedings of the 14th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada. 2002. 681–687.
15 Zhang ML, Wu L. Lift: Multi-label learning with label-specific features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 107–120. [doi: 10.1109/TPAMI.2014.2339815]
16 蒋芸, 肖潇, 侯金泉, 等. 融合标记独有属性特征的 k 近邻多标记分类新算法. 计算机工程与科学, 2019, 41(3): 513–519. [doi: 10.3969/j.issn.1007-130X.2019.03.018]
17 Chen ZS, Zhang ML. Multi-label learning with regularization enriched label-specific features. Proceedings of Machine Learning Research, Volume 101: Asian Conference on Machine Learning. Nagoya, Japan. 2019. 411–424.
18 Zhang ML, Li YK, Liu XY. Towards class-imbalance aware multi-label learning. Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina. 2015. 4041–4047.