面向多标签图数据的主动学习

李远航 刘 波 唐 侨
(广东工业大学自动化学院 广州 511495)

摘 要 主动学习已经广泛应用于图数据的研究,但应用于多标签图数据的分类较为少见。结合基于误差界最小化的主动学习,给出了一种多标签图数据的分类方法,即通过多标签分类与局部和全局的一致性学习(LLGC)得到一系列目标方程,并将其用于最小化直推式的拉德马赫复杂度,得到最小泛化误差上界,从而在图上获取少量的但蕴含巨大信息量的节点。实验证明,应用该方法的多标签分类器的输出有很高的精确度。

关键词 图数据,主动学习,复杂度,最小化
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.11.050

Active Learning for Multi-label Classification on Graphs

LI Yuan-hang LIU Bo TANG Qiao
(School of Automation, Guangdong University of Technology, Guangzhou 511495, China)

Abstract Although active learning has been extensively used in study in graph data, little research has been done on active learning on multi-label classification with graph data. We proposed a novel approach for multi-label classification with graph data by using an active learning based on error bound minimization. We first obtained a series of equations by using multi-label classification and learning with local and global consistency (LLGC), so as to make the equation apply to minimize the transductive rademacher complexity and minimize the generalization error bound. By using the approach, we obtained the most informative sample data from graph data. Experiments show that our method can obtain high performance for multi-label classification.

Keywords Data on graph, Active learning, Complexity, Minimization

1 引言

在传统的分类方法中,大部分都是面向单标签的分类问题,如C4.5算法、支持向量机算法、K近邻算法等[1,2]。而在实际应用和研究中遇到的分类问题却大多是较为复杂的多标签分类问题[3,4]。多标签分类问题比单标签分类问题更一般化或普遍。例如,一篇有关运动员刘翔的报道,它既可以是体育、娱乐新闻,也可以是人物传记;一部电影既可以划分为动作片,也可以是恐怖片和犯罪剧情片等。针对这些情况,传统的单标签分类技术已经无法解决。

近几年来,多标签数据分类越来越受到研究者的关注,随着相关研究的不断深入,多标签分类的重要性及其应用价值已逐步体现出来,如它在Rubin等人的多标签文本分类中的应用[5]。目前,人们通过与各种学习技术相结合,提出了不同类型的多标签学习方法,以解决不同的实际问题。很多专门针对多标签分类学习的方法也陆续被提了出来,如多标签文本分类方法[6]、基于核函数的多标签分类方法[7]以及基于神经网络的多标签分类方法[8]等。

尽管多标签分类已经得到了广泛的应用,但是对于面向图数据的多标签方法的研究还是比较缺乏。将数据以图的形式表示的数据分类方法比传统的基于向量的分类方法更加符合实际应用,更具有一般性。在现实生活中较为普遍的应用是数据以图的形式而存在的,例如在药物的分类中,药物的信息数据以及对各种病症产生的效果数据都是以图的形式进行存储和分析的[9]。在本文实验中,对有毒物质是否致癌进行预测与分类,亦即将各有毒物质表示成节点,以图的形式存储,并有效地利用物质间的相关性对有毒物质的致癌性进行预测。为了能利用少量有标签数据对大量未标签数据进行预测,以提高分类效率和降低分类成本,本文方法将结合主动学习对面向多标签的图数据进行分类。

本文研究了基于多标签图数据的主动学习,贡献有:(1)使用一种对图数据进行分类的方法:局部和全局的一致性学习(LLGC)[10],并给出这种方法与多标签数据分类的目标方程;(2)为了让多标签分类器获得更好的泛化误差,引入直推式的拉德马赫复杂度[11]和一种顺序优化方法[12],给出与多标签数据分类有关的求解算法,该算法可以得到少量且具有最重要信息的节点,从而得到最优泛化误差;(3)为了检验该方法的有效性,对一些数据集进行了若干实验,对数据分类的

到稿日期:2014-01-28 返修日期:2014-04-13 本文受国家自然科学基金(61070033,61203280,61202270),广东省自然科学杰出青年基金(S2013050014133),广东省自然科学基金(9251009001000005,S2011040004187,S2012040007078)资助。

李远航 男,硕士生,主要研究方向为数据挖掘等;刘 波 男,教授,硕士生导师,主要研究方向为数据挖掘等,E-mail:csbliu@gmail.com(通信作者);唐 侨 男,硕士生,主要研究方向为数据挖掘等。

准确性进行了分析与对比。通过实验发现，该方法能在使用少量标签数据的情况下，获得高的准确率。

本文第2节回顾相关工作内容；第3节提出基于多标签图数据的主动学习；第4节是实验部分；最后是对以后工作的总结。

2 相关工作

我们关注的是多标签图数据的主动学习，因此首先回顾多标签数据分类和主动学习。

2.1 多标签数据分类

构造多标签分类器的难度在于如何权衡各类别之间的关系，以准确地进行多标签分类。Gao[13]提出了为每一个类别训练相互独立的分类器，并通过一个自信度测试标准对分类器进行排序；Crammer 和 Singer[14]描述了一种模型，它为每一个分类获取标准的特征向量，并通过这个标准和文本之间的关系导出分类的等级；由于许多领域的标签之间是高度相互依赖的，因此 Ghamrawi 和 McCallum[15]提出了“有条件的随机领域”(CRF)分类器模型来直接确定标签间相关性参数；二元相关性[8]方法将多标签问题转换成多重的二元分类器问题，如 Boutell 提出的学习多标签的事件分类器[16]。ML-KNN[17]也是其中一种二元相关性的方法。

以上方法分别以两种相反的思路分析了多标签分类的问题：不考虑与考虑不同类别标签间的相关性。虽然考虑标签间的相关性可以进一步提高分类器输出的准确性，但方法的实现难度以及获得相关性的耗费将大大增加。而为每类标签学习独立的分类器，可以加强方法的灵活性，从而能更好地结合其他优秀算法，以利用其优越性。例如，本文提出的方法是将多标签分类问题转换成相互独立的二元相关性问题，它能将一种单标签主动学习方法应用于其中，使其对多标签数据的分类有更高效率和准确率。

基于图的分类也被许多人研究，如 Yan 和 Han 开发了一种深度优先查询算法——gSpan[18]，这个算法通过在图中建立一种字典式的规则，将各个子图对应地映射到唯一的最小 DFS 编码上作为它的规则标签。许多其他的关于图数据的分类方法也趋于成熟，如 AGM[19]、FSG[20]、MoFa[21]和 Gaston[22]。

以上方法虽然在研究的数据集中的图数据上增加了多标签分类方法的实用性，但这些方法却依然存在为获取大量的标签数据而耗费巨大的问题。因此有必要在研究多标签图数据分类的基础上加入主动学习方法。

2.2 主动学习

近几年，关于图数据的主动学习方法已经被提出。例如，Guillory[23]提出用任意对称子模函数替换 Guillory[24]中的图形切割，从而求解泛化误差界。由于图形切割的尺度越大，预测误差越小，因此 Guillory[23]还提出一种优化算法，这种算法应用子模函数最大化技术使图形切割最大化。Ji[25]也对图数据的分类提出了一种主动学习方法，这种方法是通过最小化 Gaussian Filed and Harmonic Function(GFHF)的预测方差来得到富含最重要信息的节点。在主动学习领域中的研究还有许多，如 Ngomo[26]提出的一种基于遗传规划的主动学习方法，Cesa-Bianchi[27]提出的以迭代的方式查询和预测图上节点标签的方法。

尽管对多标签分类或主动学习的研究已经比较深入，但是在多标签分类中结合主动学习的研究还比较缺乏。本文将对多标签的图数据进行分类，引入一种最小化直推式的拉德马赫复杂度的主动学习方法，得到一种与求解多标签数据分类有关的主动学习方法，其可以高效利用标签数据对未表标签数据进行标签。本文选择了一种半监督式的学习方法，即局部和全局的一致性方法(LLGC)[10]，并给出这种方法与多标签数据分类的目标方程。以这种半监督式的学习方法所获得的目标方程能够很好地结合直推式的拉德马赫复杂度，从而解决将主动学习方法应用于多标签的问题。

3 面向多标签图数据主动学习方法的分析

3.1 多标签 LLGC[10]的定义

定义一个数据空间 χ 和一个有限的标记集 $\omega=\{1,2,3,\cdots,Q\}$。数据集 $T=\{(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)\}(x_i\in\chi)$，其中 $Y_i=[Y_i^1,Y_i^2,\cdots,Y_i^Q]$，$Y_i^k=[Y_{i1}^k,Y_{i2}^k]\in\{0,1\}$。如果 $Y_{i1}^k=1,Y_{i2}^k=0$，那么 x_i 就被标记为 $k(1\leqslant k\leqslant Q)$；$Y_{i1}^k=0,Y_{i2}^k=1$ 则刚好相反；而 $Y_{i1}^k=0,Y_{i2}^k=0$ 表示 x_i 是未标签数据。本文从数据集中抽取一部分数据作为已标签数据，目的是通过已标签数据对未标签数据进行多标签的标记。

给出一个多标签权重图 $G=(V,\xi,Y,K)$，其中 V 为节点集，节点 $V_i\in V$ 对应于相应的数据 x_i；$\xi\subseteq V\times V$ 是一个边缘集，它可反映节点间的相关性；此时，可引入邻接矩阵 W，$W_{ij}\in W$ 反映了 $i\text{-}th$ 节点和 $j\text{-}th$ 节点的相关性。对于无向图而言，W 是一个对称矩阵，Y 是关于节点的标记集；$K:V\cup\xi\rightarrow Y$，它是一个函数，其作用是利用节点集和边集得出标记集。

根据 LLGC，如果 $i\text{-}th$ 节点和 $j\text{-}th$ 节点是相互联系的，那么这两个节点是相似的。对于邻接矩阵 $W\in R^{n\times n}$，可定义[10]：

$$W_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2) \tag{1}$$

且 $W_{ii}=0$。同时，可构建一个加权相似矩阵[10]：

$$S=D^{-1/2}WD^{-1/2} \tag{2}$$

其中，$D_{ii}=\sum_{j=1}^{n}W_{ij}$。

式(1)与式(2)用于获取节点间的相关性，S_{ij} 越大表明节点 i 与节点 j 越相关。

3.2 算法的构造

这一部分采用将多标签分类问题分成多个相互独立的单标签分类问题的方法，融合 LLGC，构建了多标签分类器，并在分类器构建成功的同时引入泛化误差界。其中我们将直推式的拉德马赫复杂度作为工具，用于求解泛化误差界问题。

定义一个 $n\times 2$ 的正定矩阵 κ，矩阵 $K^k=[K_1^k,K_2^k,\cdots,K_n^k]^T\in\kappa$。节点 x_i 第 k 个标签为：

$$y_i^k=\arg\max_{j\leqslant 2}K_{ij}^k \tag{3}$$

定义一个 $n\times 2$ 的矩阵 $Y^k\in\kappa$，其中

$$Y_{ij}^k=\begin{cases}1, & \text{如果 } x_i \text{ 的第 } k \text{ 个标签为 } j\\ 0, & \text{其他}\end{cases}$$

因此 Y_i^k 与起始的多标签数据信息保持一致。

节点的相关信息都蕴含在加权相似矩阵 S 之中，我们可以利用节点间的这一特性，对已获得的多标签信息进行传递。于是，对于与第 k 个标签有关的信息，可获得一迭代方式 $K^k(t+1)=SK^k(t)$，其中 $S_{ii}=0$ 可避免信息的自增强所带来的误差。同时，还需保证能获取起始的多标签信息 Y^k。

根据以上定义和分析，我们给出多标签的迭代公式：

$$K^k(t+1)=\alpha SK^k(t)+(1-\alpha)Y^k \tag{4}$$

其中，$\alpha\in(0,1)$是调整节点相关性和初始多标签信息之间比重的参数。

当 $K^k(t)$达到稳定时，取其极限：

$$K^{k*}=(I-\alpha S)^{-1}Y^k \tag{5}$$

于是每个 x_i 的第 k 个标签为：$y_i^k=\arg\max_{j\leqslant 2} K_{ij}^{k*}$。

K^{k*} 是关于第 k 个标签的分类器，我们的方法就将多标签分类问题分成多独立的单标签分类问题，如算法 1 所示。

算法 1 通过 LLGC 方法求多标签分类器

输入：邻接矩阵 W，矩阵 Y^k(k≤Q)，参数 α；

计算 $S=D^{-1/2}WD^{-1/2}$；

计算 $K^{k*}=(I-\alpha S)^{-1}Y^k$；

for k=0→Q−1 do

  for i=0→n−1 do

$$Y_{ij}^k=\begin{cases}1, & \text{if } y_i^k=j(j\leqslant Q)\\ 0, & \text{其他}\end{cases}$$

  end for

end for

直推式的拉德马赫复杂度与分类器的泛化误差界直接相关，以直推式的拉德马赫复杂度作为一般函数分类的工具，将经验直推式的拉德马赫复杂度最小化，可以有效地对多标签分类器的泛化误差界进行最小化。我们让 k_i 为 x_i 的分类器，则有向量 $K=[k_1,k_2,\cdots,k_n]^T$。

定义 1[11] 让 D 为 χ 上的分布概率，并且来自 χ 的样本 $\{x_i\}_{i=1}^n$ 独立于分布概率 D。让 K 作为 χ 的分类函数，那么关于多标签分类函数 F 的经验直推式的拉德马赫复杂度为：

$$\hat{R}_n(F)=\frac{2}{n}E_\sigma\{\sup_{k\in F}\sum_{i=1}^n\sigma_i k_i\} \tag{6}$$

其中 $\sigma=(\sigma_1,\sigma_2,\cdots,\sigma_n)^T$ 是一个独立的随机变量，见式(7)：

$$\sigma_i=\begin{cases}1, & p\\ -1, & p\\ 0, & 1-2p\end{cases} \tag{7}$$

其中，p 是一概率参数，$0\leqslant p\leqslant\frac{1}{2}$。

携带参数 p 直推式的拉德马赫复杂度为：

$$R_{l+u}(F,p)=(\frac{1}{l}+\frac{1}{u})E_\sigma\{\sup_{k\in F}\sigma^T F\} \tag{8}$$

其中 $l+u=n$。同时，

$$R_n(F)=E_{X\sim D^n}\{\hat{R}_n(F)\} \tag{9}$$

当 $p=\frac{1}{2}$和 $l=u$ 时，$R_{l+u}(F)=2\hat{R}_{l+u}(F)$。在 $p<\frac{1}{2}$的情况下，一些拉德马赫参数会获得零值并降低复杂度。

定理 1[11,12] 让 $c_0=\sqrt{\frac{32\ln(4e)}{3}}<5.05$，$U=(\frac{1}{l}+\frac{1}{u})$和 $\delta\in(0,1)$，然后通过随机获得的概率 $1-\delta$ 构造出 n 个样本，因此 $k\in F$ 满足：

$$e(k)\leqslant\hat{e}(k)+\hat{R}_n(F)+c_0U\sqrt{\min(l,u)}+\sqrt{2U\ln(1/\delta)} \tag{10}$$

其中，$e(k)$是未标签数据的期望误差，$\hat{e}(k)$是已标签数据的经验误差。

定理 1 说明分类器所产生的泛化误差与分类函数 F 的经验直推式的拉德马赫复杂度相关。因此，可以通过最小化经验直推式的拉德马赫复杂度来最小化期望误差 $e(k)$。

3.3 算法优化

由式(10)可知，如果我们能够让 F 的经验直推式的拉德马赫复杂度得到有效的边界，那么上述的误差界将会对许多直推式学习算法起到很好的优化作用。也即可通过优化经验的拉德马赫复杂度来优化本文提出的多标签的分类器方法。

设一个向量 $t=[t_1,t_2,\cdots,t_n]^T$，其中 $t_i\in\{\pm1\}$表示节点 i 已被标记；$t_i=0$ 则相反。

通过式(5)可以得到多标签分类函数 $F_l=\{K=(I-\beta S)^{-1}t,\|t\|_2\leqslant\sqrt{l}\}$。

l 为节点集中含有标签节点的数量，因此，$\|t\|_2\leqslant\sqrt{l}$。于是，对于式(6)有：

$$\hat{R}_n(F)=\frac{2}{n}E_\sigma\{\sup_{K:\|t\|_2\leqslant\sqrt{l}}t^T(I-\beta S)^{-1}\sigma\} \tag{11}$$

通过应用柯西-施瓦茨不等式和詹森不等式，最终得到 F 的经验直推式的拉德马赫复杂度的上界为：

$$\hat{R}_n(F)=\frac{2}{n}\sqrt{\frac{2}{u}tr((I-\beta S)^{-2})} \tag{12}$$

结合式(10)和式(12)可以得到：

$$e(k)\leqslant\hat{e}(k)+U\sqrt{\frac{2}{u}tr((I-\beta S)^{-2})}+c_0\sqrt{\min(l,u)}+\sqrt{2Q\ln(1/\delta)} \tag{13}$$

接下来，将使用顺序优化算法[16]最小化误差上界。顺序优化算法是一种主动学习方法。

首先，给出一个权重图 $G=(V,\xi,Y,K)$，V 是节点集，我们的目标就是找到包含 l 个具有最重要信息的节点的子集 $L\subset V$。而 $u=V/L$ 为剩下的未标记的节点。给出一个加权相似矩阵 S 与图上节点的标记情况相联系，S_{LL} 为关于标记集 L 的主子矩阵，S_{uu} 为未标记集 u 的主子矩阵。

通过式(13)可知，只有通过控制经验直推式的拉德马赫复杂度才能降低误差上限，因此最小化经验直推式的拉德马赫复杂度上界的主动学习准则为：

$$\arg\min_{L\subset V}tr((I-\beta S)^{-2}) \tag{14}$$

上述优化问题是一种组合优化问题，虽然找寻该类问题的全局优化方案是 NP-hard 问题，但顺序优化算法[16]成功解决了该问题，因此引入该优化算法，得到解决式(14)所需的优化算法。

本文引进一个选择矩阵 $H\in R^{n\times l}$，定义为：

$$H_{ij}=\begin{cases}1, & \text{如果 } x_i \text{ 被选择为标记节点}\\ 0,, & \text{其他}\end{cases}$$

于是，选择矩阵 S 可以被定义为：

$$H=\{H|H\in\{0,1\}^{n\times l},H^TH=I\} \tag{15}$$

于是，$S_{LL}=H^{\mathrm{T}}SH$。则式(14)等价于

$$\arg\min_{H\subset H} tr((I-\beta H^{\mathrm{T}}SH)^{-2}) \tag{16}$$

$$\arg\min_{L\subset V} tr((\Gamma^{-1}+U_L^{\mathrm{T}}U_L)^{-1}\Gamma^{-1}) \tag{17}$$

其中，$U_L=H^{\mathrm{T}}U$ 是 U 的子矩阵。

令 $D_0=\Gamma^{-1}$，k 个标记的节点被选择时可表示为 L_k，相应地有 $U_{L_k}\in R^{k\times n}$。让 $D_k=\Gamma^{-1}+U_{L_k}^{\mathrm{T}}U_{L_k}$，然后可以通过下面的优化问题对第 $k+1$ 个节点进行选择。

$$i_{k+1}=\arg\min_{i\subset V/L_k} tr((D_K+u_iu_i^{\mathrm{T}})^{-1}\Gamma^{-1}) \tag{18}$$

通过使用谢尔曼-莫里森公式[12]，可以得到：

$$(D_k+u_iu_i^{\mathrm{T}})^{-1}=D_k^{-1}-\frac{D_k^{-1}u_iu_i^{\mathrm{T}}D_k^{-1}}{1+u_i^{\mathrm{T}}D_k^{-1}u_i} \tag{19}$$

因此

$$tr((D_k+u_iu_i^{\mathrm{T}})^{-1}\Gamma^{-1})=tr(H_k^{-1}\Gamma^{-1})-\frac{u_i^{\mathrm{T}}D_k^{-1}\Gamma^{-1}D_k^{-1}u_i}{1+u_i^{\mathrm{T}}D_k^{-1}u_i} \tag{20}$$

于是，式(18)等价于

$$i_{k+1}=\arg\max_{i\subset V/L_k}\frac{u_i^{\mathrm{T}}D_k^{-1}\Gamma^{-1}D_k^{-1}u_i}{1+u_i^{\mathrm{T}}D_k^{-1}u_i} \tag{21}$$

当第 $k+1$ 个节点被选择时，我们可以通过 D_k^{-1} 获取 D_{k+1}^{-1}：

$$D_{k+1}^{-1}=D_k^{-1}-\frac{D_k^{-1}u_{i_{k+1}}u_{i_{k+1}}^{\mathrm{T}}D_k^{-1}}{1+u_{i_{k+1}}^{\mathrm{T}}D_k^{-1}u_{i_{k+1}}} \tag{22}$$

通过上述方法，最终可获得 l 个具有最重要信息的标记节点。算法 2 为主动学习算法。

算法 2 利用泛化误差界最小化的主动学习方法

输入：加权相似矩阵 S，需要选取的节点数 l，参数 α

执行特征值分解 $S=U\Lambda U^{\mathrm{T}}$

初始化 $D_0=\Gamma^{-1}$

for k=0→l−1 do

计算 $i_{k+1}=\arg\max_{i\subset V/L_k}\frac{u_i^{\mathrm{T}}D_k^{-1}\Gamma^{-1}D_k^{-1}u_i}{1+u_i^{\mathrm{T}}D_k^{-1}u_i}$

获取 $L_{k+1}=L_k\cup\{i_{k+1}\}$

计算 $D_{k+1}^{-1}=D_k^{-1}-\frac{D_k^{-1}u_{i_{k+1}}u_{i_{k+1}}^{\mathrm{T}}D_k^{-1}}{1+u_{i_{k+1}}^{\mathrm{T}}D_k^{-1}u_{i_{k+1}}}$

end for

4 实验

这一节将本文方法和图数据的多标签分类方法进行对比性实验。使用现实生活中存在的多标签数据集，通过对这些图数据的实验验证了本文算法的实用性与准确性。

4.1 数据集

使用一组标准的数据集 PTC1[28]来建立用于实验的基于图的多标签数据集，这组数据包含了 417 个化学成分在 4 种老鼠身上的致癌记录，这 4 种老鼠包括 MM(小型雄性老鼠)、FM(小型母性老鼠)、MR(大型雄性老鼠)、FR(大型母性老鼠)。每一种化学物质在其中一种老鼠身上都有一种致癌反应，这种致癌反应是{CE,SE,P,E,EE,IS,NE,N}中的一种。我们把{CE,SE,P}归为正类标签，即没有致癌效果；把{NE,N}归为负类标签，即存在致癌效果。将 4 种动物模型中不完整的数据集移除，最终获得了 253 个有效的化学成分，它们拥有 4 个标签属性{MM,FM,MR,FR}，并将构成实验所需的基于图的多标签数据集。两个样本间的边缘属性由它们对同类型老鼠造成相同致癌反应的权重值进行构造。

4.2 实验方法与参数设定

一般情况下，将所有数据作为已标签数据而产生的分类器的输出精度是最高的，而至今在还没有对多标签图数据分类使用主动学习方法的情况下，将本文方法即只利用少量标签数据作为输入数据，与将所有数据作为已标签的输入数据作对比，得到它们的精度差，从而得出本文提出的主动学习方法的有效性与实用性。实验方法如下。

1)图的多标签分类方法(ML Graph)，它是直接将节点集中所有的节点作为已标签的输入节点。

2)多标签分类器的泛化误差界最小化方法(ML Bound)，即本文中的主动学习方法。它是通过在多标签分类中引入主动学习方法，来实现用少量的已标记节点数据对大量的无标签节点数据进行标签。参数 α 是可调节的，在本实验中，根据起始标签信息对实验的影响度，将其设为 0.9。

将所有节点数据作为标签数据的多标分类方法是最优的分类方法，但现实生活中要获得大量的标签数据是相当困难且耗费巨大的事，因此这种方法并没有实用意义。然而本文通过将这种方法与主动学习的多标签分类方法作对比，来分析本文方法在输出精度上的优势。

4.3 实验的设定

为了实现实验的准确性与多样性，将用上文提及的一组数据(PTC)进行实验。主动学习方法将在节点集中分别选择{10,20,30,40,50}个节点作为标签节点。

多标签分类系统好坏的评价是不同于单标签分类系统的。接下来本文将使用两种测量方法[17]来评价分类器的性能。假设有一个图的多标签数据集 $T=\{(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)\}$，其中 x_i 被标记为 $Y_i=\{0,1\}^Q$，$f(x_i,k)$是一个输出真实值的分类器，即 x_i 在第 k 个标签的真实标记状况 (L_k)。

1)Hamming Loss：评价预测所得的所有标签与实际标签的不匹配率。$hloss(f)$越小，则分类器的性能越好。

$$hloss(f)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Q}|f(x_i)\Delta Y_i| \tag{23}$$

其中，Δ 表示两个集合的对称差。

(2)One-error：评价隶属度最高的标签与实际标签不一致的概率。评价标准也是 $one\text{-}error(f)$越小，分类器性能越好。

$$one\text{-}error(f)=\frac{1}{n}\sum_{i=1}^{n}\{[\arg\max_{l_k\in L}f(x_i,k)]\notin L_k\} \tag{24}$$

4.4 实验结果

未标签数据的标记结果分别如表 1、图 1 和图 2 所示。从表 1 中我们可以知道 ML Bound 的汉明损失和 1-错误率都比 ML Graph 的大，但数值大小很接近。如图 1、图 2 所示，横坐标表示本文方法所选取的标签节点数，纵坐标则分别表示汉明损失值和 1-误差值。从曲线图可知，随着选取的标签节点数的增加，本文方法(ML Bound)的输出效果也与 ML Graph 的输出效果逐渐接近。ML Graph 是将所有节点数据作为标签数据的多标签分类方法，是输出效果最佳的方法；而本文采用的方法是通过少量的标签数据对大量的未标签数据

进行标记的方法，在现实的应用中能大大降低标记数据的耗费。采用的主动学习方法(Bound)与 ML Graph 方法的输出效果相近，表明了本文方法有低的错误率和好的实用价值。

表 1　ML Bound 与 ML Graph 实验结果对比

	Hamming Loss	One-error
ML Bound	0.049±0.005	0.196±0.010
ML Graph	0.032±0.003	0.125±0.009

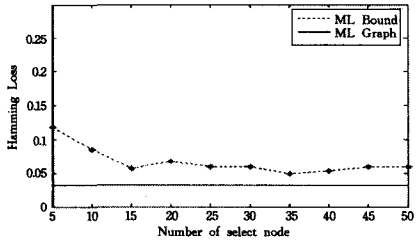

图 1　ML Bound 与 ML Graph 的汉明损失

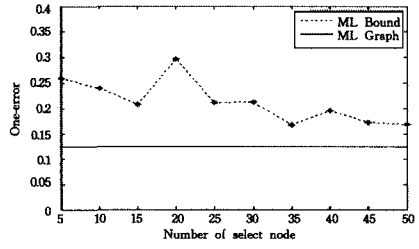

图 2　ML Bound 与 ML Graph 的 1 误差率

总之，该方法是一种结合了主动学习的图数据的多标签分类方法。通过实验证明，其在选取少量的标签数据对大量的数据进行多标签分类时也能得到高的准确率。

结束语　本文解决了结合多标签分类的 LLGC 方法的泛化误差界问题，将主动学习的方法应用于多标签分类问题中。通过将泛化误差界最小化，得到了一小部分含最重要信息的节点，并通过这部分节点让本文提出的分类器的输出有了高的准确率。在以后的研究工作中，将会继续完善本文方法，并寻找更优秀的能应用于多标签分类的主动学习方法。

参 考 文 献

[1] Prati L, Villa A, Lupini A R, et al. Gold on carbon: one billion catalysts under a single label[J]. Physical Chemistry Chemical Physics, 2012, 14(9): 2969-2978

[2] Zhao G, Xuan K, Taniar D. Path KNN query processing in mobile systems[J]. Browse Journals and Magazines, 2013, 60(3): 1099-1107

[3] Chou K-C. Some remarks on predicting multi-label attributes in molecular biosystems[J]. Molecular BioSystems, 2013, 9(6): 1092-1100

[4] Markatopoulou F, Mezaris V, Kompatsiaris I. A comparative study on the use of multi-label classification techniques for concept-based video indexing and annotation[J]. MultiMedia Modeling, 2014, 8325: 1-12

[5] Rubin T N, Chambers A, Smyth P, et al. Statistical topic models for multi-label document classification[J]. Machine Learning, 2012, 88(1/2): 157-208

[6] Kazawa H, Lzumitani T, Taira H, et al. Maximal marging labeling for multi-topic text categorization[J]. Neural Information Processing Systems, 2005, 17: 649-656

[7] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757-1771

[8] Zhang M L, Zhou Z H. Multilabel neural networks with applications to functional genomics and text categorization[J]. Knowledge and Data Engineering, 2006, 18(10): 1338-1351

[9] Chen B, Ding Y, Wild D. Assessing drug target association using semantic linked data[J]. PLoSComput Biology, 2012, 8(7): 1-10

[10] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[J]. Advances in neural information processng systems, 2004, 16(16): 321-328

[11] El-Yaniv R, Pechyony D. Transductiverademacher complexity and its applications[J]. Learning Theory, 2007, 4539: 157-171

[12] Gu Q, Han J. Towards active learning on graphs: An error bound minimization approach[C]// ICDM. 2012: 882-887

[13] Gao S, Wu W, Lee C-H, et al. A MFoM learning approach to robust multiclass multi-label text categorization [C] // ICML. 2004: 42

[14] Crammer K, Singer Y. A new family of online algorithms for category ranking[C]// SIGIR. 2002: 151-158

[15] Ghamrawi N, McCallum A. Collective Multi-Label Classification [C]// CIKM. 2005: 195-200

[16] Boutell M R, Luo J, Shen X, et al. Learning muti-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757-1771

[17] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048

[18] Yan X, Han J. gSpan: Graph-based substructure pattern mining [C]// ICDM. 2002: 721-724

[19] Inokuchi A, Washio T, Motoda H. An apriori-based algorithm for mining frequent substructures form graph data[C]// PKDD. 2000: 13-23

[20] Kuramochi M, Karypis G. Frequent subgraphdiscovery[C]// ICDM. 2001: 313-320

[21] Borgelt C, Berthold M. Mining molecular fragments: Finding relevant substructures of molecules[C]// ICDM. 2002: 211-128

[22] Nijssen S, Kok J. A quickstart in frequent structure mining can make a difference[C]// KDD. 2004: 647-562

[23] Guillory A, Bilmes J. Active semi-supervised learning using submodularfunctions[C]// UAI. 2011: 274-282

[24] Guillory A, Bilmes J A. Label selection on graphs[C]// NIPS. 2009: 669-691

[25] Ji M, Han J. A variance minimization criterion to active learning on graphs[C]// AISTATS. 2012: 556-564

[26] Ngomo A-C N, Lyko K. EAGLE: Efficient active learning of link specifications using genetic programming [J]. The Semantic Web: Research and Applications, 2012, 7295: 149-163

[27] Cesa-Bianchi N, Gentile C, Vitale F, et al. Active learning on trees and graphs[C]// COLT. 2010: 320-332

[28] Helma C, King R, Kramer S, et al. The predictive toxicology challenge 2000-2001[J]. Bioinformatics, 2001, 17(1): 107-108