基于标记依赖关系集成分类器链的多示例多标签支持向量机算法①

李村合, 王文杰

(中国石油大学 计算机与通信工程学院, 青岛 266580)

摘 要: ECC-MIMLSVM$^+$是多示例多标签学习框架下一种算法, 该算法提出了一种基于分类器链的方法, 但其没有充分考虑到标签之间的依赖关系, 而且当标签数目的增多, 子分类器链长度增加, 使得误差传播问题凸显. 因此针对此问题, 提出了一种改进算法, 将 ECC-MIMLSVM$^+$算法和标签依赖关系相结合, 设计成基于标记依赖关系集成分类器链(ELDCT-MIMLSVM$^+$)来加强标签间信息联系, 避免信息丢失, 提高分类的准确率. 通过实验将本文算法与其他算法进行了对比, 实验结果显示, 本文算法取得了良好的效果.

关键词: 多示例多标签; 支持向量机; 标签依赖关系; 分类器链

Multi-Instance Multi-Label Support Vector Machine Algorithm Based on Labeled Dependency Relation Ensemble Classifier Chain

LI Cun-He, WANG Wen-Jie

(College of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: ECC-MIMLSVM$^+$ is an algorithm of multi-instance and multi-label learning framework. This algorithm proposes a method based on classifier chain, but it does not consider the dependencies between labels. When the number of tags increases, the length of the sub classifier chain also increases, making the error propagation problem prominent. Therefore, this paper presents a kind of improved algorithm, combining ECC-MIMLSVM$^+$ algorithm and the label dependencies. ELDCT-MIMLSVM$^+$ algorithm is designed, which is based on ensembles of label dependencies classifier chain to avoid the information loss and improve the classification accuracy. The experiment results show that the algorithm has good effect.

Key words: multi-instance multi-label; SVM; ensembles of label dependencies; classifier chains

1 引言

在传统监督学习[1]中, 每个对象用一个示例进行描述, 该示例隶属于一个概念标签, 在该框架下对象不具有歧义性. 在多示例学习[2]中, 每个对象由多个示例进行描述, 并且同时隶属于一个概念标签, 其描述对象的内容存在歧义性. 在多标签[3,4]学习中, 每个对象由单个示例描述, 并且隶属于多个概念标签, 其描述对象的概念存在歧义性. 然而在现实生活中的对象通常同时具有内容、概念两方面的歧义性, 因此出现了多示例多标签学习框架(MIML)[5,6].

2007 年, 南京大学的周志华等人提出两个多示例多标签支持向量机算法, 分别是 MIMLBOOST 算法[6]和 MIMLSVM 算法[7].

2010 年, 河海大学的张敏灵提出 MIML-KNN[8]算法, 该算法在 K-NN 算法基础上的改进算法并应用于 MIML 学习. 该算法不仅考虑一个样本的近邻(称为 neighbors), 还考虑了以该样本为近邻的样本.

2011 年, 美国的 Nguyen 提出的一种新的解决 MIML 学习问题的 SVM 算法, SISL-MIML 算法[9]. 该算法假设 MIML 样本中的每一个示例只有一个最准确的与之对应的标签, 因为在传统的 MIML 样本退化为 SISL 样本时会丢失信息, 即会标错很多示例的标签.

① 收稿时间:2016-07-24;收到修改稿时间:2016-08-22 [doi:10.15888/j.cnki.csa.005686]

2013 年, 哈尔滨工业大学的 Wu 等人提出一种基于马尔科夫链的 MIML 学习算法[10].

鉴于支持向量机(SVM)[11]在解决小样本、非线性及高维模式识别问题中的优势, 因此在多示例多标签学习框架下的许多算法都采用了支持向量机技术, 如 E-MIMLSVM$^+$算法[12]等.

然而这些算法通常基于退化策略, 该策略会导致退化过程中有效信息的丢失, 降低了分类准确率. ECC 算法[13,14]是对所有的 CC 模型进行集成学习. 虽然在一定程度上改善退化过程中信息丢失, 但当标签数目的增多, 子分类器链长度增加, 使得误差传播问题凸显.

因此本文提出新算法, 改进了 ECC 算法误差传播凸显问题,提高分类准确率. ELDCT-MIMLSVM$^+$算法是在训练过程中依次加入依赖程度最大的标签信息. 主要目的在于减少其他无关标签的干扰, 避免了信息丢失, 同时也降低因增加子分类链而使误差传播凸显的问题, 从而达到较好的分类效果.

本文的组织结构如下: 第二部分介绍相关工作, 第三部分提出改进算法, 第四部分给出了实验结果, 第五部分进行了总结.

2　相关工作

在传统监督学习中, 每个对象由一个示例描述并且隶属于一个概念标签, 在该学习框架下, 对象与示例和标签都是一一对应的关系[15], 学习对象不具有歧义性. 其通过已有的部分输入数据与输出标签生成一个函数, 建立输入输出数据间的关系, 将输入数据映射到合适的输出.

在多示例学习中, 每个对象由多个示例描述并且隶属于一个概念标签, 其主要考查了对象在概念空间的歧义性. 多示例学习在给定的多示例数据集合 $\{(X_1,y_1),(X_2,y_2),...,(X_m,y_m)\}$ 上学习一个映射函数 $f_{MIL}:2^x\to\{-1,+1\}$, 其中 $x_j^{(i)}\in x(j=1,2,...,n_i)$ 是第 i 个包 Xi 中的一个示例, 每个包 $X_i\subseteq X$ 是 n_i 个示例 $\{x_1^{(i)},x_2^{(i)},...,x_{n_i}^{(i)}\}$ 的集合, $y_i\in\{-1,+1\}$ 是包 X_i 的所属类别.

在多标签学习中, 每个对象由单个示例描述并且同时隶属于多个概念标签, 其主要考察对象在语义空间的歧义性. 多标签学习在给定的多标签数据集合 $\{(x_1,Y_1),(x_2,Y_2),...,(x_m,Y_m)\}$ 上学习一个映射函数 $f_{MLL}:x\to2^y$, 其中 x_i 是单个示例描述的对象, $Y_i\subseteq Y$ 是对象 i 所属的类别标签的集合 $\{y_1^{(i)},y_2^{(i)},...,y_{l_i}^{(i)}\}$, $y_k^{(i)}\in Y(k=1,2,...,l_i)$.

在现实生活中, 对象往往同时具有概念歧义性和语义歧义性, 为同时考察这两方面的歧义性, 多示例多标签学习框架应运而生. MIML 既不像多标签学习中那样将对象仅用单一示例描述引起对象概念信息的丢失[16], 也不像多示例学习中那样仅将对象划分到单个预定义的语义类别引起对象语义信息的丢失. 在该框架下, 每个对象由多个示例表示同时隶属于多个概念标签[17], 因此它能够充分考虑到输入输出空间中的歧义性, 对歧义性对象进行有效地学习. 多示例多标签学习在给定的多示例多标签数据集合 $L=\{(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)\}$ 上学习一个从示例集合 X_i 到标签集合 Y_i 上的映射函数 $f:2^x\to2^y$, 其中 $X_i\subseteq X$ 是 n_i 个示例 $x_j^{(i)}\in x(j=1,2,...,n_i)$ 的集合 $\{x_1^{(i)},x_2^{(i)},...,x_{n_i}^{(i)}\}$, $Y_i\subseteq Y$ 是与包 X_i 相关的 l_i 个类别标签 $y_k^{(i)}\in Y(k=1,2,...,l_i)$ 的集合 $\{y_1^{(i)},y_2^{(i)},...,y_{l_i}^{(i)}\}$. 传统监督问题、多示例问题、多标签问题都是多示例多标签问题的特殊表示形式, 上述三种问题都可在多示例多标签框架下进行求解.

目前, 在 MIML 框架下问题的解决方式主要有两种:

一种是基于退化的策略, 将多示例多标签问题退化为多示例或多标签, 再转化为传统监督学习框架下的等价形式进行求解, 如 MIMLSVM[7]、MIMLBOOST[6]等算法, 两者分别以多标签学习和多示例学习为桥梁, 将多示例多标签问题退化为传统监督问题进行求解. 但这两种方式在退化过程中都会引起有效信息的丢失, 导致分类效果不理想, 而且针对大规模机器学习问题时有明显不足;

另一种是考察示例与标签之间的关系, 直接设计针对多示例多标签样本的学习算法, 如 M3MIML 算法[18]. 但这种算法设计难度大, 样本训练时间长, 而且实验证明分类效果不好.

Ying-xin Li 和 Shui-wang Ji 等在果蝇基因表达模式注释的问题中提出基于支持向量机的多示例多标签算法的 MIMLSVM$^+$算法[12], 这是一种针对大规模学习问题提出的算法. 该算法具有较低的训练时间和较好的分类效果, 但其没有考虑标签之间的依赖性, 忽略了标签内在联系, 影响力分类准确率.

3 改进的算法

3.1 MIMLSVM+算法(Multi-instance Multi-label SVM for Large-scale Learning)

MIMLSVM+算法同 MIMLSVM 和 MIMLBOOST 算法相似是一种基于退化的算法, MIMLSVM+将多示例多标签问题退化为多示例单标签问题进行求解, 该算法主要针对大规模的数据问题. MIMLSVM+算法每次为单个标签训练分类器, 收集所有具有该标签的包为正包, 不具有该标签的包为负包, 得到一系列二类分类任务, 每个任务利用支持向量机处理. 为处理类不平衡问题, MIMLSVM+采用不同的惩罚参数分别应用于正类和负类的松弛条件.

3.1.1 MIMLSVM+算法主要包括以下步骤:

① 将多示例多标签退化为二分问题, 对每个二分类问题设计 SVM 算法进行处理

② 处理数据不平衡问题, 在训练过程中优化 SVM, 采用“rescaling”(尺度改变)方法来调节惩罚参数大小.

③ 不同的核函数的计算结果不同, 选取恰当的核函数.

④ 设计总的分类模型.

3.1.2 MIMLSVM+的惩罚参数

对于每个标签 $y \in Y$, 若 $\varphi(X_i, y)=1$, 则说明包 X_i 具有标签 y, 若 $\varphi(X_i, y)=-1$, 表明包 X_i 不具有标签 y. MIMLSVM+的相关优化问题为:

$$\min_{w,b,z} \frac{1}{2}\| w \|^2 + C^+ \sum \varphi(X_i, y)\varepsilon_i + C^- \sum \varphi(X_i, y)\varepsilon_i \quad (1)$$

$$\varphi(X_i, y)(w'\phi(X_i)+b) \geq 1-\varepsilon_i$$

$$\varepsilon_i \geq 0 \ (i=1,2,....,n)$$

其中, $\phi(X_i)$ 是将示例包 X_i 映射到核空间的映射函数, $\varphi(X_i, y)$ 表示包 X_i 是否具有标签 y. ε_i 是 hinge loss, n 是示例包的数目, w 和 b 是用于表示核空间线性描述函数的参数. C^+ 和 C^- 分别为正类和负类的惩罚因子.

3.1.3 MIMLSVM+的改进

为了增强分类模型的容错率和减少训练样本被错误分类, 我们对算法引入非负松弛变量 ξ_{iy}. 引入之后将公式转化为:

$$\min_{w,b,z} \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n} \xi_{iy} T_{iy} \quad (2)$$

经过上述改进, SVM 的优化变为以上问题(MIMLSVM+步骤 2).

3.1.4 核函数

MIMLSVM+算法采用多示例核函数[19] $K_{MI}(X,X')$:

$$K_{MI}(X,X') = \sum_{i-1}^{n} \sum_{j-1}^{m} K^p(x_i x_j) \quad (3)$$

该核函数是通过对标准设置内核函数

$$K_{SET}(X,X') = \sum_{i-1}^{n} \sum_{j-1}^{m} K(x_i x_j) \quad (4)$$

取合适的 p 值得到的, 其中 $p \geq 1$, $K(.,.)$ 为示例级别的内核.

为进一步利用本地视觉特征和空间特征的信息, MIMLSVM+算法重新定义了核函数

$$K_{MI_D}(X_i, X_j) = \frac{1}{n_i n_j} \sum_{(x_{i0}, x_{i1}) \in X_i} \sum_{(x_{k0}, x_{k1}) \in X_j} e^{-\gamma_1 \| x_{i0} - x_{k0} \|^2 - \gamma_2 \| x_{i1} - x_{k1} \|^2} \quad (5)$$

其中 $\| x_{t0} - x_{k0} \|^2$ 衡量了两个图像补丁间本地视觉特征的相似性, $\| x_{ti} - x_{ki} \|^2$ 衡量了两个图像补丁间的的空间距离. 通过调节参数 γ_1 和 γ_2 可对本地视觉特征和空间信息进行显式利用.

MIMLSVM+的核函数十分恰当, 所以我们采用其核函数. 最终的判别函数为:

$$f_y(X) = \sum_{i-1}^{n} \alpha_i \varphi(X_i, y) K_{MI_D}(X_{i,} X) + b_y \quad (6)$$

3.2 ELDCC(Ensembles of Label Dependencies Classifier Chain)

MIMLSVM+算法为每个标签建立一个二类分类器, 忽略了标签之间的联系信息, 退化过程中信息丢失, 因此本篇论文提出了基于标记依赖关系的多示例多标签分类器链算法. 本篇文章采用 ECC 技术对标签间的联系信息加以利用, 并依据某种策略计算标签间的依赖程度值, 根据获得的标签依赖程度组织基分类器链.

ELDCC 主要目标是依据标签间的依赖程度的大小依次训练基分类器, 在训练时依次加入依赖程度最大的标签信息, 以达到较好的分类效果. 这样在保持了 ECC 算法低时间复杂度、低空间复杂度优势的同时, 又能够对标签间的依赖关系加以利用, 进一步提高分类的准确率. 该算法不仅能降低时间、空间复杂度, 还能消除标签顺序对分类的影响.

以下是 ELDCC 算法的主要步骤:

第一步: 计算标签间的依赖程度值.

首先依据表 1 统计数据集中相应样本的数目. 其中 N 表示数据集中的样本总数, 各个变量(a/b/c/d)对应在数据集中与两个标记相关的样本的统计量, 例如 a

表示数据集中同时与标签 i 和标签 j 相关的样本个数.

表 1　标记和的关联表

	l_i	$\neg l_i$	合计
l_j	a	b	$a+b$
$\neg l_j$	c	d	$c+d$
合计	$a+c$	$b+d$	$a+b+c+d=N$

获得表 1 求得的变量值, 依据公式(1)量化标签 i 与标签 j 的依赖程度.

$$\chi^2=\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} \tag{7}$$

第二步: 随机选择 m 个标签作为初始根节点, 利用 prim 算法生成 m 个有序分类器链.

取依赖程度值的倒数, 利用 prim 算法获得最小生成树. 集合 U 是已加入链中的标签, 集合 V 是待加入标签, 每次从集合 V 中选取到集合 U 中依赖程度最大的标签, 并将其加入到集合 U 中, 更新集合 V 到集合 U 中的依赖程度值, 循环直到集合 V 中所有的标签都加入到集合 U 中. 将依次加入集合 U 中标签结点的顺序作为一个分类器链的训练标签的顺序.

第三步: 利用训练样本训练基分类器.

假设共有 L 个标签, 依据第二步获得的标签顺序 $C(c_1,c_2,....c_L)$ 为每个标签构建一个二类分类器, L 个二类分类器组成一个有序分类器链. 分类器链中第一个基分类器的输入为初始样本, 其余基分类器的输入为上一个基分类器的输入样本以及该样本相应输出标签的组合, 即:

$$X_i^{'}=\left[x_{i1},x_{i2},\ldots,x_{in},y_1^{'},y_2^{'},\ldots,y_{k-1}^{'}\right](i=1,2,\ldots,m)$$

其中 X_i 表示第 i 个包, x_{ij} 为包 X_i 中的第 j 个示例, y_k 是第 k 个基分类器的输出标签. 鉴于 ECC 算法是解决单示例多标签问题的算法, 本篇论文采用与 ECC-MIMLSVM$^+$算法相同的策略, 对 ECC 算法加以改造使其适用于解决多示例多标签问题. 即在第 k 个基分类器训练之前, 首先将标签 y_{k-1} 扩展为 d 维向量 $y_{k-1}=(y_{k-1},y_{k-1},.....,y_{k-1})_d^T(k=1,...,L)$, 其中 d 为示例的维数, L 为标签的数目.

然后将扩展后的向量加入到包 $y_{k-1}^{'}$ 中形成新的包 $X_i^{'}=\left[x_{i1},x_{i2},\ldots,x_{in},y_1^{'},y_2^{'},\ldots,y_{k-1}^{'}\right](i=1,2,\ldots,m)$, 进而得到新的数据集合

$$S_k=\{(X_i^{'},\varphi(X_i,y_k))\}(k=1,2,...L),$$

最终基于新的数据集合训练第 k 个基分类器.

第四步: 置信度的计算.

为取得更准确的分类结果, ELDCC 对多个链的输出结果进行汇总. 假设有 m 个有序分类器链: $h_1,h_2.....,h_m$, $\hat{y}_1,\hat{y}_2,.....\hat{y}_m$ 为 m 个有序分类器链上的输出. 根据公式计算该样本在所有分类器链的平均值.

$$\hat{w}_j=\frac{1}{m}\sum_{k=1}^{m}\hat{y}_{j,k} \tag{8}$$

其中 $\hat{y}_{j,k}$ 表示样本在第 k 个分类器链标签 j 的输出值.

第五步: 确定阈值 t.

$$t=\arg\min_t\left\|LCARD(S)-\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}1_{\hat{w}_j\geq t}\right)\right\| \tag{9}$$

$LCARD(S)=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}y_j^i$ 为测试集 S 上每个样本关联的平均标签个数, N 为测试样本数目.

第六步: 计算样本的最终输出.

利用第四步的置信度, 结合第五步的阈值函数计算该样本的最终输出标签 $\hat{y}=(\hat{y}_1,\hat{y}_2,.....\hat{y}_j,....,\hat{y}_L)$. 当第 j 个标签的置信度大于或等于阈值 t 时, 第 j 个标签的最终输出为 1, 表明该样本具有第 j 个标签, 否则该样本在标签 j 上的输出为 0, 表明该样本不具有标签 j.

$$\hat{y}_j=\begin{cases}1 & if \quad \hat{w}_j\geq t\\ 0 & if \quad \hat{w}_j<t\end{cases} \tag{10}$$

对于一个未知标签的测试样本 X, 将样本输入到 m 个分类器链中. 对应其中的一个分类器链, 在进行第 j 个标签预测时, 获得对应链第 j-1 个基分类器的输出值并进行 d 维扩展获得新的样本 $X^{'}=\left[x_1,x_2,...,x_n,\hat{y}_1,\hat{y}_2,...,\hat{y}_{j-1}\right]$, 将该样本代入第 j 个标签的分类函数中 $f_j(X^{'})$, 获得样本 X 关于第 j 个标签的估计值 y_j. 对于所有的有序分类器链都重复上述过程, 将 m 个分类器链的结果进行计算获得置信向量, 利用置信向量和阈值确定样本的最终输出标签.

改进算法的伪代码:

Y=ELDCC-MIMLSVM$^+$(S,X,N,O)

输入: S –训练样本

X –测试样本

N –集合迭代次数

O –样本总数

输出: Y – X 集合预测的标签

1) 将标签数据集 O 随机分 N 次, 得到不同的 n 个数据集 $L_i(i=1,2,\ldots,N)$.

2) 循环训练集 $S=\{(X_i,Y_i)\}(i=1,2,\cdots,n)$, 计算多示

例核矩阵

$$\left[K_{MI}\left(X_i, X_j\right)\right](i, j=1,2,\ldots,n).$$

3) for 每一个有序的标签集

for $L_k\ (k=1,2,\ldots,N)$

(a) 对每一个标签 $y_i \in O(i=1,2,....,l)$ (l 为关于标签集标签的数量)

do $y_i^{'} \leftarrow \left(y_i, y_i, \ldots, y_i\right)_d^T$

$S_i \leftarrow \{\ \}$

for each $\left(X_j, Y_j\right) \in S$

do $X_j^{'} \leftarrow \left[X_j, y_1^{'}, y_2^{'}, \ldots, y_{i-1}^{'}\right]$

$S_i \leftarrow S_i \cup \left(X_j^{'}, \varphi\left(X_j, y_i\right)\right)$

end for

end for

(b) 训练一个 SVM $f_{ki} = SVMTrain(S_i)$ based on $\left[K_{MI}\left(X_i, X_j\right)\right]$, 用公式(3)去训练.

end for

4) 训练每一个测试包 X

(a) for i=1,….,*N*

$\hat{Y}_i \leftarrow \{\}$

for j=1,…..,*l*

$\hat{y}_{ij} \leftarrow f_{ij}(X)$

$\hat{Y}_i \leftarrow \hat{Y}_i \cup \hat{y}_{ij}$

end for

end for

(b) 用 $\hat{Y}_i\ (i=1,2,\ldots,N)$ 计算 $\hat{w}$.

(c) 用公式(8)获得 X 标签: $Y=\left\{y_j \middle| \hat{y}_{ij}=1\right\}$.

4 实验

4.1 实验设置

本文使用周志华等人提供的两个数据集(即表 2 图像样本集和文本样本集特征)进行实验.

表 2 的 scene 数据集是图形图像数据集. 该数据集包含 2000 个场景图像, 其中每个图像都被分配一组标签. 总共 5 个类别, 分别是海、沙漠、山、树和日落. 其中单标签样本数目 1544 个, 约占整个样本集数量 77%左右; 双标签 441, 约占整个样本集数量的 22%左右, 同时属于三个类的样本数目极少. 平均每个样本与 1.24±0.44 个类标签有关联. 每幅图片所对应的多示例多标签样本的示例数为 9, 本文用一个 15 维的特征向量表示每一个示例[20].

表二的 Reuters 是文本数据集, 从 Reuters-21578 样本集[21]中获得. 我们基于最常用的 7 个类, 删除部分只属于一个类别的文本, 再删除其中没有类别和没有正文的文本, 总共得到 8848 个文本. 抽出一部分单标签样本和所有双标签和三标签样本, 得到该样本集所包含 2000 个文本. 属于多个类的文本数占该数据集的 15%左右, 平均每个文本所属的类别数是 1.15±0.37. 每个文本通过滑动窗口术用一组实例向量表示, 滑动窗口的大小设置为 50. 包中的示例采取基于词频的词袋模型进行表示, 将词频为前 3%的词汇予以保留[22], 最终包中的每个示例都由 243 维的特征向量进行表示.

把本文提出的 ELDCT-MIMLSVM$^+$ 算法跟、MIMLSVM$^+$、MIMLBOOST 与 MIMLSVM 算法进行对比. MIMLBOOST 和 MIMLSVM 算法的参数分别根据文献[6]和[12]设置为它们的最佳值, MIMLSVM 的高斯核为 $\gamma = 0.2^2$, MIMLBOOST 的 boosting rounds 的值设为 25. 为了保证实验客观正确, ELDCT-MIMLSVM$^+$ 算法和 MIMLSVM+ 算法的 *gamma* = *le* − 5 . 比较四种算法的平均分类表现.

这四种多示例多标签算法的评价指标采用五个标准的多示例评价指标: one-error、average precision、hamming loss、ranking loss 和 coverage. 对于这五个评价指标, 简单来说 one-error、hamming loss、ranking loss 和 coverage 这四个值越小说明算法效果越好; 而 average precision 则值越大说明算法效果越好.

表 2 图像样本集和文本样本集特征

样本集	样本总数	类别总数	包中示例个数		样本包所含标签数 (*k*)			训练集样本数	测试集样本数
			Min	Max	*k*=1	*k*=2	*k*≥3		
Scene	2000	5	9	9	1544	441	15	1500	500
Reuters	2000	7	2	26	1701	290	9	1500	500

4.2 实验结果

表 3 和表 4 分别显示了四种不同的多示例多标签算法在图像数据集和文本数据集上的实验结果.

表 3　场图形图像样本集实验结果

	ELDCT-MIMLSVM$^+$	**MIMLSVM$^+$**	**MIMLBOOST**	**MIMLSVM**
Hamming Loss	0.190±0.012	0.198±0.004	0.248±0.008	0.331±0.004
One Error	0.317±0.014	0.354±0.011	0.748±0.026	0.712±0.018
Coverage	0.957±0.048	1.091±0.045	2.250±0.047	2.095±0.051
Ranking Loss	0.172±0.010	0.200±0.012	0.998±0.001	0.999±0.001
Average Precision	0.794±0.009	0.769±0.008	0.489\±0.014	0.517±0.017

从表 3 可以看出，在图像数据集上，ELDCT-MIMLSVM$^+$算法略好于 MIMLSVM$^+$，对于 MIMLBOOST 和 MIMLSVM 有明显的优势.

图形图像样本的分类复杂，计算数据量大. MIMLSVM$^+$算法主要是为大规模机器学习设计的算法，采用了 SVM 来提高分类，所以其效率要好于 MIMLBOOST 和 MIMLSVM. ELDCT-MIMLSVM$^+$采用了分类器链技术，也是将问题退化为传统的机器学习. ELDCT-MIMLSVM$^+$与其它算法不同在于，在退化过程中依次加入标签之间的依赖关系，增强了标签之间联系，所以可以取得比较好的效果.

表 4　文本样本集实验结果

	ELDCT-MIMLSVM$^+$	**MIMLSVM$^+$**	**MIMLBOOST**	**MIMLSVM**
Hamming Loss	**0.024±0.005**	0.033±0.011	0.176±0.002	0.170±0.007
One Error	**0.049±0.007**	0.059±0.005	0.540±0.008	0.525±0.042
Coverage	**0.279±0.023**	0.278±0.011	1.553±0.055	1.521±0.071
Ranking Loss	**0.012±0.016**	0.018±0.001	0.130±0.005	0.135±0.024
Average Precision	**0.970±0.014**	0.963±0.004	0.658±0.007	0.667±0.013

从表 4 可以看出，在文本数据集上，ELDCT-MIMLSVM$^+$算法的前四项指标均小于其他各项算法，Average precision 大于其他算法，所有指标在四种算法中表现最佳.

文本分类相对简单，各算法之间差距比较小. SVM 在二分问题上有巨大的优势，ELDCT-MIMLSVM$^+$采用了 SVM 技术，将问题二分然后再求解，体现 SVM 优势，效果明显.

表 5　算法在图像数据集上的训练时间对比

	ELDCT-MIMLSVM$^+$	**MIMLSVM$^+$**	**MIMLBOOST**	**MIMLSVM**
Training(minute)	9.17±0.07	**6.64±0.32**	2994.18±42.26	9.18±0.38
Testing(minute)	5.94±0.04	**0.007±0.011**	354.01±3.87	1.93±0.13

从表 5 可以看出，在图形图像数据集上 ELDCT-MIMLSVM$^+$ 的训练时间和测试时间比 MIMLSVM$^+$ 差，略微优于 MIMLSVM，对 MIMLBOOST 有比较明显的优势.

ELDCT-MIMLSVM$^+$采用分类器链设计模式，当子分类器链增加，耗时就会比较多. 更重要的是 ELDCT-MIMLSVM$^+$比 MIMLSVM 多一步计算标签之间依赖关系，这增加了计算量. 所以在训练时间上会比 MIMLSVM$^+$多.

表 6　四种算法在文本数据集上的训练时间对比

	ELDCT-MIMLSVM$^+$	**MIMLSVM$^+$**	**MIMLBOOST**	**MIMLSVM**
Training(minute)	5.26±0.24	**0.93±0.02**	3009.85±59.17	5.33±0.06
Testing(minute)	2.25±0.17	**0.009±0.01**	675.77±14.62	1.29±0.35

从表 6 可以看出，在文本数据集上，MIMLSVM$^+$ 的训练时间和测试时间最少，ELDCT-MIMLSVM$^+$的训练时间小于 MIMLSVM，但测试时间比 MIMLSVM 多. 效率最低的是 MIMLBOOST.

与在图形图像数据集一样，ELDCT-MIMLSVM$^+$ 多一步计算标签依赖关系，所以耗时较多，但这一步骤增强标签之间联系，提高了分类准确率. 如果在强调准确率而不是对时间有特别严格的要求，ELDCT-MIMLSVM$^+$算法是首选.

5 总结

从上面数据分析可知, 对于图像数据集和文本数据集, 在保证训练时间和测试时间的前提下, ELDCT-MIMLSVM$^+$算法的准确率略优于 MIMLSVM$^+$, 明显高于另外两种算法, 并且 ELDCT-MIMLSVM$^+$在各项指标中表现优异. 因此, 我们可推出结论: ELDCT-MIMLSVM$^+$算法在训练过程中可以依据标签依赖关系, 逐步加入依赖程度较大标签的信息来辅助分类器训练学习, 从而进一步提高分类器分类的准确率.

参考文献

1 Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Informatica, 2007, 31(3): 249–268

2 Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1-2): 31–71.

3 Boutell MR, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757–1771.

4 Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000, 39(2-3): 135–168.

5 Zhou ZH. A Framework for machine learning with ambiguous objects. Proc. Brain Informatics, International Conference(BI 2009). Beijing, China. 2009. 5819. 6–6

6 Zhou ZH, Zhang ML, Huang SJ, et al. Multi-instance multi-label learning. Artificial Intelligence, 2012, 176(1): 2291–2320

7 Zhou ZH, Zhang ML. Multi-instance Multi-label learning with application to scene classification. The Neural Information Processing Systems, 2006: 1609–1616.

8 Zhang ML. A k-nearest neighbor based multi-instance multi-label learning algorithm. 22ND International Conference on Tools with Artificial Intelligence. 2010, 2. 207–212.

9 Nguyen N. A new SVM approach to multi-instance multi-label learning. 2010 IEEE International Conference on Data Mining. 2010. 109.

10 Wu QY, Ng MK, Ye YM. Markov-MIML: A Markov chain-based multi-instance multi-label learning algorithm. Knowledge and Information System, 2013, 37(1): 83–104.

11 Vapnik V. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

12 Li YX, Ji SW, Kumar S, Ye JP, Zhou ZH. Drosophila gene expression pattern annotation through multi-instance multi-label learning. Trans. on Computational Biology and Bioinformatics, 2012: 98–111.

13 Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification. Machine Learning, 2011, 85(3): 254–269.

14 Briggs F, Fern XZ, Raich R. Context-aware MIML instance annotation: exploiting label correlations with classifier chains. Knowledge and Information Systems, 2015, 43(1): 53–79

15 Platt JC. Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods-Support Vector Rning. MIT Press, 1999: 185–208

16 Gărtner T, Flach PA, Smola AJ. Multi-instance kernels. Proc. of the 19th Intenational Conference on Machine Learning. Sydney, Australia. 2002. 179–186.

17 Yakhnenko O, Honavar V. Multi-instance multi-label learning for image classification with large vocabularies. BMVC. 2011. 1–12

18 Pei Y, Fern XZ. Constrained instance clustering in multi-instance multi-label learning. Pattern Recognition Letters, 2014, 37(1): 107–114.

19 Zhang ML, Zhou ZH. M3MIML: A maximum margin method for multi-instance multi-label learning. Proc. of the 8th IEEE International Conference on Data Mining (ICDM'08). Pisa, Italy. 2008. 688–697.

20 Haussler D. Convolution kernels on discrete structures, [Technical Report]. UCSC-CRL-99-10. Dept. of Computer Science, Univ. of California at Santa Cruz. Santa Cruz, CA. July 1999.

21 Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1–47.

22 Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. Advances 696 in Neural Information Processing Systems 15, 2003: 561–568.

23 Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proc. of the 14th International Conference on Machine Learning. Nashville, TN. 1997. 412–420.