一种基于聚类的图卷积多示例学习算法

王江晴,毕建权,帖军,孙翀,艾勇
(中南民族大学 计算机科学学院 & 湖北省制造企业智能管理工程技术研究中心,武汉 430074)

摘　要　基于图结构的多示例学习可用于解决挖掘包中示例间相关性问题.然而,现有的方法大多随机选择包中示例构建图结构,忽略了包中具有代表性示例对图结构的影响;同时都是间接在包图结构上建立分类器,造成了模型运行效率低下的问题.针对上述问题,提出了一种基于聚类的图卷积多示例学习算法 MIL-GCC,首先通过聚类方法获取每个包中的超示例,作为包图结构中的节点;然后通过挖掘超示例间关系构建包图的边,确定包图结构;最后利用图卷积对包图节点重要度分数进行学习,直接在包图结构上建立分类器.实验表明:MIL-GCC 可以充分表示包图结构,有效提高分类模型的质量.

关键词　多示例学习;聚类;包图;图卷积
中图分类号　TP391　**文献标志码**　A　**文章编号**　1672-4321(2021)01-0089-06
doi:10.12130/znmdzk.20210114
引用格式　王江晴,毕建权,帖军,等.一种基于聚类的图卷积多示例学习算法[J].中南民族大学学报(自然科学版),2021,40(1):89-94.
WANG Jiangqing,BI Jianquan,TIE Jun,et al.A multi-instance learning algorithm of graph convolution based on clustering [J].Journal of South-Central University for Nationalities(Natural Science Edition),2021,40(1):89-94.

A multi-instance learning algorithm of graph convolution based on clustering

WANG Jiangqing,BI Jianquan,TIE Jun,SUN Chong,AI Yong
(College of Computer Science & Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises,South-Central University for Nationalities,Wuhan 430074,China)

Abstract　Multi-instance learning based on graph structure can be used to solve the problem of correlation between instances in mining bag.However,most of the existing methods randomly select the instances in the bag to construct the graph structure,ignoring the influence of the representative instances in the bag on the graph structure,and all of them build classifiers on the bag graph structure indirectly,resulting in the low efficiency of the model.Aiming at these problems,a multi-instance learning algorithm of graph convolution based on clustering is proposed(MIL-GCC). Firstly,the super-instances in each bag are obtained by clustering method as the nodes in the bag graph structure.Secondly,the edge of the bag graph is constructed by mining the relationship between super-instances to determine the structure of the bag graph. Finally,graph convolution is used to learn the node importance score of the bag graph,and the classifier is directly established on the bag graph structure.Experiments show that MIL-GCC can fully represent the bag graph structure and effectively improve the quality of the classification model.

Keywords　multi-instance learning;clustering;bag graph;graph convolution

收稿日期　2020-07-30
作者简介　王江晴(1964-),女,教授,博士,研究方向:智能算法,E-mail:wjqing2000@ mail.scuec.edu.cn
基金项目　湖北省技术创新专项重大项目(2019ABA101);武汉市科技计划应用基础前沿项目(2020020601012267);中央高校基本科研业务费专项资金资助项目(CZY18015)

多示例学习的概念源自医学领域制药过程中何种分子适合制药问题[1],一个对象被定义为一个包,每个包由多个示例组成并由一个标记与之对应,学习的目的是建立一个学习器,对未知标记的包进行分类[2].目前,多示例学习算法被成功地应用到图像分类与标注[3]、自然语言处理[4]、股票趋势预测[5]等领域.

近年来,多示例学习的研究可以分为挖掘示例间关系、挖掘包与标记间关系和关键示例检测[6,7]三种.MIGraph、miGraph 方法[8]是典型的基于图结构挖掘示例间关系的多示例学习方法,核心思想是将包视为图,包中的每个示例视为图中的节点,分别利用 ε-graph 和亲和度矩阵构建包图结构,设计一个图核函数来捕获包图间各节点的相似性,用以作为构建分类器的依据.由于随机选择包中的示例构建图结构,因此在模型分类准确率上还有较大的提升空间.SCPMK_MIL 方法[9]利用谱聚类方法获取潜在的正示例代表,利用径向基函数和金字塔核分别挖掘正示例间和负示例间的相似性,但未充分考虑包间的相似性问题.MI-SVM 方法[10]和 DD-SVM 方法[11]使用有监督学习的支持向量机 SVM 挖掘包与标记间关系,虽然这些方法具有较好的泛化能力和小样本学习能力,但是求解的目标函数很难直接计算,导致训练效率较低.MGML-ELM 方法[12]、BELM-MIL 方法[13]和 RBF-MIP 方法[14]在此基础上利用神经网络提高模型的训练效率,但是降低了模型的可解释性,同时在基于图结构的方法中存在子图转为特征向量造成模型运行效率低下等问题.

本文针对随机选择包中的示例构建图结构和模型运行效率低下的问题,提出了一种基于聚类的图卷积多示例学习算法 MIL-GCC(Multi-Instance Learning Algorithm of Graph Convolution Base on Clustering).MIL-GCC 主要分为 2 个步骤:(1)构建包的图结构,利用 k-means 聚类方法获取包中具有代表性的示例作为包图中的节点,然后挖掘超示例间关系构建包图的边;(2)构建图分类器,将构建好的包图结构转为邻接矩阵的形式作为图卷积层的输入,利用图卷积对节点重要度分数进行学习,筛选重要度分数排序靠前的节点以及这些节点组成的包图结构作为模型分类的依据.

1 问题描述与形式化定义

MIL-GCC 的目标是从训练集中学习一组包图结构与标记间的关系,用于对未知标记包分类.

令 $X=\mathbb{R}^d$ 表示示例空间.定义 $D=\{(X_1,y_1),\cdots,(X_i,y_i),\cdots,(X_m,y_m)\}$ 表示具有 m 个包的 MIL 数据集,其中 $X_i=\{x_{i1},\cdots,x_{ij},\cdots,x_{ip_i}\}\subseteq X$ 被称为一个包,$y_i\in Y=\{0,1\}$ 是 X_i 所属的标记;$x_{ij}\in X$ 是一个由 d 维特征向量表示的示例,即 $x_{ij}=[x_{ij1},\cdots,x_{ijl},\cdots,x_{ijd}]'$;$p_i$ 表示 X_i 中示例个数的总数.如果存在 $index\in\{1,\cdots,j,\cdots,p_i\}$,使得 $x_{i,index}$ 是一个正示例,则 X_i 是正包且 $y_i=1$;否则 X_i 是负包且 $y_i=0$.该模型的目标是从数据集 D 中学习一个多示例分类器 $f:2^X\to Y$.

2 构建包图结构

MIL-GCC 算法的第 1 步是构建包图结构.包图结构的构建可以分为两部分,一部分是选取每个包中的超示例作为包图的节点,另一部分是根据超示例间关系在定义好的约束条件下创建包图的边.本节首先对超示例、包图的概念进行了定义,然后对构建包图的具体实现进行详细的描述.

定义 1 超示例.超示例是指对包中所有示例进行聚类后所得到的簇中心点,即:对数据集 D 中某一个样本 (X_i,y_i) 而言,对 X_i 中所有的示例 $\{x_{i1},\cdots,x_{ij},\cdots,x_{ip_i}\}$ 聚类得到的簇中心点 $x'_{i1},\cdots,x'_{is},\cdots,x'_{ik_i}$,此处称 x'_{is} 为 X_i 中的一个超示例,x'_{is} 所在的集合 $\{x'_{i1},\cdots,x'_{is},\cdots,x'_{ik_i}\}$ 称为 X_i 的超示例的集合.

定义 2 包图.对数据集 D 而言,包图是指每个包 X_i 的图结构信息,由节点集 $V_i=\{v_{i1},\cdots,v_{ij},\cdots,v_{in_i}\}$、边集 $E_i=\{(v_{ia},v_{ib}),\cdots,(v_{ic},v_{id})\}$ 组成,其中,$a,b,c,d\in\{1,2,\cdots,n_i\}$ 且 $a\neq b,c\neq d$,$|E_i|=e_i$ 表示包图中边个数的总数.包图简记为:$g_i,i\in\{1,2,\cdots,m\}$.

本文采用基于超示例的方法构造包图.首先,通过 k-means 聚类方法对包图中的示例进行聚类,获取包图中超示例的集合.由于 k-means 聚类方法的作用域是全局的,而包图中示例间的邻域关系具有较强的关联度,因此,本文利用局部 k-means 聚类方法(算法 1 中 2-15 行)对 X_i 中所有的 x_{ij} 聚类得到簇中心集合 $\{x'_{i1},\cdots,x'_{is},\cdots,x'_{ik_i}\}$,根据簇中心集合确定包图 g_i 的顶点集合 V_i 如公式(1):

$$v_{i1}=x'_{i1},\cdots,v_{ij}=x'_{is},\cdots,v_{in_i}=x'_{ik_i},\qquad(1)$$

其中,算法 1 中 $C(2S\times 2S)$ 表示每个包聚类的搜索范围,$S=\mathrm{sqrt}(p_i/k)$ 表示相邻聚类中心的步长,p_i 代表第 i 个包中示例总数,k 为第 i 个包中簇中心总数;

$l(ij)$ 表示第 i 个包中第 j 个示例所属的簇类别,初始值为 -1;$d(ij)$ 表示第 i 个包中第 j 个示例到任意簇中心的距离,初始值为 ∞;$dist$ 表示包中任意示例与簇中心之间的欧式距离,计算公式如下:

$$dist = \sqrt{(\tilde{x}_{ij} - \tilde{x}'_{ik})^2 + (\tilde{y}_{ij} - \tilde{y}'_{ik})^2}, \tag{2}$$

其中,$(\tilde{x}_{ij},\tilde{y}_{ij})$、$(\tilde{x}'_{ik},\tilde{y}'_{ik})$ 分别表示第 i 个包中第 j 个示例和第 k 个簇中心的位置.

然后,假设用一个二维坐标系表示每个 x'_{is} 的位置,则可以用 $coordinate_{is}[x][y]$ 表示第 i 包中第 s 个超示例所在的位置值,其中,$x,y \in \mathbb{R}^{n_i}$ 表示任意超示例的位置,例如,$coordinate_{i1}[1][1] = 1$,表示第 i 个包中第 1 个超示例在坐标(1,1) 处的位置值为 1,如果存在任意 2 个超示例 x'_{ia} 和 $x'_{ib}(a \neq b)$ 的 $coordinate_{ia}[x][y]$,$coordinate_{ib}[x][y]$ 满足(3) 式的条件,则在 x'_{ia} 和 x'_{ib} 间建立一条边,通过遍历 X_i 中所有的 x'_{is} 确定 g_i 中的 E_i.

$$\begin{cases} coordinate_{ia}[x][y] - coordinate_{ib}[x][y] = \pm 1, \\ \max\{coordinate_{ia}[x][y], coordinate_{ib}[x][y]\} \leqslant \\ \qquad \max(coordinate_{is}[x][y]). \end{cases} \tag{3}$$

通过算法 1 将数据集 D 中的带标记的包映射成为一组带标记的包图之后,可以采用很多已有的方法挖掘包图间的相关关系并建立分类器,例如建立一个图核函数表示包图之间的相似性,然后利用支持向量机 SVM 解决分类问题,或者通过挖掘所有包图的频繁项信息子图表示其相似性,然后利用极限学习机 ELM 解决分类问题.虽然上述方法可以有较好的分类准确率,但是间接在包图结构上建立分类器会造成模型运行效率低下问题,因此,本文基于直接在包图结构上建立分类器的思想,利用图卷积进行图分类器的构建.

算法 1 基于超示例的包图结构构建算法

输入:数据集 D,聚类簇数 k 和阈值 β
输出:带有标记的一组包图 $g_i, i \in \{1,2,\cdots,m\}$

```
for i in {1,2,…,m} do
    随机选择 k 个超示例簇中心{x'_i1,…,x'_is,…,x'_ik_i}
    初始化 x_ij 所属类别为 l(ij) = -1,距离 d(ij) = ∞
    repeat
        for x'_il in {x'_i1,…,x'_is,…,x'_ik_i} do
            for x_ij in C(2S × 2S) do
                计算 x'_il 与 x_ij 之间的欧式距离 dist
                if dist < d(ij) then
                    d(ij) = dist, l(ij) = k
                end if
            end for
        end for
        重新计算 k 个超示例簇中心
        利用 L2 范数计算残差 E
    until E ≤ β
    for x'_il in {x'_i1,…,x'_is,…,x'_ik_i} do
        for x'_i(l+1) in {x'_i1,…,x'_is,…,x'_ik_i} - {x'_il} do
            根据(3) 式成立条件在 E_i 中建立一条边
            (x'_il, x'_i(l+1))
        end for
    end for
    return g_i
end for
```

3　构建图分类器

本文利用公式(4) 获得每个包图 g_i 的邻接矩阵信息,并作为图卷积层的输入.

$$A_{ab} = \begin{cases} 1, \text{if}(v_{ia}, v_{ib}) \in E_i, \\ 0, \text{others}. \end{cases} \tag{4}$$

其中,$A \in \mathbb{R}^{n_i \times n_i}$ 表示 g_i 的邻接矩阵,A 是对称矩阵 $(A_{ab} = A_{ba})$;v_{ia}, v_{ib} 表示 g_i 中任意 2 个节点$(a \neq b)$.然后,利用经典的图卷积神经网络 GCN[15,16] 中的卷积层提取包图中所有节点的特征,从而得到每个节点的重要度分数 z_{scorej},计算公式如下:

$$z_{score} = \sigma(\tilde{L}_{sym} VW),$$

其中,$z_{score} = \{z_{score1}, \cdots, z_{scorej}, \cdots, z_{scoren_i}\}$;$\sigma$ 表示激活函数;$\tilde{L}_{sym}$ 为重归一化拉普拉斯矩阵,$\tilde{L}_{sym} = \tilde{D}^{1/2}\tilde{A}\tilde{D}^{-1/2}$,$\tilde{A} = A + I$,$\tilde{D}$ 为 $\tilde{A}$ 的度矩阵,$A \in \mathbb{R}^{n \times n}$ 表示包图的邻接矩阵,$N = \sum_{i=1}^{m} n_i$;V 表示包图中节点的特征;$W \in \mathbb{R}^{F \times 1}$ 表示权重参数,F 表示当前图卷积层的输入特征维度.

在计算每个节点重要度分数的基础上,还考虑了包图间的不同尺度问题,即对于任意 2 个包图 g_i 和 g_j,其中 $i \neq j$,存在 $n_i \neq n_j$ 和 $e_i \neq e_j$ 的情况,通过设置一个保留节点比例的超参数 $\delta \in (0,1]$,对每个 z_{scorej} 进行降排序,筛选前 δn_i 个节点进行特征的学习.每次筛选相当于对包图结构和特征进行更新,具体的包图结构更新计算公式如下:

$$A' = A_{mask,mask}, V' = V_{mask,:},$$

其中,A' 和 V' 分别表示保留节点之间的邻接矩阵和节点特征矩阵;$A_{mask,mask}$ 表示根据前 δn_i 个节点的节点索引 $mask$ 对 A 同时进行行切片和列切片;$V_{mask,:}$ 表示按照 $mask$ 对 V 进行列切片.

虽然通过层层丢弃节点的方式可以提高包图中远距离节点的融合效率,但是会降低对所有节点信

息的有效融合,因此,本文采用全局最大池化与全局平均池化[17] 拼接的方式对包图的全局信息进行一次性融合,拼接过程如图 1 所示.

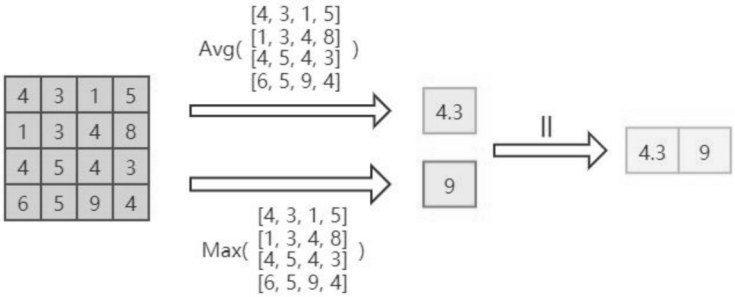

图 1 全局平均池化与全局最大池化拼接过程

Fig.1 Splicing process of global average pooling and global maximum pooling

最后将学习到的包图全局信息用于分类.基于图卷积的图分类器模型相比于已有的基于 SVM、ELM 等分类模型,忽略了包图中节点数和边数对模型的影响,因此,在包图中节点数和边数较大时具有一定的优势.

4 实验与分析

4.1 数据集与实验环境

本文选取 5 个多示例学习基准数据集(Musk1、Musk2、Elephant、Fox、Tiger)和 1 个真实图像数据集(2000-Image)对提出的算法进行评价.多示例学习基准数据集的具体属性信息见表 1.2000-Image 图像数据集汇总包含 20 类 COREL 图像,每个类别由 100 张像素为 64×96 的彩色图像组成,每个图像都视为一个包,图像中的每个段被视为一个示例.

表 1 多示例学习基准数据集具体属性信息

Tab.1 Specific attribute information of multi-instance learning benchmark datasets

数据集名称	维度	正包数/个	负包数/个	正示例数/个	负示例数/个
Musk1	166	47	45	207	269
Musk2	166	39	63	1017	5581
Elephant	230	100	100	762	629
Fox	230	100	100	647	673
Tiger	230	100	100	544	676

在实验中,本文使用 10 倍交叉验证来比较结果.将数据集分为 10 份,轮流将其中 9 份作为训练集,一份作为测试集,进行实验,将 10 次结果的准确率的平均值作为算法的评判指标,具体的计算公式如下:

$$Accuracy = \frac{1}{N}\sum_{q=1}^{N}\frac{s_q}{t_q},$$

其中,$N = 10$,s_q 表示第 q 次结果中所有 X_i 被正确分类的总数,t_q 表示第 q 次结果中样本总数.

实验环境为 16 G 内存的 Windows10 操作系统,其 CPU 为 AMD Ryzen 5 4600U with Radeon Graphics,主频为 2.1 GHz,编程语言为 Python 3.7.6.

4.2 实验结果与分析

本文的实验过程主要分为 4 个部分,第 1 部分是对数据集进行预处理,即确定每个输入数据集中示例规模的一致;第 2 部分是构建包图结构,本文采用局部 k-means 方法获取每个包中的超示例,然后根据边成立的约束条件确定包图结构,并根据训练/测试集所占比随机划分训练/测试集;第 3 部分为了保证实验对比的公平性,实验对基于多示例多标记的 MGML-ELM 方法进行了单标记的条件约束,然后和 MIL-GCC 分别进行分类器的构建;第 4 部分则是通过评判指标对实验结果进行分析与总结.

影响实验结果的参数主要有:示例的聚类数 n,

阈值 β,图卷积隐藏层层数 h,节点保留比 δ.为了确定模型的最优分类准确率,本文实验依据表 2 的实验各参数取值范围分别对模型进行准确率的计算,其中,对于基准数据集,在 $n = 20, \beta = 1, h = 60, \delta = 0.6$ 时,可达到最优分类准确率;对于图像数据集,在 $n = 100, \beta = 1, h = 120, \delta = 0.8$ 时,可达到最佳分类准确率.

表 2　实验各参数取值范围

Tab.2　The range of parameters in the experiment

数据集	示例的聚类数 n	阈值 β	图卷积隐藏层层数 h	节点保留比 δ
基准数据集	{20,40,60,80,100}	(0,1]	{20,40,60,80,100}	{0.4,0.5,0.6,0.7,0.8}
图像数据集	{50,100,150,200}	(0,1]	{80,100,120,140,160}	{0.5,0.6,0.7,0.8,0.9}

实验的对比结果见表 3,通过在基准数据集和图像数据集上与 MIGraph[8]、miGraph[8]、MIKI[6]、MGML-ELM*(MGML-ELM 进行单标记条件约束后的算法简称)[12] 的比较,可以发现 MIL-GCC 在 Musk1、Musk2 数据集上没有 MIGraph、miGraph 和 MIKI 分类准确率高,但是,在图像类数据集上相对其他 3 种方法具有很好的分类准确率,同时从所有数据集平均准确率的角度看,MIL-GCC 具有一定的分类准确率优势.

表 3　基于各数据集下的各算法准确率对比

Tab.3　Comparison of accuracy of each algorithm based on each dataset　　%

数据集	MIGraph	miGraph	MIKI	MGML-ELM*	MIL-GCC
Musk1	84.3	88.9	88.2	78.5	81.3
Musk2	81.3	87.4	84.3	75.6	75.6
Elephant	77.8	82.3	81.4	84.3	85.3
Fox	61.2	61.6	68.5	78.6	82.1
Tiger	81.9	78.4	82.5	87.8	86.2
2000-Image	72.1	70.5	85.6	81.7	89.2
平均值	76.4	78.2	81.8	81.1	83.3

经实验验证,MIL-GCC 在 Musk1、Musk2 数据集上准确率不高的主要原因是,MIL-GCC 在数据集预处理过程选择局部 k-means 算法不利于挖掘 Musk1、Musk2 数据集中超示例相关性,采用 k-means 算法准确率可达 90%,但 MIL-GCC 在图像数据集准确率和总数据集的平均准确率上未占优势.

同时,为了验证 MIL-GCC 可以有效地提高算法的执行效率,分别在基准数据集和图像数据集上与 MIGraph、miGraph、MGML-ELM* 算法进行比较,对比结果如图 2 所示,可以清晰地发现基于图像数据集,MIL-GCC 算法相对于其他 3 种算法需要的运行时间较少,因此,MIL-GCC 算法在处理规模较大的数据集时也具有一定的优势.

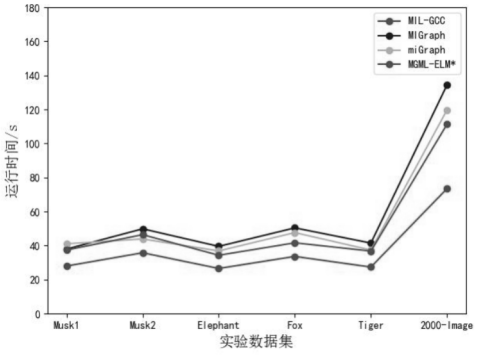

图 2　各算法执行效率对比

Fig.2　Comparison of execution efficiency of each algorithm

5　结语

本文提出的 MIL-GCC 算法通过利用局部 k-means 聚类方法确定构建包图的超示例的集合,然后通过挖掘示例间的相关性确定包图结构,最后基于图卷积思想直接在包图上建立图分类器,忽略了包图中节点数和边数对图分类模型的影响.实验验证了 MIL-GCC 的性能,同时在图像分类领域有明显的成效.然而,MIL-GCC 在分子数据集上还存在提升

的空间,因此,如何更充分地选择构建包图的超示例成为下一步的主要工作.

参 考 文 献

[1] DIETTERICH T G,LATHROP R H,LOZANOPEREZ T, et al. Solving the multiple instance problem with axis-parallel rectangles [J]. Artificial Intelligence, 1997, 89 (1): 31-71.

[2] ZHOU Z H. Multi-instance learning: a survey [R]. Nanjing: Nanjing University,2004.

[3] HONG R, WANG M,GAO Y,et al.Image annotation by multiple-instance learning with discriminative feature mapping and selection [J]. IEEE Transactions on Cybernetics,2014,44(5):669-680.

[4] LIN Y,SHEN S,LIU Z,et al. Neural relation extraction with selective attention over instances[C]//ACL.Meeting of the Association for Computational Linguistics. Berlin: ACL,2016: 2124-2133.

[5] ZHANG X, QU S, HUANG J, et al. Stock market prediction via multi-source multiple instance learning[J]. IEEE Access,2018(6): 50720-50728.

[6] ZHANG Y L,ZHOU Z H.Multi-instance learning with key instance shift [C]//IJCAI. Proceeding of the 26th International Joint Conference on Artificial Intelligence. Melbourne: IJCAI,2017: 3441-3447.

[7] LIU G,WU J,ZHOU Z H.Key instance detection in multi-instance learning [J]. Journal of Machine Learning Research,2012,25: 253-268.

[8] ZHOU Z H,SUN Y Y,LI Y F.Multi-instance learning by treating instances as non-I. I. D samples [C]//ICML. Proceedings of the 26th International Conference on Machine Learning.Montreal: ICML,2009: 1249-1256.

[9] 李展,彭进业,温超.基于谱聚类和多示例学习的图像检索方法[J].华南理工大学学报(自然科学版),2011,39(7):156-162.

[10] ZHOU Z H, ZHANG M L. Multi-instance multi-label learning with application to scene classification[C]// NIPS. Advances in Neural Information Processing Systems 19.Vancouver: NIPS,2006: 1609-1616.

[11] ANDREWS S, TSOCHANTARIDIS I, HOFMANN T. Support vector machines for multiple-instance learning [J]. Advances in Neural Information Processing Systems,2003,15(2): 561-568.

[12] ZHU Z X,ZHAO Y H.Multi-graph multi-label learning based on entropy[J].Entropy,2018,20(4): 245-259.

[13] WANG P P,ZHENG X, KU J, et al. Multiple-instance learning approach via bayesian extreme learning machine [J].IEEE Access,2020(8): 62458-62470.

[14] ZHANG M L,ZHOU Z H.Adapting RBF neural networks to multi-instance learning[J].Neural Processing Letters, 2006,23(1): 1-26.

[15] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs [C]// ICLR. International Conference on Learning Representations.Banff: ICLR,2014: 6203-6217.

[16] LI Q M,HAN Z C,WU X M.Deeper insights into graph convolutional networks for semi-supervised learning [C]//AAAI.AAAI Conference on Artificial Intelligence. New Orleans: AAAI,2018: 3538-3545.

[17] LEE J,LEE I,KANG J,et al.Self-attention graph pooling [C]//ICML.36th International Conference on Machine Learning.Los Angeles: ICML,2019: 3734-3743.

(责任编辑 曹东)