第58卷 第1期
2022年1月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 58, No. 1
Jan., 2022

DOI:10.13232/j.cnki.jnju.2022.01.010

多示例学习的两阶段实例选择和自适应包映射算法

杨 梅[1],曾雯喜[1],方 宇[1],闵 帆[1,2*]

(1.西南石油大学计算机科学学院,成都,610500;2.西南石油大学人工智能研究院,成都,610500)

摘 要:多示例学习(Multi-Instance Learning,MIL)研究对象的内部结构比单示例学习更加复杂.已有的MIL方法大都基于原始空间中的实例进行包映射,但这些方法通常忽略包的内部结构信息,难以保证所选实例与包在新特征空间中的关联性.提出一种多示例学习的两阶段实例选择和自适应包映射(TAMI)算法.首先,实例选择技术根据包中实例的密度值和关联性,挖掘包内结构特征,选取实例原型;其次,实例选择技术选取具有峰值密度的实例原型作为代表实例;最后,自适应包映射技术通过定义新的映射函数将包转换为单向量进行学习.实验利用显著性检验从统计学的角度验证了TAMI在图像检索、文本分类等基本数据集上的有效性.结果表明,TAMI在图像检索和医学图像数据集上取得了比其他MIL算法更好的效果,并在文本分类数据集上表现良好.

关键词:自适应映射,关联性,密度,实例选择,多示例学习

中图分类号:TP181 **文献标志码:**A

Two-stage instance selection and adaptive bag mapping algorithm for multi-instance learning

Yang Mei[1], Zeng Wenxi[1], Fang Yu[1], Min Fan[1,2*]

(1. School of Computer Science, Southwest Petroleum University, Chengdu, 610500, China;
2. Institute for Artificial Intelligence, Southwest Petroleum University, Chengdu, 610500, China)

Abstract: Compared with single-instance learning, multi-instance learning (MIL) has a more complex internal structure of its research objects. Most of the existing MIL methods map bags based on instances in the original space. They hardly consider the internal structure information of the bags. It is difficult to guarantee the affinity between the selected instance and the bag in the new feature space. In this paper, we propose a two-stage instance selection and adaptive bag mapping algorithm for multi-instance learning (TAMI) to handle this issue. Firstly, the first-stage instance selection technique excavates structural features and selects instance prototypes based on the density and affinity of the instances in the bag. Secondly, the second-stage instance selection technique chooses instance prototypes with the peak density as representatives. Finally, the new adaptive bag mapping technique converts each bag into a single vector. Experiment verifiy the effectiveness of TAMI on the basic dataset from a statistical point of view. The results show that TAMI has achieved better results than other MIL algorithms on image retrieval and medical image datasets, and it performs well on text classification datasets.

Key words: adaptive mapping, affinity, density, instance selection, multi-instance learning

多示例学习(Multi-Instance Learning,MIL)由Dietterich et al[1]提出,目的是通过分析已知分子的集合来预测新分子是否可以用来制造某种药物.与单示例学习相比,MIL的输入是一组带有

基金项目:国家自然科学基金(62006200),四川省自然科学基金(2019YJ0314),四川省青年科学技术创新团队(2019JDTD0017),西南石油大学研究生全英文课程建设项目(2020QY04)

收稿日期:2021-06-28

*通讯联系人,E-mail:minfan@swpu.edu.cn

标签的包,而不是带有标签的实例,每一个包是由多个实例组成的集合.对于二分类问题,如果一个包至少包含一个正实例,则该包为正,反之为负.近年来,多示例学习被广泛地应用于多个领域,如图像分类[2-4]、医学诊断[5-6]、文本分类[7-8]和情感分析[9-10].

将包从原始空间转换到新特征空间进行学习的分类算法是近年来MIL的研究热点之一.Constructive Clustering-Based Ensemble (CCE)[11]首先将所有包中的实例聚类为d簇;然后将包映射为d维单向量,如果包中实例属于第i个簇,则第i个特征值设置为1,否则设置为0. Multi-Instance Learning via Embedded Instance Selection (MILES)[12]首先构建包含所有训练实例的特征空间,然后通过特征空间中实例与包中实例的相似关系,将包转化为单向量. Image Annotation by Multi-Instance Learning with Discriminative Feature Mapping and Selection (MILFM)[13]首先从正负包中选择原型实例并从负包中选择聚类实例,再利用包与正负实例原型的相关性进行映射.然而,上述算法没有考虑包内部的结构特征,且只能利用固定的空间关系进行映射.

本文提出一种多示例学习的两阶段实例选择和自适应包映射算法(The Two-stage Instance Selection and Adaptive Bag Mapping Algorithm for Multi-Instance Learning,TAMI).实例选择技术分两个阶段来选择数据空间的代表实例:第一阶段,利用包中实例的密度值和关联性,分析包内结构特征,选取实例原型;第二阶段,根据实例原型分布的紧密程度,从中选出具有峰值密度[14]的实例作为代表实例.自适应包映射技术基于包与代表实例的自适应距离关系,通过差值处理将其转化为单向量.实验结果表明,TAMI在图像检索和医学图像数据集上取得了比其他MIL算法更好的效果,并在文本分类数据集上表现良好.

针对MIL分类任务,本文的主要贡献如下:

(1)提出一种高效的两阶段实例选择技术:第一阶段,实例选择技术最大限度地保留了包的内部结构特征,同时大幅度地缩减了代表实例的求解范围;第二阶段,实例选择技术从实例原型中寻找更具有代表性的实例,使得包在新特征空间更具有可区分性.

(2)提出一种自适应的包映射函数.包中实例与代表实例间以最佳相似度进行映射,使新的特征空间能最大程度地保留原始空间的特征.

1　相关工作

根据不同的处理策略,MIL算法主要包括基于统计的算法、基于包的算法、基于核的算法和基于实例的算法.

1.1　基于统计的算法　基于统计的算法使用一个或多个统计变量来对包进行量化,试图捕获包中实例的特征分布. Simple-Multi Instance (Simple-MI)[15]统计包中所有实例属性的平均值,即用均值向量来代表该包,具有简单快速的特点.

1.2　基于包的算法　基于包的算法在包空间寻找最具信息量的包节点,然后利用包与包之间的距离关系进行映射. Bag Level Multi-Instance Clustering (Bamic)[16]在训练包上建立聚类模型,将训练包分成k簇;根据k个簇的中心,将每个包转换成单向量.向量的第j个属性值则是该包到第j个簇中心的距离.

1.3　基于核的算法　基于核的算法专注于设计或使用核来进行实例之间的相似性描述,例如高斯核、Fisher核或隔离核. Multi-Instance Graph (miGraph)[8]考虑包内实例的关联性,设计了一种直接计算包与包之间相似性的graph核. Multi-Instance Learning Based on the Fisher Vector Representation (miFV)[8]使用预训练的高斯混合模型获取不同分布的权重、均值和协方差,然后基于这些信息和Fisher核将每个包编码为高维单向量. Isolation Set-Kernel (ISK)[18]从训练实例空间随机取样并构建多个维诺图,使用隔离核为包中的实例分配权重,最终基于包中带权重的实例和维诺图进行映射.

1.4　基于实例的算法　基于实例的算法在实例级进行处理.首先,通过无监督学习、指标分析或其他方法在实例空间中找到关键实例;其次,通过包与关键实例之间的距离,将每个包映射为单个向量. Multi-Instance Learning Based on the Vector of Locally Aggregated Descriptors Representation (miVLAD)[19]算法在整个实例空间中,通过

k-Means 获得聚类中心,然后使用映射函数将包转化为单向量. 单向量的维度与 k-Means 聚类中心数量直接关联,即 k 值越大,映射向量的维度越高. Discriminative Mapping Approach for Multi-Instance Learning (MILDM)[20]在整个实例空间构建满足关键实例判别标准的映射函数,尽可能保留包在新特征空间中的可区分性,然而算法的时间开销与实例空间的大小呈平方关系. Multi-Instance with Key Instance shift (MIKI)[21]首先训练一个加权的多类模型以选择具有高正性的实例作为原型实例,然后将包转换为具有原型实例信息的向量,其次将原型实例的权重合并到转换后的包向量中,以缩小训练/测试分布之间的差距,最后使用新的包向量及其权重来学习.

基于实例的算法在很大程度上保留了实例的信息,因此在特定的数据集上有更好的效果.

2 两阶段实例选择和自适应包映射算法

本算法的主要思想:通过两阶段实例选择技术快速选取代表实例,再采用自适应的映射方法将包转换成单示例. 以图像数据集的处理过程为例,如图 1 所示(图片来源为 Corel 数据集):第一阶段,计算原始数据包中实例的密度值和关联性,分析其内部结构特征,选出实例原型;第二阶段,从实例原型池中选取出具有峰值密度特点的代表实例;最后,根据包中实例与代表实例的相似性,通过自适应差值处理,实现包映射.

图 1 TAMI 算法的运行实例

Fig. 1 Running example of TAMI algorithm

2.1 符号系统 令 $\mathcal{T}=\{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_i,\cdots,\boldsymbol{X}_n\}$ 表示给定数据集,其中 $\boldsymbol{X}_i=\{\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{ij},\cdots,\boldsymbol{x}_{in_i}\}$ 表示一个包,$y_i\in\{-1,+1\}$ 表示 $\boldsymbol{X}_i$ 的标签,n 表示 $\mathcal{T}$ 中包的个数,n_i 则表示 $\boldsymbol{X}_i$ 中实例的数量. MIL 的任务就是在标准多示例的假设下,通过对已有的数据进行分析得到一个有效的分类模型,从而实现对未知包的标签预测.

2.2　两阶段实例选择技术　两阶段实例选择技术首先在包内选取实例原型,然后在实例原型池中选出代表实例.

2.2.1　实例原型选择　第一阶段实例选择技术基于包内实例的优先级,选出具有包内部结构特征的实例原型.令$\boldsymbol{p}_i$表示$\boldsymbol{X}_i$中所有实例的优先级,如式(1)所示:

$$\boldsymbol{p}_i=\boldsymbol{\rho}_i\times\boldsymbol{s}_i \tag{1}$$

其中,$\boldsymbol{\rho}_i$表示$\boldsymbol{X}_i$中实例的密度值向量,$\boldsymbol{s}_i$表示$\boldsymbol{X}_i$中实例的关联值向量.具有最高优先级的实例被选作实例原型$\boldsymbol{t}_i$.

(1)为了使实例$\boldsymbol{x}_{ij}$的密度$\rho_{ij}\in\boldsymbol{\rho}_i$能反映邻域范围内的实例聚集程度,$\rho_{ij}$的计算如式(2)所示:

$$\rho_{ij}=\sum_{k\neq j}^{n_i}\boldsymbol{e}^{\left(d_{jk}/d_c\right)^2} \tag{2}$$

其中,d_{jk}表示任意两个实例之间的距离,$d_c=\mu\times\max\{d_{jk}\}$表示实例$\boldsymbol{x}_{ij}$邻域范围的半径,$\mu\in[0,1]$表示给定阈值.如图2所示,在以$d_c$为半径的邻域范围内分析每个实例的密度值:实例3周围聚集了更多的实例,因此具有较大的密度值;处于边缘的实例6周围存在较少的实例,因此密度值较小.这说明邻域范围内的实例聚集程度越高,密度值越大.

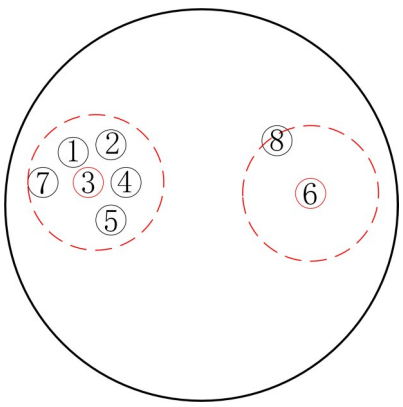

图2　实例密度图

Fig. 2　Diagram of instance density

(2)MIL使用多个实例来表示对象,实例之间则存在一种特殊的内在结构.本文使用包中实例的关联性来分析包的内部结构特征.实例$\boldsymbol{x}_{ij}$与实例$\boldsymbol{x}_{ik}$之间的关联性如式(3)所示:

$$s_{ijk}=\begin{cases}1, & d_{jk}\leqslant d^{\text{ave}}\\ 0, & \text{else}\end{cases} \tag{3}$$

其中,d^{ave}是包内实例的平均距离.实例$\boldsymbol{x}_{ij}$的关联值s_{ij}则用式(4)求出,表示与该实例相连边的数量.如图3所示,与实例4具有关联性的实例有1,5,6,7,则其关联值为4;实例2的关联值则为2.关联值越高,则代表该实例与其他实例之间的关联性越强,就更能代表整个包的重要信息.

$$s_{ij}=\sum_{1\leqslant k\leqslant n_i}s_{ijk} \tag{4}$$

最后,利用式(1)得到$\boldsymbol{X}_i$中每个实例的优先级,选取优先级最高的实例作为实例原型$\boldsymbol{t}_i$,则$T=\{\boldsymbol{t}_1,\cdots,\boldsymbol{t}_i,\cdots,\boldsymbol{t}_n\}$表示$\boldsymbol{\mathcal{T}}$中所有实例原型的集合.

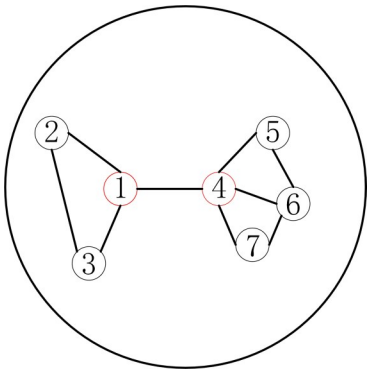

图3　实例关联图

Fig. 3　Diagram of instance affinity

2.2.2　代表实例选择　第二阶段实例选择技术从T中选择一组峰值密度较大的实例原型作为代表实例,并构建代表实例池$\boldsymbol{R}$.

对于任意的$\boldsymbol{t}_i$,需要计算其两个指标:$\boldsymbol{t}_i$的密度值δ_i以及$\boldsymbol{t}_i$与更高密度点的最近距离β_i.$\boldsymbol{t}_i$的密度δ_i通过式(2)计算获得,区别在于计算区间从包内迁移至T中.β_i则是通过计算$\boldsymbol{t}_i$与其他密度更高点之间的最小距离来测量:

$$\beta_i=\min_{j:\delta_j>\delta_i}\{d_{ij}\} \tag{5}$$

特别地,密度最大的实例原型的$\beta_i=\max_j\{d_{ij}\}$.

基于实例原型的δ值和β值,可以将其分为三种不同的类型,如图4所示:(1)峰值点:两个红色实例分布在图的右上角,具有非常高的δ值和β值,表明在两个点的较大区域内不存在密度更高

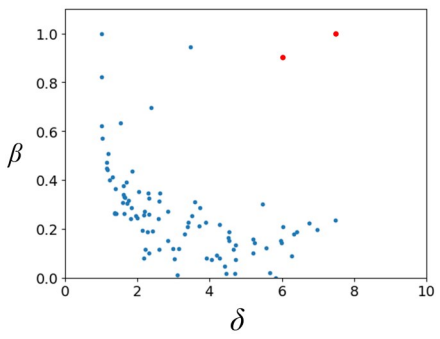

图 4 决策图

Fig. 4 Decision diagram

的实例原型,因此这两个点适合被选取作为代表实例;(2)正常点:处于右下角的实例的 δ 值虽然高,但是 β 值却很小,表明它们附近存在密度更高的点;(3)离群点:分布在左上角的实例点的 δ 值非常低,表明它们是离群点.

因此,本文从 T 中选取前 n_r 个 λ_i 值最大的实例原型,构成代表实例池 $\boldsymbol{R}=\{\boldsymbol{r}_1,\cdots,\boldsymbol{r}_i,\cdots,\boldsymbol{r}_{n_r}\}$,其中,$\lambda_i=\delta_i\times\beta_i$.

2.3 自适应包映射技术 自适应包映射技术根据包中实例与代表实例间的最佳相似度进行映射,既突出了包的内部结构特征,也保证了包在新特征空间中的可区分性.

对于给定的包 $\boldsymbol{X}_i\in\mathcal{T}$,其映射过程如图 5 所示:(1)计算包中实例 $\boldsymbol{x}_{ij}$ 与代表实例池 $\boldsymbol{R}$ 的距离向量 $\boldsymbol{d}_j=\left[d_{j1},\cdots,d_{jm},\cdots,d_{jn_r}\right]$,其中,$d_{jm}\in\boldsymbol{d}_j$ 表示 $\boldsymbol{x}_{ij}$ 与 $\boldsymbol{r}_m\in\boldsymbol{R}$ 的欧式距离;(2)计算 $\boldsymbol{x}_{ij}$ 的自适应目标索引 $\tau_j=\underset{1\leqslant m\leqslant n_r}{\operatorname{argmin}}\{d_{jm}\}$;(3)计算 $\boldsymbol{x}_{ij}$ 的自适应映射向量 $\boldsymbol{v}_{ij}$:

$$\boldsymbol{v}_{ij}=\boldsymbol{x}_{ij}-\boldsymbol{r}_{\tau_j} \tag{6}$$

(4)将所有 $\boldsymbol{v}_{ij}$ 进行叠加得到 $\boldsymbol{X}_i$ 的映射向量 $\boldsymbol{V}_i$:

$$\boldsymbol{V}_i=\sum_{j=1}^{n_i}\boldsymbol{v}_{ij} \tag{7}$$

进一步,对于 $\boldsymbol{V}_i$ 的每一个元素 V_{il} 通过 $V_{il}\leftarrow\operatorname{sign}(V_{il})\sqrt{|V_{il}|}$ 进行处理,再由 $\boldsymbol{V}_i\leftarrow\boldsymbol{V}_i/\|\boldsymbol{V}_i\|_2$ 对映射向量进行二范归一化[22].最后利用 $(\boldsymbol{V}_i,y_i)$ 训练单示例分类器 $\mathcal{F}(\cdot)$.

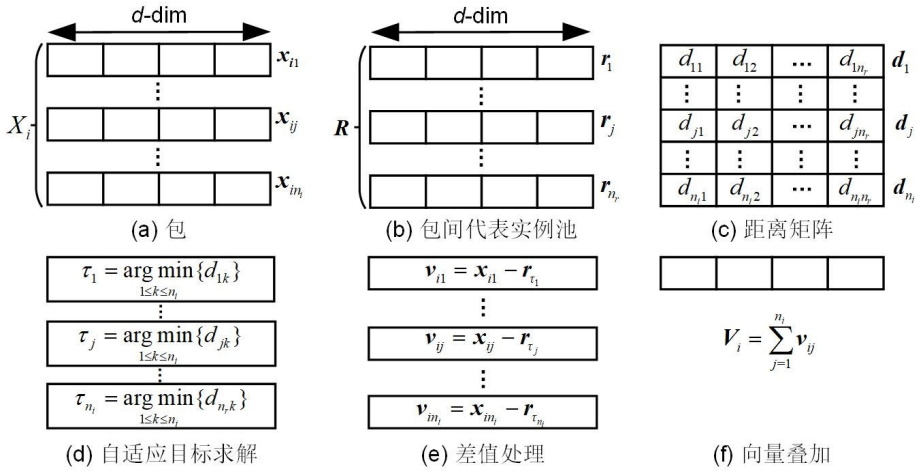

图 5 自适应包映射技术示意图

Fig. 5 Schematic diagram of adaptive bag mapping technique

2.4 算法描述 TAMI 算法分为两个阶段,如 Algorithm 1 所示:算法输入是训练集 B_{tr}(其中 n_{tr} 表示 B_{tr} 中包的个数)和代表实例的数量 n_r;输出是训练好的单示例分类器 $\mathcal{F}(\cdot)$ 和学习到的代表实例池 $\boldsymbol{R}$. 在训练阶段,第 1~6 行为训练集 B_{tr} 中的每一个包 $\boldsymbol{X}_i$ 选择出实例原型,获得 T. 第 7~12 行通过计算 T 中实例的优先级,选出一组具有峰值密度的实例构成代表实例池 $\boldsymbol{R}$. 第 13~18 行使用 $\boldsymbol{R}$ 训练单示例分类器 $\mathcal{F}(\cdot)$. 在测试阶段,第 19~20 行将测试集 B_{te} 中的每一个包 $\boldsymbol{X}_i'$ 通过自适应包映射技术转换成新特征向量 $\boldsymbol{V}_i'$,然后预测出包的标签 $\mathcal{F}(\boldsymbol{V}_i')$.

Algorithm 1　The TAMI algorithm

Input: Training data $B_{tr}=\left\{\left(\boldsymbol{X}_i,y_i\right)\right\}_{i=1}^{n_{tr}}$; the number of representative instances n_r;

Output: The a single-instance classifier $\mathcal{F}(\cdot)$; the representative instance pool $\boldsymbol{R}$;

Train:

//Step 1. Find instance prototypes $\boldsymbol{T}$ based on the density and affinity of instances in the bag.

1. $\boldsymbol{T}=\varnothing$;

2. for $\left(i\in\left[1..n_{tr}\right]\right)$ do

3. Compute $\boldsymbol{p}_i=\left\{p_{i1},\cdots,p_{ij},\cdots,p_{in_i}\right\}$ according to Eq. (1);

4. $m=\underset{1\leqslant j\leqslant n_i}{\operatorname{argmax}}\left\{p_{ij}\right\}$;

5. $\boldsymbol{T}\leftarrow\boldsymbol{T}\cup\left\{\boldsymbol{t}_i\right\}$, where $\boldsymbol{t}_i=\boldsymbol{x}_{im}$;

6. end for

//Step 2. Find representative instance pool $\boldsymbol{R}$ based on the peak density of $\boldsymbol{T}$.

7. $\boldsymbol{R}=\varnothing$;

8. $\boldsymbol{\lambda}=\boldsymbol{\delta}\times\boldsymbol{\beta}=\left\{\lambda_1,\cdots,\lambda_i,\cdots,\lambda_{n_{tr}}\right\}$, where $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are computed according to Eqs. (2) and (5);

9. for $\left(j\in\left[1..n_r\right]\right)$ do

10. $m=\underset{i\in\left[1..n_{tr}\right]\backslash m}{\operatorname{argmax}}\left\{\lambda_i\right\}$, where $\lambda_i\in\boldsymbol{\lambda}$;

11. $\boldsymbol{R}\leftarrow\boldsymbol{R}\cup\left\{\boldsymbol{r}_j\right\}$, where $\boldsymbol{r}_j=\boldsymbol{t}_m$;

12. end for

//Step 3. Map each bag into a single vector.

13. for $\left(i\in\left[1..n_{tr}\right]\right)$ do

14. Compute $\boldsymbol{V}_i$ according to Eq. (7);

15. $V_{il}\leftarrow sign\left(V_{il}\right)\sqrt{\left|V_{il}\right|}$, where V_{il} represents the l-th attribute of $\boldsymbol{V}_i$;

16. $\boldsymbol{V}_i\leftarrow\boldsymbol{V}_i/\left\|\boldsymbol{V}_i\right\|_2$;

17. end for

18. Train a single-instance classifier $\mathcal{F}(\cdot)$ with vector-label pairs $\left(\boldsymbol{V}_i,y_i\right)$;

Test:

19. For each test bag $\boldsymbol{X}_i'\in B_{te}$, perform (13)~(17) steps to get the new vector $\boldsymbol{V}_i'$;

20. Output the prediction $\mathcal{F}\left(\boldsymbol{V}_i'\right)$.

3　实验与结果

为了验证TAMI算法的分类性能,选取三个领域的MIL数据集进行实验,包括图像检索、医学图像以及文本分类,并与Simple-MI,Bamic,miVLAD,miFV和MILDM四类MIL算法进行实验对比.表1给出了以上数据集的关键信息,采用支持向量机(Support Vector Machine,SVM)作为单示例级别分类器.

表1　实验使用的数据集信息

Table 1　Detailed characteristics of datasets used in experiments

Name	Bag	Instance	Attribute	Class
Elephant	200	1391	230	2
Fox	200	1320	230	2
Tiger	200	1220	230	2
Messidor	1200	12352	687	2
Ucsb_breast	58	2002	708	2
Newsgroups	2000	80137	200	20

实验的性能度量建立在混淆矩阵的基础上.对于二分类问题,可将样本分为真正例(TP)、假正例(FP)、真反例(TN)和假反例(FN)四种情形.最终选取$Accuracy$作为性能度量指标,如式(8)所示:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}\tag{8}$$

3.1　数据集简介　Elephant,Fox和Tiger[7]是图像检索中常用的数据集,其任务是确定图像中是否包含感兴趣的对象.每类数据集包含200张图像,每个图像看作一个包,图像的不同区域则视为包中的实例.如果包中包含至少一个该类别的实例,则视为正包,反之为负包.

Messidor[23]和Ucsb_breast[24]属于医学图像数据集.Messidor数据集来自654位糖尿病患者和546位健康患者的1200个眼底图像.原始数据的每个图像都重新缩放为700×700像素,并分成135×135像素的块.Ucsb_breast数据集由58位TMA图像摘录(896×768像素)组成,摘录自32位良性和26位恶性乳腺癌患者.

Newsgroups数据集属于文本分类问题[8],包含20个来自不同新闻帖子的子数据集,每个子数

据集都包含50个正包和50个负包. 每个正包中包含目标类别中3%的帖子以及从其他类别中随机抽取的97%的帖子. 每个负包则只包含从其他类别中随机均匀抽取的帖子,不含目标类别的任何数据.

3.2 实验结果 实验通过10次10折交叉验证来比较TAMI与其他算法的性能. 表2给出了在图像检索和医学图像任务上的实验结果,表3给出了在文本分类任务上的实验结果,表中加粗的数字表示在当前数据集上得到的最好结果. TAMI在这些数据集上都取得了不错的分类性能,尤其在三个图像检索和一个医学图像数据集上显著优于其他算法;文本数据集的分类结果显示TAMI在八个子数据集上取得最优,且在余下子数据集上的分类性能与其他MIL算法相近.

表2 图像检索和医学图像数据集上不同算法的平均准确率

Table 2 The average accuracy of different algorithms on image retrieval and medical image datasets

Datasets	Simple-MI	Bamic	miFV	miVLAD	MILDM	TAMI
Elephant	$82.4_{\pm 0.90}$	$82.9_{\pm 1.57}$	$84.3_{\pm 1.08}$	$84.7_{\pm 1.18}$	$84.2_{\pm 1.25}$	$89.6_{\pm 1.30}$
Fox	$54.0_{\pm 1.50}$	$59.8_{\pm 2.08}$	$60.4_{\pm 1.20}$	$63.3_{\pm 2.21}$	$62.9_{\pm 2.30}$	$64.1_{\pm 2.10}$
Tiger	$80.5_{\pm 0.88}$	$80.6_{\pm 2.23}$	$76.5_{\pm 1.21}$	$84.9_{\pm 0.52}$	$80.1_{\pm 2.09}$	$86.0_{\pm 1.04}$
Messidor	$61.5_{\pm 0.48}$	$62.7_{\pm 0.69}$	$69.4_{\pm 0.55}$	$67.9_{\pm 0.24}$	$57.97_{\pm 1.71}$	$80.7_{\pm 0.79}$
Ucsb_breast	$76.1_{\pm 1.79}$	$76.2_{\pm 3.16}$	$\mathbf{87.0_{\pm 2.57}}$	$81.2_{\pm 2.56}$	$74.80_{\pm 4.24}$	$83.6_{\pm 2.65}$

表3 文本分类数据集上不同算法的平均准确率

Table 3 The average accuracy of different algorithms on text classification datasets

Datasets	Simple-MI	Bamic	miFV	miVLAD	MILDM	TAMI
alt. atheism	$59.6_{\pm 0.80}$	$67.3_{\pm 1.27}$	$82.4_{\pm 0.80}$	$85.1_{\pm 2.17}$	$55.5_{\pm 3.03}$	$\mathbf{88.2_{\pm 0.87}}$
comp. graphics	$52.1_{\pm 1.30}$	$80.6_{\pm 0.66}$	$80.4_{\pm 1.02}$	$79.2_{\pm 1.66}$	$51.1_{\pm 1.85}$	$\mathbf{82.9_{\pm 0.94}}$
comp. os. ms	$50.2_{\pm 1.25}$	$59.5_{\pm 1.02}$	$\mathbf{73.5_{\pm 1.69}}$	$68.6_{\pm 2.20}$	$48.3_{\pm 3.95}$	$71.4_{\pm 1.28}$
comp. sys. ibm	$55.3_{\pm 1.19}$	$74.6_{\pm 0.66}$	$78.8_{\pm 1.89}$	$\mathbf{80.7_{\pm 1.62}}$	$50.4_{\pm 1.78}$	$79.6_{\pm 1.02}$
comp. sys. mac	$52.1_{\pm 2.02}$	$75.6_{\pm 0.66}$	$78.1_{\pm 1.22}$	$78.2_{\pm 2.56}$	$44.3_{\pm 2.58}$	$\mathbf{79.8_{\pm 0.87}}$
comp. window. x	$61.2_{\pm 1.08}$	$74.1_{\pm 0.83}$	$\mathbf{84.8_{\pm 1.66}}$	$81.4_{\pm 1.50}$	$55.0_{\pm 2.94}$	$84.3_{\pm 1.49}$
misc. forsale	$54.7_{\pm 2.49}$	$63.5_{\pm 1.75}$	$\mathbf{73.7_{\pm 1.85}}$	$72.1_{\pm 2.59}$	$47.0_{\pm 4.64}$	$68.9_{\pm 3.18}$
rec. autos	$54.4_{\pm 0.92}$	$72.7_{\pm 0.64}$	$78.8_{\pm 1.17}$	$\mathbf{81.3_{\pm 2.57}}$	$46.4_{\pm 4.74}$	$78.2_{\pm 1.66}$
rec. motorcycles	$54.4_{\pm 1.11}$	$52.0_{\pm 3.79}$	$\mathbf{86.6_{\pm 1.28}}$	$81.4_{\pm 1.20}$	$57.3_{\pm 4.27}$	$83.4_{\pm 1.69}$
rec. sport. baseball	$56.5_{\pm 1.02}$	$78.1_{\pm 0.54}$	$\mathbf{84.7_{\pm 1.19}}$	$82.8_{\pm 1.33}$	$51.1_{\pm 1.66}$	$80.0_{\pm 1.00}$
rec. sport. hockey	$61.2_{\pm 0.75}$	$82.4_{\pm 0.80}$	$87.4_{\pm 1.50}$	$89.8_{\pm 1.40}$	$45.7_{\pm 3.06}$	$\mathbf{90.2_{\pm 1.72}}$
sci. crypt	$60.7_{\pm 0.78}$	$68.0_{\pm 1.73}$	$76.1_{\pm 1.37}$	$82.2_{\pm 2.32}$	$51.2_{\pm 5.45}$	$\mathbf{83.2_{\pm 2.18}}$
sci. electronics	$53.0_{\pm 0.00}$	$92.0_{\pm 0.00}$	$92.7_{\pm 0.78}$	$92.3_{\pm 0.78}$	$52.6_{\pm 1.26}$	$\mathbf{93.9_{\pm 0.70}}$
sci. med	$59.9_{\pm 1.22}$	$79.6_{\pm 0.80}$	$\mathbf{84.3_{\pm 1.42}}$	$82.6_{\pm 2.54}$	$50.5_{\pm 2.68}$	$83.4_{\pm 1.50}$
sci. religion	$60.6_{\pm 1.56}$	$75.8_{\pm 0.98}$	$\mathbf{80.5_{\pm 1.12}}$	$79.6_{\pm 1.36}$	$52.7_{\pm 4.14}$	$79.2_{\pm 2.36}$
sci. space	$52.8_{\pm 0.40}$	$77.4_{\pm 0.49}$	$\mathbf{87.3_{\pm 1.00}}$	$85.0_{\pm 1.10}$	$49.3_{\pm 3.71}$	$86.9_{\pm 2.07}$
talk. politics. guns	$52.8_{\pm 0.40}$	$75.3_{\pm 1.73}$	$77.9_{\pm 1.14}$	$\mathbf{81.4_{\pm 1.28}}$	$48.1_{\pm 2.33}$	$78.3_{\pm 2.15}$
talk. politics. mideast	$64.7_{\pm 0.90}$	$74.3_{\pm 1.10}$	$79.1_{\pm 1.22}$	$\mathbf{83.9_{\pm 1.04}}$	$46.0_{\pm 3.27}$	$82.5_{\pm 0.67}$
talk. politics. misc	$65.5_{\pm 2.54}$	$58.4_{\pm 2.73}$	$73.7_{\pm 1.79}$	$76.5_{\pm 2.29}$	$57.4_{\pm 3.34}$	$\mathbf{82.1_{\pm 1.76}}$
talk. religion. misc	$59.5_{\pm 0.67}$	$68.0_{\pm 0.77}$	$75.3_{\pm 1.73}$	$\mathbf{79.6_{\pm 2.20}}$	$51.5_{\pm 2.84}$	$72.8_{\pm 3.03}$

用统计显著性检验[25]进一步分析TAMI算法的性能. 采用Friedman检验比较各个算法准确率的平均排名. 表4显示TAMI算法与其他五个算法的等级排名,等级越高取值越低. 根据平均

序值能够得到六个算法的分类性能排名,从高到低依次为TAMI,miVLAD,miFV,Bamic,Simple-MI,MILDM. 图6直观地展现了各个算法之间的性能差异,实心点表示各算法的平均序值,以实心点为中心的水平线段表示临界值范围的大小. 结果显示,TAMI算法与MILDM和Simple-MI的水平线段没有交叠区域,因此TAMI算法显著优于MILDM和Simple-MI;与Bamic有很少的交叠区域,因此TAMI优于算法Bamic;与miVLAD和miFV有部分交叠区域,因此它们存在一定差异.

表4　六种算法在三类数据集上的平均排名

Table 4　The mean ranking of the six algorithms in the three types of data sets

Datasets	Simple-MI	Bamic	miFV	miVLAD	MILDM	TAMI
图像检索	5.3	4.3	4.3	2	4	1
医学图像	5	4	1.5	3	6	1.5
文本分类	4.95	4.05	2.1	2.15	5.9	1.85
平均等级	5.1	4.13	2.64	2.38	5.3	1.45

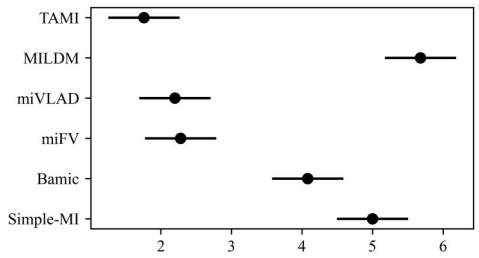

图6　六种算法的Friedman检验图

Fig. 6　Friedman test chart of six algorithms

4　总结与展望

本文提出一种多示例学习的两阶段实例选择和自适应包映射(TAMI)算法. 该算法首先计算包内实例的密度值和关联性,选出具有包内部结构信息的实例原型;然后基于实例原型的峰值密度特点,选择一定数目的代表实例;最后根据包与代表实例的相似关系来构建自适应背景,从而实现自适应包映射. 该算法的优势在于:(1)代表实例的选取无需迭代优化,计算速度快;(2)通过考虑包中实例的密度值和相关性,可以有效保留包的内部结构特征;(3)代表实例的数目可以自由选择,使用灵活. 实验结果表明,TAMI在图像检索和医学图像数据集上取得了比其他MIL算法更好的效果,并在文本分类数据集上的表现良好.

通过实验也发现:(1)TAMI算法没有在所有数据集上都表现最优,例如在文本数据集上取得了和miFV相近的结果;(2)实例的关联性计算仅考虑了实例之间的距离关系. 这些都是本文进一步研究的方向,将在未来的工作中尝试解决.

参考文献

[1] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1−2):31−71.

[2] Maron O, Ratan A L. Multiple-instance learning for natural scene classification//Proceedings of the 15th International Conference on Machine Learning. San Francisco, CA, USA: IEEE, 1998:341−349.

[3] Song X F, Jiao L C, Yang S Y, et al. Sparse coding and classifier ensemble based multi-instance learning for image categorization. Signal Processing, 2013, 93(1):1−11.

[4] Wei X S, Ye H J, Mu X, et al. Multi-instance learning with emerging novel class. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(5):2109−2120.

[5] Zhu L, Zhao B, Gao Y. Multi-class multi-instance learning for lung cancer image classification based on bag feature selection//2008 5th International Conference on Fuzzy Systems and Knowledge Discovery. Ji′nan, China: IEEE, 2008:487−492.

[6] Wang Z Y, Poon J, Sun S D, et al. Attention-based multi-instance neural network for medical diagnosis from incomplete and low quality data//2019 International Joint Conference on Neural Networks. Budapest, Hungary: IEEE, 2019:1−8.

[7] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning// Proceedings of the 15th International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2002:561−568.

[8] Zhou Z H, Sun Y Y, Li Y F. Multi-instance learning by treating instances as non-I.I.D. samples//Proceedings of the 26th Annual International Conference on

Machine Learning. New York, NY, USA: ACM, 2009:1249－1256.

[9] Angelidis S, Lapata M. Multiple instance learning networks for fine - grained sentiment analysis. Transactions of the Association for Computational Linguistics, 2018(6):17－31.

[10] Zhang D, He J R, Lawrence R. MI2LS: Multi - instance learning from multiple informationsources// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2013: 149－157.

[11] Zhou Z H, Zhang M L. Solving multi - instance problems with classifier ensemble based on constructive clustering. Knowledge and Information Systems, 2007, 11(2):155－170.

[12] Chen Y X, Bi J B, Wang J Z. MILES: Multiple - instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):1931－1947.

[13] Hong C, Wang M, Gao Y, et al. Image annotation by multiple-instance learning with discriminative feature mapping and selection. IEEE Transactions on Cybernetics, 2014, 44(5):669－680.

[14] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344(6191): 1492－1496.

[15] Amores J. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence, 2013(201):81－105.

[16] Zhang M l, Zhou Z H. Multi-instance clustering with applications to multi - instance prediction. Applied Intelligence, 2009, 31(1):47－68.

[17] Wei X S, Wu J X, Zhou Z H. Scalable multi-instance learning//2014 IEEE International Conference on Data Mining. Shenzhen, China: IEEE, 2014: 1037－1042.

[18] Xu B C, Ting K M, Zhou Z H. Isolation set-kernel and its application to multi - instance learning// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: ACM, 2019: 941－949.

[19] Wei X S, Wu X J, Zhou Z H. Scalable algorithms for multi - instance learning. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(4): 975－987.

[20] Wu J, Pan S R, Zhu X Q, et al. Multi - instance learning with discriminative bag mapping. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6):1065－1080.

[21] Zhang Y L, Zhou Z H. Multi-instance learning with key instance shift//Proceedings of the 26th International Joint Conference on Artificial Intelligence Main Track. Melbourne, Australia: IJCAI, 2017:3441－3447.

[22] Sánchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision, 2013, 105(3):222－245.

[23] Decencière E, Zhang X W, Cazuguel G, et al. Feedback on a publicly distributed image database: The messidor database. Image Analysis & Stereology, 2014, 33(3):231.

[24] Kandemir M, Hamprecht F A. Computer - aided diagnosis from weak supervision: A benchmarking study. Computerized Medical Imaging and Graphics, 2015(42):44－50.

[25] Demšar J. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 2006(7):1－30.

(责任编辑 杨可盛)