基于半监督学习的多示例多标记E-MIMLSVM$^+$算法

李村合,朱红波
LI Cunhe, ZHU Hongbo

中国石油大学(华东) 计算机与通信工程学院,山东 青岛 266580
School of Computer and Communication Engineering, China University of Petroleum, Qingdao, Shandong 266580, China

LI Cunhe, ZHU Hongbo. MIML algorithm E-MIMLSVM$^+$ based on semi-supervised learning. Computer Engineering and Applications, 2018, 54(2):149-154.

Abstract: Multi-instance multi-label learning is a new machine learning framework. In MIML framework, an example is described by multiple instances and associated with multiple class labels. Algorithm MIMLSVM$^+$ decomposes the MIML problem into multiple independent binary classification problems. However, the degeneration process may lose information, which will influence the classification performance. Algorithm E-MIMLSVM$^+$ by using multitask learning techniques is utilized to incorporate label correlations to improve the algorithm MIMLSVM$^+$. In order to make full use of the unlabeled samples to improve the classification accuracy, this paper improves the E-MIMLSVM$^+$ algorithm by using the semi-supervised support vector machine TSVM. In this paper, the algorithm is compared with other MIML algorithms. The experimental results show that the algorithm has achieved good classification results.

Key words: machine learning; Multi-Instance Multi-Label(MIML); Support Vector Machine(SVM); semi-supervised learning

摘 要: 多示例多标记是一种新的机器学习框架,在该框架下一个对象用多个示例来表示,同时与多个类别标记相关联。MIMLSVM$^+$算法将多示例多标记问题转化为一系列独立的二类分类问题,但是在退化过程中标记之间的联系信息会丢失,而E-MIMLSVM$^+$算法则通过引入多任务学习技术对MIMLSVM$^+$算法进行了改进。为了充分利用未标记样本来提高分类准确率,使用半监督支持向量机TSVM对E-MIMLSVM$^+$算法进行了改进。通过实验将该算法与其他多示例多标记算法进行了比较,实验结果显示,改进算法取得了良好的分类效果。

关键词: 机器学习;多示例多标记;支持向量机(SVM);半监督学习
文献标志码: A **中图分类号:** TP393 doi:10.3778/j.issn.1002-8331.1608-0140

1 引言

在机器学习中,多示例多标记学习(Multi-Instance Multi-Label learning,MIML)是一种新的学习框架[1]。传统的监督学习框架是用一个示例表示一个对象,同时该示例对应一个类别标记,利用某种学习算法学得从示例空间到标记空间的一个映射。但是真实世界中的对象往往具有丰富的含义,比如在网页分类中,一篇新闻报道可能同时属于“科技”、“旅游”或“财经”等几个类别,这时只用一个示例来表示一个对象过于简化,在表示阶段就失去了许多有用的信息。

多示例多标记框架将一个对象用多个示例来表示,同时该对象与多个类别标记相关联。与其他机器学习框架相比,多示例多标记框架对于真实世界中对象的表示能力更强,传统的单示例单标记框架其实可以看作是多示例多标记框架的一种简化,因此可以将多示例多标记转化为单示例多标记或者是多示例单标记,进而转化为传统的单示例单标记进行学习[2],但是在转化过程中会丢失有用的信息,比如标记与标记之间的联系信息,造成分类器的学习效果变差。

支持向量机(SVM)是一种有监督的机器学习方法[3],它在本质上是一个二分类的分类算法。SVM基于统计学习理论,目的是要寻找一个间隔最大化的超平

作者简介: 李村合(1966—),男,教授,硕士生导师,研究领域为计算机网络、机器学习与计算智能、智能信息处理;朱红波(1990—),男,硕士研究生,主要研究领域为机器学习与计算智能(机器学习、支持向量机),E-mail:1506094650@qq.com。

收稿日期: 2016-08-02 **修回日期:** 2016-10-08 **文章编号:** 1002-8331(2018)02-0149-06
CNKI **网络优先出版:** 2017-02-10, http://www.cnki.net/kcms/detail/11.2127.TP.20170210.0844.032.html

面,它被广泛地应用在诸如文本分类、图像分类、基因分析和计算机入侵检测等多种应用中。监督学习只是利用有标记的样本进行训练,但是在现实生活中,有标记样本的数目是很少的,而没有标记的样本是大量存在并且比较容易获得的,比如说网络上存在着的大量网页,这些没有标记的样本往往能够提供更加准确的样本分布信息[4]。因此,半监督学习实际上就是在已有监督学习的基础上,结合这些没有标记的样本来训练分类函数,从而获得更好的分类效果。

2 相关工作

在多示例多标记学习框架提出之前,已经有了传统的单示例单标记监督学习框架、多示例学习框架和多标记学习框架。

多示例学习[5](Multi-Instance Learning,MIL)是用一个示例集合(示例包)表示一个对象,该示例集合和一个类别标记相关联,学习的目的是预测未见示例包的合适类别标记,形式化的描述如下[1]:X 表示示例空间,Y 表示类别标记空间,给定数据集 $\{(X_1,y_1),(X_2,y_2),\cdots,(X_m,y_m)\}$,学习任务是学得一个映射函数 $f_{\mathrm{MIL}}:2^X\to Y$,其中,$X_i\subseteq X$ 为一组示例集合 $\{x_{i1},x_{i2},\cdots,x_{in_i}\},x_{ij}\in X(j=1,2,\cdots,n_i)$,$y_i\in Y$ 为示例集合 X_i 对应的一个合适类别标记,n_i 为 X_i 中所含示例的个数。

多标记学习[6](Multi-Label Learning,MLL)是用一个示例表示一个对象,该示例和多个类别标记相关联,学习的目的是预测未见示例的合适类别标记集合,形式化的描述如下[1]:给定数据集 $\{(x_1,Y_1),(x_2,Y_2),\cdots,(x_m,Y_m)\}$,学习任务是学得一个映射函数 $f_{\mathrm{MLL}}:X\to 2^Y$,其中,$x_i\in X(i=1,2,\cdots,m)$ 是示例空间 X 中的一个示例,$Y_i\subseteq Y$ 为示例 x_i 对应的一组合适类别标记 $\{y_{i1},y_{i2},\cdots,y_{i,l_i}\},y_{ik}\in Y(k=1,2,\cdots,l_i)$,$l_i$ 为 Y_i 中所含类别标记的个数。

多示例多标记学习框架学习的目的是要学得一个从示例集合到类别标记集合的映射,该框架形式化的定义如下[1]:给定数据集 $\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_m,Y_m)\}$,目标是学得 $f:2^X\to 2^Y$。其中,$X_i\in X$ 为一组示例集合 $\{x_{i1},x_{i2},\cdots,x_{in_i}\},x_{ij}\in X(j=1,2,\cdots,n_i)$ 是 d 维示例空间 X 中的一个向量,而 $Y_i\subseteq Y$ 为 X_i 所对应的一组合适类别标记集合 $\{y_{i1},y_{i2},\cdots,y_{i,l_i}\},y_{ik}\in Y(k=1,2,\cdots,l_i)$。$l_i$ 为 Y_i 中所含标记的个数,n_i 为 X_i 中所含示例的个数。好的表示往往至关重要,因为如果采用了不合适的表示,那么学习任务从一开始就失去了许多重要的信息,导致整个学习任务变得不容易完成,因此,好的表示从一定程度上决定了学习任务是否能够成功完成[7]。使用MIML对具有复杂含义的多义性对象进行表示,能够更多地利用原始信息,从而有助于学习任务的解决。

在多示例多标记学习框架下已经有了许多适合该框架的算法。基于退化策略,周志华等人提出了MIMLBOOST算法和MIMLSVM算法[8],基于正则化机制以及最大化间隔策略提出了D-MIMLSVM算法[7]和M^{3}MIML算法[9]。MIMLBOOST算法和MIMLSVM算法分别以多示例学习和多标记学习为桥梁将MIML问题转化为传统的监督学习问题进行求解。这两种算法都是基于退化策略的,因此在建模和学习的过程中会丢失对分类有用的信息。D-MIMLSVM算法假设类别标记集合 Y 中共含有 T 个类别标记,同时假设分类系统由 T 个函数 $f=(f_1,f_2,\cdots,f_T)$ 构成,然后为每一个标记建立一个分类器。D-MIMLSVM算法在训练集上定义了一个损失函数 V,使用某种定义在包上的核函数,最后利用CCCP的迭代优化策略对其求解。M^{3}MIML算法假设分类系统包含 T 个线性模型,每个线性模型对应于一个可能的概念类,一个测试样本是否属于第 i 类是由其包含的所有示例在上述对应的某个线性模型上的最大输出决定。M^{3}MIML基于最大化间隔策略,最后得到一个二次规划问题,利用KKT条件求解对偶问题获得模型参数。

张敏灵等人还基于神经网络提出了MIMLRBF算法[10]。第一层神经元是由中心点组成,而中心点则是通过在MIML样本集上调用k-MEDOIDS算法获取的,其中两个包之间的距离度量采用的是Hausdorff距离。隐含层和输出层之间的权值矩阵利用奇异值分解最小化误差平方和函数求出。

Ying-Xin Li等人在研究果蝇基因标注问题时将该问题采用了多示例多标记框架进行训练,提出了两种基于SVM的多示例多标记算法MIMLSVM$^+$和E-MIMLSVM$^+$[11]。MIMLSVM$^+$算法首先基于退化策略将多示例多标记问题退化为一系列二类分类问题,其中每个二类分类问题对应于标记空间 Y 中的一个类别。MIMLBOOST算法和MIMLSVM算法在退化成传统的监督学习问题时使用的是基于示例的高斯核函数,而MIMLSVM$^+$算法在退化过程中使用了基于包的多示例核函数。由于该算法是基于退化策略的,并没有考虑每一个标记之间的联系,在退化过程中就会丢失对分类有用的信息,因此E-MIMLSVM$^+$算法使用了多任务技术[12-14]对MIMLSVM$^+$算法进行了改进,增加了标记间的联系信息,从而获得了更好的分类效果,但同时也增加了训练时间。

3 改进的算法

3.1 MIMLSVM$^+$算法和E-MIMLSVM$^+$算法

给定MIML训练样本集 $D=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_m,Y_m)\}$,类别标记集合 Y。MIMLSVM$^+$算法基于退化策略,为每个类别标记建立一个SVM,从而将多示例多标记问题转化成为 $|Y|$ 个独立的二类分类问题。假

设该训练样本集的样本个数为 n，X_i 是训练集中的第 i 个包，$y\in Y$ 是类别标记空间中的一个标记，$Y_i\subseteq Y$ 是示例集合 X_i 对应的类别标记集合。$\phi(X_i,y)$ 作为指示函数，对于任意的 $y\in Y,\phi(X_i,y)=+1$ 当且仅当 $y\in Y_i$，否则 $\phi(X_i,y)=-1$。由此得到的SVM分类模型为：

$$\min_{w_y,b_y,\xi_{iy}}\frac{1}{2}\|w_y\|^2+C\sum_{i=1}^{n}\xi_{iy}\tau_{iy}$$
$$\text{s.t. }\phi(X_i,y)\left(\langle w_y,\varphi(X_i)\rangle+b_y\right)\geqslant 1-\xi_{iy}$$
$$\xi_{iy}\geqslant 0(i=1,2,\cdots,n) \tag{1}$$

其中，w_y 和 ξ_{iy} 都是要优化的参数。w_y 代表的是该超平面的法线方向，$\|w_y\|^2$ 反应了模型的复杂度，y 是示例集合 X_i 对应的类别标记。函数 $\varphi(X_i)$ 可以将包 X_i 从输入空间映射到某个高维特征空间，ξ_{iy} 是松弛变量，对应样本点允许偏离的函数间隔的量。参数 C 是用来平衡模型复杂度和训练样本误差的权重。如果不同类别的训练样本数目差别较大就会影响到分类效果，甚至使分类器没有价值。通常使用“再缩放”[15]的策略处理不平衡问题，即上式中的“缩放因子” τ_{iy}。τ_{iy} 的定义如下：

$$\tau_{iy}=\frac{1+\phi(X_i,y)}{2}R_y+\frac{1-\phi(X_i,y)}{2} \tag{2}$$

其中，R_y 是类别 y 的不平衡级别，是根据训练集中正样本的个数和负样本的个数估计出来的。可以看出，当 $\phi(X_i,y)=+1$ 时 τ_{iy} 就是 R_y，因此正包的惩罚损失被 R_y 放大了。

核函数的使用使得SVM避免了直接在高维空间中的复杂计算，使用不同的核函数也会产生不同的分类效果。MIMLSVM⁺算法使用了基于包的多示例核函数 $K(X_i,X_j)=<\phi(X_i),\phi(X_i)>$，其定义如下：

$$K_{\mathrm{MI}}(X_i,X_j)=\frac{1}{n_in_j}$$
$$\sum_{(x_{is\_0},x_{is\_1})\in X_i}\sum_{(x_{jz\_0},x_{jz\_1})\in X_j}\mathrm{e}^{-\gamma_1\|x_{is\_0}-x_{jz\_0}\|^2-\gamma_2\|x_{is\_1}-x_{jz\_1}\|^2} \tag{3}$$

其中，n_i 和 n_j 分别表示包 X_i 和 X_j 中的示例个数，$\|x_{is\_0}-x_{jz\_0}\|^2$ 用来衡量两个示例间的表达模式视觉特征相似性，$\|x_{is\_1}-x_{jz\_1}\|^2$ 衡量了两个示例的空间距离，而 γ_1 和 γ_2 分别表示视觉信息和空间信息的权重。

求解式(1)，最终的分类模型为：

$$f_y(X)=\langle w_Y,\varphi(X)\rangle+b_y=$$
$$\sum_{i=1}^{n}a_{iy}\phi(X_i,y)K_{\mathrm{MI}}(X_i,X)+b_y \tag{4}$$

上述退化的过程中，没有考虑标记之间的联系信息，因此E-MIMLSVM⁺算法引入多任务学习技术[12-14]对MIMLSVM⁺算法进行了改进。假设学习所得标记 $y\in Y$ 的分类模型为第 y 个任务，并且标记 y 的分类函数为：

$$f_y(X)=\langle w_Y,\varphi(X)\rangle+b_y=\langle(w_0+v_y),\varphi(X)\rangle+b \tag{5}$$

其中，w_0 表示不同任务间的共性，而 v_y 则用来表示任务 y 跟其他任务的区别。多任务学习的目标就是训练优化 w_0，$v_t(t=1,2,\cdots,|Y|)$ 和 b 的值从而利用标记之间的联系信息，获得更好的分类效果，其优化问题变为：

$$\min_{w_0,v_y,b_y,\xi_{iy}}\frac{1}{2}\left(\sum_{y\in Y}\|v_y\|^2+\mu\|w_0\|^2\right)+C\sum_{y\in Y}\sum_{i=1}^{n}\xi_{iy}\tau$$
$$\text{s.t. }\phi(X_i,y)\left(\langle(w_0+v_y),\varphi(X_i)\rangle+b\right)\geqslant 1-\xi_{iy}$$
$$\xi_{iy}\geqslant 0 \tag{6}$$

其中，参数 μ 用来调节参数 w_0 和参数 v_y 之间的关系，即平衡各个任务的相似性。通过求解优化问题式(6)，分类函数变为：

$$f_y(X)=\sum_{i=1}^{n}\sum_{t\in Y}\alpha_{it}\phi(X_i,t)K_{ty}(X_i,X)+b \tag{7}$$

其中，K_{ty} 是核函数，其定义如下：

$$K_{ty}(X_i,X_j)=\left(\frac{1}{\mu}+\delta(t=y)\right)\langle\varphi(X_i),\varphi(X_j)\rangle=$$
$$\left(\frac{1}{\mu}+\delta(t=y)\right)K_{\mathrm{MI}}(X_i,X_j)\ (t,y\in Y) \tag{8}$$

其中，当任务 t 和任务 y 是同一任务时，即 $t=y$ 时，$\delta(t=y)=1$，否则 $\delta(t=y)=0$。MIMLSVM⁺算法中的核函数衡量的是同一个任务中两个包的相似性，而 K_{ty} 则可以用来衡量任务 t 和任务 y 中两个包的相似性。

从式(7)建立的模型可以看出，每一个分类模型 f_y 都有一个共同的参数 w_0，也就是模型假设每一个标记之间都是有关联的，但是这可能与实际的情况并不相符，因为有的标记之间可能并不存在关系。因此可以先在标记空间中聚类，从而将标记空间划分为一些具有关联性的标记子集，每一个示例包和标记之间的标记指示矩阵可表示为 $Y=\left[\phi(X_i,y)\right]_{n\times|y|}$。在聚类时使用的是 Y 的列上计算出的皮尔逊相关系数来衡量标记间联系。如果标记空间的聚类个数 K 等于标记空间 Y 中的标记数，即 $K=|Y|$，在这种情况下E-MIMLSVM⁺算法会退化成MIMLSVM⁺算法。

3.2 基于半监督学习的E-MIMLSVM⁺算法TMIMLSVM⁺

在实际问题中想要获取大量有标记的示例往往是很难的，因为获得这些有标记示例可能需要耗费很高的代价。如果只使用少量的有标记示例，那么利用它们所训练出的分类系统往往很难具有很强的泛化能力。而在真实问题中大量的未标记示例是容易获得的，这些未标记示例能够提供更加准确的关于示例分布的信息。因此，在有标记示例数目较少时，可以充分利用大量的未标记示例进行训练，从而可以提高分类器的学习能力

和泛化能力。

半监督SVM基于SVM和半监督学习的聚类假设[15],试图找到可以将两类有标记样本分开,且穿过数据低密度区域的划分超平面[4],这样就可以同时利用有标记样本和未标记样本。

半监督SVM中最著名的是TSVM(Transductive Support Vector Machine,TSVM)[16],它试图对所有未标记样本进行各种标记,然后在所有样本中训练SVM找到间隔最大的分类超平面。TSVM的形式化的定义如下:假设给定有标记的样本集 $D_l=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, $x_i\in\mathbf{R}^n$, $y_i\in\{-1,+1\}$,给定未标记的样本集 $D_u=\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$,那么TSVM的目标优化为:

$$\begin{aligned}&\min_{w,b,y,\xi_i}\frac{1}{2}\|w\|^2+C_l\sum_{i=1}^{l}\xi_i+C_u\sum_{i=l+1}^{m}\xi_i\\&\text{s.t. } y_i(w\cdot x_i+b)\geqslant 1-\xi_i,i=1,2,\cdots,l\\&\quad y_i^*(w\cdot x_i+b)\geqslant 1-\xi_i,j=l+1,l+2,\cdots,m\\&\quad \xi_i\geqslant 0,i=1,2,\cdots,m\end{aligned}\tag{9}$$

其中,$\xi_i(i=1,2,\cdots,l)$是有标记样本的松弛变量,$\xi_i(j=l+1,l+2,\cdots,m)$则对应于未标记样本的松弛变量,$C_l$、$C_u$用于平衡模型复杂度、有标记样本和未标记样本重要程度的平衡参数,由用户指定,为了使有标记样本发挥更重要的作用,开始时C_u要设置得比C_l的值小。

TSVM的主要训练步骤如下:

(1)忽略未标记样本,利用有标记样本训练出一个初始SVM,利用这个SVM对未标记样本进行标记指派得到 y_i^*。

(2)将未标记样本的指派标记 y_i^* 带入式(9),得到了一个标准的SVM问题,于是可求解出新的划分超平面和松弛变量。

(3)注意到此时未标记样本的标记可能已经发生了变化,从正负两个类中分别找出标记发生错误的样本,交换它们的标记,再重新求解出新的划分超平面和松弛变量,重复(3)直至标记指派调整完成。

(4)逐渐增大C_u以提高未标记样本对优化目标的影响,进行下一轮标记指派。直至$C_u=C_l$。

为了充分利用未标记样本提供的样本空间分布信息,这里利用TSVM对E-MIMLSVM$^+$算法进行改进得到TMIMLSVM$^+$,使其能够提高分类准确率,则优化问题变为:

$$\begin{aligned}&\min_{w_y,b_y,\xi_{iy}}\frac{1}{2}\left(\sum_{y\in Y}\|v_y\|^2+\mu\|w_0\|^2\right)+C_l\sum_{y\in Y}\sum_{i=1}^{l}\xi_{iy}\tau_{iy}+\\&\qquad C_u\sum_{y\in Y}\sum_{i=1+1}^{m}\xi_{iy}\tau_{iy}\\&\text{s.t. }\phi(X_i,y)\left(\left\langle(w_0+v_y),\varphi(X_i)\right\rangle+b_y\right)\geqslant 1-\xi_{iy}\\&\quad\phi(X_i^*,y)\left(\left\langle(w_0+v_y),\varphi(X_i^*)\right\rangle+b_y\right)\geqslant 1-\xi_{iy}\\&\quad\xi_{iy}\geqslant 0(i=1,2,\cdots,m)\end{aligned}\tag{10}$$

但是注意到,可能会出现样本不可分的情况,比如某个样本在所有分类器上的输出都是负数,因此该样本没有与之对应的类别标记。出现这种情况时,采用T-Criterion[17]准则来进行预测,即用所有输出为正的分类器对应的类别来标记样本,当出现分类器输出都是负值时,用最大输出值对应的类别来标记样本。

TMIMLSVM$^+$的算法描述如下:

$Y=TMIMLSVM^+(S,X,C_l,C_u)$

输入:有标记样本集:$S=\{(X_i,Y_i)|1\leqslant i\leqslant N\}$,测试集$X=\{X_i^*|1\leqslant i\leqslant M\}$,平衡参数$C_u$、$C_l$

输出:Y测试集X的预测类别

1. 根据标记指示矩阵$Y=\left[\phi(X_i,y)\right]_{n\times|y|}$将标记空间里的标记重新划分为个$k$个子集$L=L_1\cup L_2\cup\cdots\cup L_K$,对于每一个标记子集$L_k(k=1,2,\cdots,K)$可以得到该标记子集的训练样本集$S_k=\{(X_i,Y_i\cap L_k)|i=1,2,\cdots,N\}$

2. 基于S_k计算多任务的多示例核函数矩阵$\left[K_h(X_i,X_j)\right](i,j=1,2,\cdots,N;s,t\in L_k)$

3. 在核矩阵$\left[K_h(X_i,X_j)\right]$上使用式(10)训练初始的分类函数$f_y=SVMTrain(S_k),y\in L_k$,使用$f_y$对$X$进行预测

4. 得到标记指派$y^*=(\hat{y}_{l+1},\hat{y}_{l+2},\cdots,\hat{y}_{l+u})$,初始化$C_u<<C_l$,

while $C_u<C_l$ do

4.1 基于S、X、C_u、C_l、y^*求解式(10)得到新的划分超平面和松弛变量

4.2 while $\exists\{i,j|(\hat{y}_i\hat{y}_j<0)\wedge(\xi_i>0)\wedge(\xi_j>0)\wedge(\xi_i+\xi_j>2)\}$ do

$\hat{y}_i=-\hat{y}_i,\hat{y}_j=-\hat{y}_j$使用新的标记指派$y^*$以及$S$、$X$、$C_u$、$C_l$,求解式(10)

end while

4.3 $C_u=\min\{C_u,C_l\}$

end while

5. 对于未知标记的样本集X,使用T-Criterion准则的标记预测函数为:

$$Y=\left\{\arg\max_{y\in L}f_y(X)\middle|f_y(X)<0,\forall y\in L\right\}\cup\left\{y\middle|f_y(X)\geqslant 0,y\in L\right\}$$

4 实验

4.1 实验设置

为了评估改进算法的分类效果,使用周志华等人提供的场景样本集和文本样本集进行实验。第一个样本集是场景样本集,这个样本集由2 000个自然场景图片构成,来源于COREL图像库和互联网。场景样本集上的图片共分为5个类别,分别是日落、海、沙漠、树、和山,样本集中有多个类别标记的图片超过了22%,平均来说,每一张图片与1.24个标记有关,如表1所示。其中,每一张图片用一个包来表示,每个包中包括9个示

表2 场景样本集上实验结果

Metric	MIMLSVM	MIMLSVM+	E-MIMLSVM+	TMIMLSVM+
hamming loss	0.236 0±0.013 9	0.237 8±0.022 9	0.239 6±0.019 9	0.213 5±0.026 5
one-error	0.332 0±0.027 9	0.427 5±0.051 0	0.414 0±0.040 3	0.325 3±0.017 2
coverage	1.094 5±0.066 7	1.229 5±0.133 4	1.182 5±0.121 5	1.028 9±0.035 5
ranking loss	0.257 9±0.019 0	0.238 7±0.025 7	0.229 3±0.024 7	0.181 4±0.032 7
average precision	0.726 6±0.021 8	0.722 1±0.028 7	0.731 4±0.027 0	0.783 5±0.036 8

表3 文本样本集上实验结果

Metric	MIMLSVM	MIMLSVM+	E-MIMLSVM+	TMIMLSVM+
hamming loss	0.172 9±0.013 3	0.038 4±0.005 1	0.041 9±0.006 5	0.023 0±0.002 6
one-error	0.534 5±0.040 2	0.064 0±0.015 8	0.064 5±0.015 7	0.051 3±0.031 8
coverage	1.572 5±0.088 3	0.289 5±0.060 2	0.290 0±0.064 1	0.272 1±0.054 3
ranking loss	0.238 1±0.017 8	0.021 0±0.005 5	0.021 2±0.006 3	0.021 1±0.005 6
average precision	0.658 4±0.021 8	0.959 5±0.008 7	0.959 3±0.009 6	0.968 2±0.027 1

例,每个示例通过SBN方法[18]用一个15维的特征向量表示。

表1 场景样本集特征

标记集合	图片数量	标记集合	图片数量
沙漠	340	山+日落	19
山	268	山+树	106
海	341	海日落	172
日落	213	海+树	14
树	378	日落+树	28
沙漠+山	19	沙漠+山+日落	1
沙漠+海	5	沙漠+日落+树	3
沙漠+日落	21	山+海+树	6
沙漠+树	20	山+日落+树	1
山海	38	海+日落+树	4

第二个样本集是文本样本集,这个样本集来源于被广泛研究的Reuters-21578[19]。去除没有标记和没有正文的文本,再随机去除一些只有一个类别标记的文本,最终得到2 000个文本样本,分为7个类别。其中大约有15%的文本样本具有多个类别标记,平均每个文本样本与1.15±0.37个类别标记相关联。通过使用滑动窗口[20]技术,每一篇文档用一个包表示,每个包中包括一组243的特征向量,每一个向量代表了这篇文档的某一个部分。

为了评估本文改进的E-MIMLSVM+算法,将这个算法同MIMLSVM、MIMLSVM+、E-MIMLSVM+进行了比较。其中MIMLSVM、MIMLSVM+、E-MIMLSVM+算法中的参数分别根据文献[8,11]中的实验设置为最优,即MIMLSVM的高斯核设置为$\gamma=0.2^2$、聚类个数k设置为训练样本个数的20%,MIMLSVM+和E-MIMLSVM+算法的两个高斯核函数的参数为$\gamma_1=10^{-5}$和$\gamma_2=10^{-2}$。

4.2 实验结果

对于"分类"任务而言,采用基于样本的评价指标hamming loss、one-error、coverage、ranking loss和average precision[7]。简单来说,对于hamming loss、one-error、coverage和ranking loss的值越小说明算法效果越好;对于average precision值越大说明算法效果越好。实验采用10折交叉验证,表2和表3分别显示了MIMLSVM、MIMLSVM+、E-MIMLSVM+和TMIMLSVM+算法在场景样本集和文本样本集上的实验表现。

可以看出,总体上文本样本集上的分类效果要比场景样本集上的分类效果好,在每一个样本集内,改进算法的性能要优于其他的多示例多标记分类算法。

5 总结

对于多义性对象的学习,多示例多标记框架因为具有更强的表达能力,从而表现出了良好的分类效果,本文介绍了该框架下基于退化策略并且使用SVM分类的MIMLSVM+算法和E-MIMLSVM+算法。通过在E-MIMLSVM+算法中引入可以充分利用未标记样本进行学习和分类的半监督支持向量机TSVM,从而对该算法进行了改进,提高了分类准确率,增强了分类器的泛化能力。

参考文献:

[1] Zhou Zhihua,Zhang Minling,Huang Shengjun,et al.MIML:a framework for learning with ambiguous objects[J].CORR abs/0808.3231,2008.

[2] Zhou Z H,Zhang M L,Huang S J,et al.Multi-instance multi-label learning[J].Artificial Intelligence,2012,176(1):2291-2320.

[3] Tong S,Koller D.Support vector machine active learning with applications to text classification[C]//Seventeenth International Conference on Machine Learning,2000,2(1):999-1006.

[4] Zhou Zhihua.Disagreement-based semi-supervised learning[J].Acta Automatica Sinica,2013,39(11):1871-1878.

[5] Dietterich T G,Lathrop R H,Lozano-Pérez T.Solving the multiple-instance problem with axis-parallel rectan-

gles[J].Artificial Intelligence,1997,89(1/2):31-71.
[6] Tsoumakas G,Katakis I.Multi-label classification:an overview[J].International Journal of Data Warehousing and Mining,2007,3(3):1-13.
[7] Zhou Z H,Zhang M L,Huang S J,et al.Multi-instance multi-label learning[J].Artificial Intelligence,2011,176(1):2291-2320.
[8] Zhou Z H,Zhang M L.Multi-instance multi-label learning with application to scene classification[C]//Advances in Neural Information Processing Systems,2006:1609-1616.
[9] Zhang M L,Zhou Z H.M3MIML:a maximum margin method for multi-instance multi-label learning[C]//Proceedings of the 8th IEEE International Conference on Data Mining(ICDM'08),Pisa,Italy,2008:688-697.
[10] Zhang Minling,Wang Zhijian.MIMLRBF:RBF neural networks for multi-instance multi-label learning[J].Neurocomputing,2009.
[11] Li Y X,Ji S W,Kumar S,et al.Drosophila gene expression pattern annotation through multi-instance multi-label learning[J].Transactions on Computational Biology and Bioinformatics,2012,9(1):1445-1450.
[12] Evgeniou T,Pontil M.Regularized multi-task learning[C]//Proc 10th ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining,2004:109-117.
[13] Zhang J,Ghahramani Z,Yang Y.Flexible latent variable models for multi-task learning[J].Machine Learning,2008,73(3):221-242.
[14] Evgeniou T,Micchelli C A,Pontil M.Learning multiple tasks with kernel methods[J].Machine Learning Research,2005,6(4):615-637.
[15] 周志华.机器学习[M].北京:清华大学出版社,2016.
[16] Joachims T.Transductive inference for text classification using support vector machines[C]//Proceedings of the Sixteenth International Conference on Machine Learning,1999,117(827):200-209.
[17] Boutell M R,Luo J,Shen X,et al.Learning multi-label scene classification[J].Pattern Recognition,2004,37(9):1757-1771.
[18] Maron O,Ratan A L.Multiple-instance learning for natural scene classification[C]//Proceedings of the 15th International Conference on Machine Learning,1998:341-349.
[19] Sebastiani F.Machine learning in automated text categorization[J].Computer Science,2015,34(1):1-47.
[20] Andrews S,Tsochantaridis I,Hofmann T.Support vector machines for multiple-instance learning[C]//Advances in Neural Information Processing Systems,2003:561-568.

(上接130页)

[5] Zheng Y,Imai H.How to construct efficient signcryption schemes on elliptic curves[J].Information Processing Letters,1998,68(5):227-233.
[6] Malonelee J,Mao W.Two birds one stone:signcryption using RSA[C]//Proceedings of RSA Conference on Cryptology Track.Berlin/Heidelberg:Springer-Verlag,2003:211-226.
[7] Li C K,Yang G,Wong D S,et al.An efficient signcryption scheme with key privacy and its extension to ring signcryption[J].Journal of Computer Security,2010,18(3):451-473.
[8] Libert B,Quisquater J J.A new identity based signcryption scheme from pairings[C]//Proceedings of IEEE Information Theory Workshop,2003:155-158.
[9] Chow S S M,Yiu S M,Hui L C K,et al.Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity[C]//Proceedings of International Conference on Information Security and Cryptology(ICISC 2003),2003:352-369.
[10] Boyen X.Multipurpose identity-based signcryption:a Swiss army knife for identity-based cryptography[C]//Proceedings of Advances in Cryptology(CRYPTO 2003),2003:383-399.
[11] Chen L,Malonelee J.Improved identity-based signcryption[C]//Proceedings of Conference on Public Key Cryptography(PKC 2005).Berlin/Heidelberg:Springer-Verlag,2005:362-379.
[12] Barreto P S L M,Libert B,Mccullagh N,et al.Efficient and provably-secure identity-based signatures and signcryption from bilinear maps[C]//Proceedings of International Conference on Theory and Application of Cryptology and Information Security.Berlin/Heidelberg:Springer-Verlag,2005:515-532.
[13] Sun Yinxia,Li Hui.Efficient signcryption between TPKC and IDPKC and its multi-receiver construction[J].Sciece China Information Sciences,2010,53(3):557-566.
[14] Li F,Zhang H,Takagi T.Efficient signcryption for heterogeneous systems[J].IEEE Systems Journal,2013,7(3):420-429.
[15] Huang Q,Wong D S,Yang G.Heterogeneous signcryption with key privacy[J].Computer Journal,2011,54(4):525-536.
[16] 张玉磊,李臣意,周冬瑞,等.高效的可撤销无证书签名方案[J].计算机工程,2015,41(7):157-162.