一种基于随机游走模型的多标签分类算法

郑 伟 王朝坤 刘 璋 王建民

(清华大学软件学院 北京 100084)
(清华信息科学与技术国家实验室 北京 100084)
(信息系统安全教育部重点实验室 北京 100084)

摘 要 在数据挖掘领域,传统的单分类和多分类问题已经得到了广泛的研究.但是多标签数据的普遍存在性和重要性直到近些年来才逐渐得到人们的关注.在多标签分类问题中,由于标签相关性的存在,传统的单分类和多分类问题的解决方法,无法简单地应用于多标签分类问题.文中提出了一种基于随机游走模型的多标签分类算法,称为多标签随机游走算法.首先,将多标签数据映射成为多标签随机游走图.当输入一个未分类数据时,建立一个多标签随机游走图系列.而后,对图系列中的每个图应用随机游走模型,得到遍历每个顶点的概率分布,并将这个点概率分布转化成每个标签的概率分布.最后,基于多标签随机游走算法,文中给出了一种新的阈值学习算法.真实数据集上的实验表明,多标签随机游走算法可以有效地解决多标签分类问题.

关键词 多标签;分类算法;随机游走;阈值学习
中图法分类号 TP181 **DOI 号**: 10.3724/SP.J.1016.2010.01418

A Multi-Label Classification Algorithm Based on Random Walk Model

ZHENG Wei WANG Chao-Kun LIU Zhang WANG Jian-Min

(*School of Software, Tsinghua University, Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology, Beijing* 100084)
(*Key Laboratory for Information System Security, Ministry of Education, Beijing* 100084)

Abstract There are extensive literatures related to traditional single-class and multi-class classification problems, in which each data point is assigned to one category. But in many applications, a data point may belong to more than one category. This kind of problem is called the Multi-Label Classification (MLC) problem. Due to the existing of label relevance, the traditional data-mining methods cannot be directly applied to the MLC problems. This paper proposes a novel MLC algorithm based on the random walk model, called Multi-Label Random Walk (MLRW) algorithm. Firstly, a multi-label random walk graph is built on the training set. As an unlabeled data arrives, a multi-label random walk graph system will be built, on which the random walk processing is carried out. After that, a probability distribution among all labels is obtained. At last, a threshold learning algorithm is proposed based on the MLRW algorithm so that the final prediction on each label is presented. Experimental results on actual data set show that the MLRW algorithm provides an effective solution to the MLC problems.

Keywords multi-label; classification; random walk; threshold learning

---

收稿日期:2010-06-11.本课题得到国家自然科学基金(60803016)、国家"九七三"重点基础研究发展规划项目基金(2007CB310802,2009CB320706)和国家"八六三"高技术研究发展计划项目基金(2008AA042301,2007AA040602)资助.**郑 伟**,男,1986 年生,硕士研究生,主要研究方向为多标签数据的分类和聚类、数字音乐信息检索.E-mail: zhengw04@mails.tsinghua.edu.cn. **王朝坤**,男,1976 年生,博士,讲师,主要研究方向为音乐数据管理与云计算. **刘 璋**,男,1985 年生,博士研究生,主要研究方向为非结构化数据管理和音乐数据管理. **王建民**,男,1968 年生,博士,教授,博士生导师,主要研究领域包括数据管理与信息系统、云环境下非结构化数据管理技术、业务过程与产品生命周期管理、数字版权与系统安全技术、数据库测试技术.

1 引　言

数据分类(data classification)是数据挖掘(data mining)的一个重要研究方向.一直以来,数据分类问题和方法受到了人们的广泛关注和研究.

传统数据分类问题的研究目标是如何将每条数据准确地划分到某一类中.如果候选类别只有一个,则分类目标转化为判断未分类数据是否属于该类别,这类问题被称作单分类问题(single-class classification)或二值分类问题(binary classification).如果候选类别有多个,在传统的分类问题中,分类器仅能在这些候选类别中选择一个作为输出,这类问题被称作多分类问题(multi-class classification).多分类问题可以比较容易地转化成单分类问题.单分类问题和多分类问题统称为单标签分类问题(single-label classification).它们和本文研究的多标签分类(multi-label classification)问题有着本质的区别[1].

在实际应用中,普遍存在如下情况:一条数据可能同时属于多个不同的类别.这类数据被称作多标签数据.例如,Lewis 等研究了路透社的 804414 条新闻,发现平均每条新闻同时属于 2.6 个不同的类别[2];在 ACM Computing 分类体系中,存在着一级类别 11 个、二级类别 81 个,而作者可以为每篇文章选择多个不同的类别[3];Snoek 等人通过分析 43907 个从非洲、中国和美国收集的音频片段以及与这些音频片段相关的 101 个标签,发现平均每个音频片段具有 4.4 个不同的标签[4].这样的分类问题被称作多标签分类问题(见定义 1).和传统的单标签分类问题相比,多标签分类问题存在着显著的区别,类别间的相关性(relevance)和共现性(co-occurrence)直接导致传统的单标签分类方法不能被直接应用到多标签分类问题中[1,5].多标签分类问题正逐渐成为当前的一个研究热点.

多标签分类问题的形式化定义如下所示.

定义 1.　已知一个定义在实数域 R 上的 d 维输入数据空间,记作 $X=R^d$;一个包含 q 个标签的标签集合,记作 $Y=\{\lambda_1,\lambda_2,\cdots,\lambda_q\}$ 和一个包含 m 个训练数据的训练集合,记作

$$D=\{(x_i,Y_i)\mid 1\leq i\leq m,\ x_i\in X,Y_i\subset Y\} \quad (1)$$

其中 x_i 是输入空间 X 中的一个训练数据,Y_i 是 x_i 的真实标签集合(actual label set).

多标签分类问题指:根据训练数据 D 学习分类函数 $f:X\to 2^Y$,当输入一个未分类数据 $x\in X$ 时,通过函数 f 得到 x 的预测标签集合 $P_x\subset Y$,使得 P_x 与 x 的真实标签集合 Y_x 最为接近.

易知,单标签分类问题是多标签分类问题的一个特例.当训练和测试数据都满足 $|Y_i|=1$ 时,多标签分类问题退化成多分类问题.特别地,当 $q=1$ 时,多标签分类问题退化为单分类问题.

多标签排序问题是与多标签分类问题直接相关的一类问题,其形式化定义如下.

定义 2.　已知输入空间 X、标签集合 Y 和训练数据集合 D 如定义 1 所示.

多标签排序问题指:根据训练数据 D 学习函数 $g:X\times Y\to R$,当输入一个未分类数据 $x\in X$ 时,对于任意的 $y\in Y$,得到一个置信系数 $g(x,y)$,并根据该置信系数对 Y 中的所有标签进行排序,使得此排序结果与真实结果最为接近.

不同的指标被用于度量分类或排序结果的正确性,例如 Precision/Recall/F-Measure、Subset accuracy、Hamming Loss、One-Error、Ranking Loss、Coverage、Average Precision 等[5](见 5.1 节).

多标签分类问题主要有两大类解决方法:基于问题转化的方法和基于算法转化的方法[5](见第 2 节).本文提出了一种基于随机游走模型的多标签分类算法 MLRW(A multi-label classification algorithm based on the random walk model).主要贡献有

(1)提出了一种新的多标签分类算法 MLRW.MLRW 在预测未分类数据时,除了能够给出分类结果,还可以结合条件概率模型,给出该数据具有每个标签的概率分布.

(2)基于 MLRW 算法,提出了一种分类阈值学习方法,该方法可以解决多标签分类算法中的阈值设置问题.

(3)真实数据集上的实验结果表明,MLRW 算法和分类阈值学习方法能有效解决多标签分类问题和多标签排序问题.

本文第 2 节介绍与本文有关的研究工作;第 3 节和第 4 节分别给出 MLRW 算法和阈值学习方法;第 5 节给出 MLRW 算法的相关讨论;第 6 节介绍实验方法和实验结果;最后总结全文.

2 相关工作

近年来,多标签分类和排序问题受到了人们的广泛关注和研究.其解决方法主要分为基于问题转

化的方法(Problem Transformation based methods,PT)和基于算法转化的方法(Algorithm Adaptation based methods,AA).

2.1 基于问题转化的方法

PT 类方法的主要目标是将一个多标签分类问题转化成一个或一组单标签分类问题,从而运用已有的单标签分类方法解决该问题.

BR(Binary Relevance)是一种典型的 PT 方法,它将每个标签的预测看作一个独立的单分类问题,并为每个标签训练一个独立的分类器,用全部的训练数据对每个分类器进行训练.这种算法忽略了标签之间的相互关系,往往无法达到令人满意的分类效果.文献[6]通过拷贝(copy)和带权重拷贝(copy-weight)的方法,对 BR 进行改进,将原训练集合中的一条多标签数据拆分成多条单标签数据,并给予相应的权重.

Hullermeier 等提出了基于标签对比(pairwise comparison)的分类方法.通过对比标签集合中任意两个标签之间的关系,建立 $q(q-1)/2$ 个分类器.每个分类器在两个标签 λ_i 和 λ_j 之间投票,然后组合这些投票结果作为最终的多标签分类结果[7].假设多标签分类算法中采用的基础分类器(base classifier)的复杂度为 $O(t(D))$,其中函数 $t(D)$ 表示分类器在训练集合 D 上建立分类模型的复杂度,则基于标签对比的多标签分类算法的复杂度为 $O(q(q-1)/2 \cdot t(D))$.

LP(Label Powerset)是另外一种被广泛使用的 PT 方法.它将训练数据中的每种标签组合进行二进制编码,从而形成新的标签.在 LP 中,多标签数据被以这种方式转化成单标签数据.LP 算法的显著缺点是不能预测新的标签组合.Read 等将概率分布模型应用到 LP 中,当对未分类数据进行预测时,可以预测出训练集合中未出现的标签组合[8].但是 LP 算法的复杂度较高,达到 $O(\min\{2^q,m\} \cdot t(D))$,可以通过剪枝[8]或随机标签组合[9]的方法在一定程度上降低复杂度,但降低的幅度有限.

2.2 基于算法转化的方法

AA 类方法的主要目标是,通过改变已有的单标签分类算法,使其能够处理多标签数据.典型的 AA 算法有以下几种:

基于单标签分类算法 AdaBoost.M1,Schapire 等提出了适用于多标签数据的 AdaBoost.MH 算法[10],该算法使用每个多标签训练数据生成 q 个新的单标签训练数据.该算法的主要缺点是,显著地增加了训练数据的数量,进而增加了建模时的消耗.

人工神经元网络也可以应用到多标签分类问题中.Zhang 等人通过定义针对多标签数据的全局优化函数,使得人工神经元网络能够处理多标签数据[11].该算法基本思想是,如果很多实例同时具备两个标签,那么这两个标签中的一个出现了,另外一个也很可能同时出现.

经典的 kNN 方法也可以应用到多标签分类问题中,例如文献[12]中介绍的 MLkNN 算法.MLkNN 通过统计方法,得出每个标签的先验概率.当输入一个未分类数据 x 时,对标签集合 Y 中的每个标签 λ,分别计算 x 具有标签 λ 和不具有标签 λ 的概率,进而预测 x 是否具有标签 λ.

C4.5 决策树也可应用于多标签分类问题中,只需要将单标签分类问题中熵的定义扩展到多标签分类问题.Clare 等定义多标签分类问题中的熵为

$$MLEntropy=\sum_{y\in Y}p(y)\log p(y)+(1-p(y))\log(1-p(y)),$$

而后便可以基于熵计算信息增益,从而对多标签数据建立决策树[13].此外,经典的 Bayes 等算法也可以通过修改而被用于多标签分类问题中.

此外,基于已有的多标签分类算法,Tsoumakas 等提出了二层的多标签分类模型,第一层中采用 BR、决策树或 SVM 等进行 k-fold 交叉训练;在第二层中,采用 BR、SVM 等算法,使用第一层训练后得到的各标签的得分或概率分布作为输入,来预测最终的标签输出结果[14].

2.3 随机游走模型

随机游走模型的基本思想是,从一个或一系列顶点开始遍历一张图.在任意一个顶点,遍历者将以概率 $1-\alpha$ 游走到这个顶点的邻居顶点,以概率 α 随机跳跃(teleport)到图中的任何一个顶点,称 α 为跳转发生概率.每次游走后得出一个概率分布,该概率分布刻画了图中每一个顶点被访问到的概率.用这个概率分布作为下一次游走的输入并反复迭代这一过程.当满足一定前提条件时,这个概率分布会趋于收敛.收敛后,即可以得到一个稳定的概率分布.

随机游走模型广泛应用于数据挖掘和互联网领域,PageRank 算法可以看作是随机游走模型的一个实例[15].Zhang 等人使用该模型从书评中挖掘关键词[16];Zhu 等人提出了有吸收状态的随机游走模型,该模型可以用于文本自动摘要(text summarization)和基于社会网络的分析与挖掘[17].本文使用收

敛后的概率分布来刻画未分类数据具有每个标签的概率.

3　MLRW 和阈值学习算法

3.1　随机游走图的生成

MLRW 算法首先将训练集合 D 映射成 d 维度空间中的多标签随机游走图. 我们使用随机游走模型的原因是：该模型通过点与点之间的连通性准确地刻画训练数据之间的相关性，进而刻画候选标签之间的相关性. MLRW 的基本思路是：将集合 D 中的每个训练数据 $x \in X$ 映射为图中的一个点 v. 如果两个训练数据 x_i、x_j 具有相同的标签，则将这两个训练数据对应的顶点 v_i、v_j 相连. 形式化地，已知训练集合 D 如式(1)所示，则由训练集合 D 导出的多标签随机游走图记作：

$$G=(V,E) \tag{2}$$

$$V=\{v_i \mid x_i \in X, 1\leq i\leq m\} \tag{3}$$

$$E=\{(v_i,v_j) \mid v_i,v_j \in V, Y_i \bigcap Y_j \neq \varnothing, i\neq j\} \tag{4}$$

如无特别说明，本文余下部分使用 v_i 表示训练数据 x_i 在随机游走图上对应的顶点. 例如，式(3)表示每个训练数据 x_i 将对应图 G 中的一个顶点 v_i，这些顶点构成了图 G 的顶点集合 V.

接下来，我们计算随机游走图 G 上的权重矩阵 $\boldsymbol{W}$. 如式(5)，边的权值即为训练数据对应顶点在 d 维空间中的距离，记作 $dis(v_i,v_j)$. 本文采用欧式距离作为距离函数.

$$W_{ij}=\begin{cases}0, & v_i=v_j\\ \infty, & v_i\neq v_j,\ (v_i,v_j)\notin E\\ dis(v_i,v_j), & v_i\neq v_j,\ (v_i,v_j)\in E\end{cases} \tag{5}$$

不失一般性，可以假定图 G 是连通的. 如果 G 中存在不连通的子图，则说明标签集合 Y 中存在相互独立的标签子集，G 中的每个连通分量对应 Y 中的一个独立子集. 此时，我们可以根据 G 中的连通分量，将标签集合 Y 拆分成多个互不相交的子集，并对每个子集分别应用 MLRW 算法. 因此，本文后面的内容都将基于图 G 是连通图这一前提展开.

例如，给定一个标签集合 $Y=\{\lambda_1,\lambda_2\}$，训练集合中包含 6 条数据(如表 1)，训练数据 x_1，x_2，x_3 有相同的标签 λ_1，则将这 3 个点两两相连，连接它们的边的权重即为这 3 个数据的特征向量在输入空间中的欧式距离. 同理，x_3，x_4，x_5，x_6 同时具有标签 λ_2，把它们两两相连，则由训练集合 D 导出的随机游走图如图 1 所示.

表 1　训练集合 D 示例

	特征向量	标签集合
1	$\boldsymbol{x}_1$	$Y_1=\{\lambda_1\}$
2	$\boldsymbol{x}_2$	$Y_2=\{\lambda_1\}$
3	$\boldsymbol{x}_3$	$Y_3=\{\lambda_1,\lambda_2\}$
4	$\boldsymbol{x}_4$	$Y_4=\{\lambda_2\}$
5	$\boldsymbol{x}_5$	$Y_5=\{\lambda_2\}$
6	$\boldsymbol{x}_6$	$Y_6=\{\lambda_2\}$

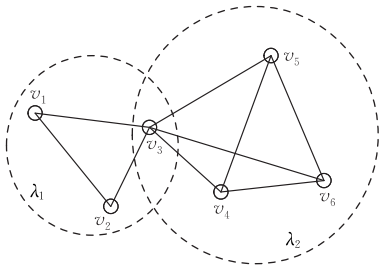

图 1　训练集合 D 对应的图

可以看出，随机游走模型无法直接应用于传统的单标签分类问题. 如果将单标签数据映射为随机游走图，得到的将会是不连通图，这不满足随机游走算法的收敛条件(见定理 1). 然而，将多标签数据映射为随机游走图时，满足随机游走模型的收敛条件，因而得以应用.

3.2　多标签随机游走过程

3.2.1　随机游走过程

随机游走过程需要 4 个输入参数：邻接矩阵 $\boldsymbol{P}$ (adjacent matrix)，初始概率分布向量 $\boldsymbol{s}_0$，跳转发生概率 α(teleporting probability)，发生跳转时跳转到图中每个顶点的概率分布向量 $\boldsymbol{d}$. 每次游走过程后的输出概率分布向量记作 $\boldsymbol{s}$，则 $\boldsymbol{s}$ 的计算法方法为

$$\boldsymbol{s}=(1-\alpha)\cdot\boldsymbol{P}^{\mathrm{T}}\cdot\boldsymbol{s}_0+\alpha\cdot\boldsymbol{d},\ 0<\alpha<1 \tag{6}$$

将向量 $\boldsymbol{s}$ 作为式(6)的输入 $\boldsymbol{s}_0$，反复迭代式(6)直至收敛，将此时的概率分布向量记作 $\boldsymbol{\pi}$，满足

$$\boldsymbol{\pi}=(1-\alpha)\cdot\boldsymbol{P}^{\mathrm{T}}\cdot\boldsymbol{\pi}+\alpha\cdot\boldsymbol{d} \tag{7}$$

式(7)中的向量 $\boldsymbol{\pi}$ 即为稳定的概率分布向量.

为了应用式(6)，首先基于权重矩阵 $\boldsymbol{W}$ 计算邻接矩阵 $\boldsymbol{P}$. 基本思想是，对任意顶点 v，在 v 的所有邻居顶点中，如果一个顶点距离 v 越远，则游走到这个顶点的概率就越低，如式(8)所示.

$$M_{ij}=\begin{cases}0, & w_{ij}=\infty\\ \dfrac{w_{ij}}{\max\limits_{1\leq k\leq m}\{w_{kj} \mid w_{kj}\neq\infty\}}, & w_{ij}\neq\infty\end{cases} \tag{8}$$

对矩阵 $\boldsymbol{M}$ 进行归一化处理：

$$M'_{ij}=\frac{M_{ij}-avg_i\{M_{ij}\}}{std_i\{M_{ij}\}} \tag{9}$$

$$P_{ij}=\frac{M'_{ij}}{\sum\limits_i M'_{ij}} \tag{10}$$

此时的概率分布矩阵 $\boldsymbol{P}$ 即为算法输入的邻接矩阵.

3.2.2 多标签随机游走图系列

当输入一个未分类数据 x 时，将 x 对应的顶点记作 u，MLRW 将以 u 作为起始点应用 q 次随机游走模型. 具体地，在第 k 次应用随机游走模型时，将 u 与所有具有标签 λ_k 的点相连得到多标签随机游走图 G_k. 我们将这些图的集合 $\{G_k\}(k=1,2,\cdots,q)$ 定义为多标签随机游走图系列(如图 2 所示).

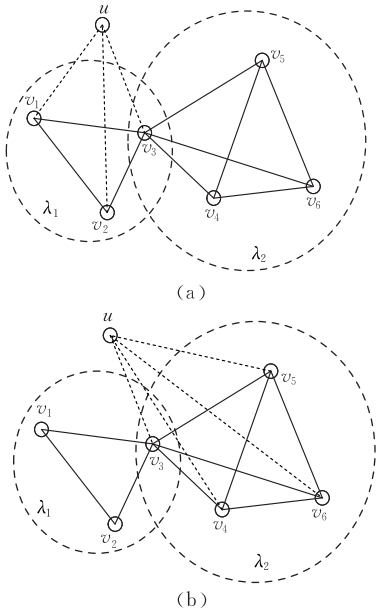

图 2 输入未分类数据 x 时(将 x 对应的点记作 u)：(a)将 u 与具有标签 λ_1 的点(v_1，v_2，v_3)相连，得到图 G_1；(b)将 u 与具有标签 λ_2 的点(v_3，v_4，v_5，v_6)相连，得到图 G_2

定义 3. 已知多标签随机游走图 G、标签集合 Y 和一个未分类数据 x，则定义由 G 和 x 导出的多标签随机游走图系列为 $T=\{G_k \mid k=1,2,\cdots,q\}$，其中

$$G_k=(V_k,E_k) \tag{11}$$

$$V_k=V\cup\{u\} \tag{12}$$

$$E_k=E\cup\{(u,v_i)\mid \lambda_k\in Y_i,1\leq i\leq m\} \tag{13}$$

其中，u 是未分类数据 x 对应的顶点，v_i 是训练数据 x_i 对应的顶点，Y_i 是 x_i 的真实标签集合.

接下来，我们对 T 中的每个图以 u 为起点应用本文 3.2.1 节所描述的随机游走过程. 此时，我们还需要计算初始向量 $\boldsymbol{s}_0$. 首先计算 $\boldsymbol{s}_0'$，$\boldsymbol{s}_0'$ 是一个 m 维向量，它的第 i 个元素为

$$\boldsymbol{s}_0'(i)=\begin{cases} dis(u,v_i), & (u,v_i)\in E_k \\ 0, & \text{其它} \end{cases} \tag{14}$$

对 $\boldsymbol{s}_0'$ 使用类似于式(8)～(10)的方法进行归一化，即得到初始向量 $\boldsymbol{s}_0$.

在本文中，我们假设从某个顶点出发跳转到图中任意一个顶点的概率是相等的，得到随机跳转到每个顶点的概率分布向量：

$$\boldsymbol{d}=\left(\frac{1}{m},\frac{1}{m},\cdots,\frac{1}{m}\right) \tag{15}$$

此外，我们还得知 α 的一般取值为 0.15[16]，本文的实验部分将对 α 取值对结果的影响进行讨论.

至此，我们已经得到了随机游走过程所需的所有输入，将它们代入式(6)中，可以得到概率分布向量 $\boldsymbol{s}$，通过反复迭代式(6)直至收敛，可得出最终的概率分布向量 $\boldsymbol{\pi}$. $\boldsymbol{\pi}$ 刻画了将未分类数据 x 对应的顶点 u 与具有标签 λ_k 的数据对应的顶点相连(记作 $x<\lambda_k$)时，以 u 点为起点游走到图 G_k 中每个顶点的条件概率. 我们将该条件概率记作

$$P(v_i\mid x<\lambda_k)=\pi(i),\ 1\leq i\leq m \tag{16}$$

其中，$\pi(i)$ 表示向量 $\boldsymbol{\pi}$ 的第 i 个元素. 将每个标签对应的点取其条件概率的平均值，即为以 u 点为起点遍历图 G_k 时游走到每个标签的平均条件概率：

$$P(\lambda_j\in Y_x\mid x<\lambda_k)=avg\{P(v_i\mid x<\lambda_k)\mid \lambda_j\in Y_i\} \tag{17}$$

3.2.3 条件概率模型

根据条件概率模型，未分类数据 x 具有标签 λ_j 的概率可以采用以下公式计算：

$$P(\lambda_j\in Y_x)=\sum_{1\leq k\leq q}P(\lambda_j\in Y_x\mid x<\lambda_k)P(x<\lambda_k) \tag{18}$$

因此，我们还需要求出未知数据与具有标签 λ_k 的点相邻的先验概率 $P(x<\lambda_k)$.

在本文中，使用 u 点和具有标签 λ_k 的数据对应顶点的平均距离来刻画这个先验概率. 即平均距离越大，则该先验概率的值就越小. 为此，首先计算一个临时变量，记作

$$w(x<\lambda_k)=avg\{dis(x,v_i)\mid \lambda_k\in Y_i\} \tag{19}$$

而后，使用类似于式(8)～(10)的方法对式(19)进行归一化，即可得到所需的先验概率 $P(x<\lambda_k)$，将其代入式(18)，得到最终的概率分布结果.

MLRW 算法的形式化描述如图 3 所示.

3.3 图剪枝

标签集的势指平均每条数据具有的标签数[5]，记作 c. 我们发现，当训练集中标签的势较大时，图 G 中边的数量会大大增加. 这是因为，平均每个标签关联的数据为 $O(mc/q)$，则每个点平均具有边 $O(mc^2/q)$，图 G 中边的总数为 $O(m^2c^2/2q)$. 由此可以看出图 G 中的总边数随 c 的增大而快速增大，当 $c\geq\sqrt{q}$ 时，MLRW 算法的空间消耗快速上升. 因此，我们对图 G 进行如下剪枝，以降低算法的空间消耗.

定义 4. 已知图 $G=(V,E)$，其上的权重矩阵为 $\boldsymbol{W}$，则图 G 上的 Top-k 剪枝指的是，对每个顶点 $v_i\in V$，将其相关联的所有边 $\{(v_i,v_j)\mid i\neq j,(v_i,v_j)\in$

$E\}$，按照其权重 W_{ij} 排序，保留其中权重最小（即距离最近）的 k 条边，将其它边从图 G 中删除.

我们将在本文实验部分对剪枝的影响进行讨论.

算法 1.　多标签随机游走算法 MLRW.
输入：训练数据集合 D，随机跳转发生概率 α，
　　未分类数据 x，标签集合 Y
输出：x 具有 Y 中每个标签的概率分布 $P(\lambda_j \in Y_x)$
MLRW(D,α,x,Y)
1. 初始化数组 PC、PP　　//临时保存条件概率、先验概率
2. 构造随机跳转到每个顶点的概率分布向量 $\boldsymbol{d}$ //式(15)
3. for $k \leftarrow 1$ to q do
4.　根据 D,x,λ_k 构造随机游走图 G_k　　//式(11)～(13)
5.　根据 G_k 计算邻接矩阵 $\boldsymbol{P}$　　//式(8)～(10)
6.　根据 G_k 计算随机游走初始向量 $\boldsymbol{s}_0$　　//式(14)
7.　应用随机游走模型得出条件概率分布向量，记作 $\boldsymbol{Q}$：
　　$\boldsymbol{Q}(j)=P(\lambda_j \in Y_x | x<\lambda_k), j=1,2,\cdots,q$ //式(16)～(17)
8.　将向量 $\boldsymbol{Q}$ 保存到数组 PC 中，即 $PC[k]=\boldsymbol{Q}$
9.　计算先验概率并保存在数组 PP 中：$PP[k]=P(x<\lambda_k)$
10. end for
11. foreach $\lambda_j \in Y$ do
11.　根据 PP,PR 计算概率分布 $P(\lambda_j \in Y_x)$ //式(18)
12. end for

图 3　多标签随机游走算法 MLRW

4　分类阈值学习方法

由式(18)，当输入一个未分类数据 x 时，可求出 x 具有每个标签的概率分布. 通过该概率分布，可以得到一个排序后的标签集合. 此时，为了决定每个标签的取舍，还需要为每个标签给定一个阈值，将概率大于阈值的标签集合作为 x 的预测标签集合 P_x.

多标签分类中的阈值确定问题，同样得到了人们的广泛研究. 例如，Fan 等提出了 SCutFBR 算法[18]，Tang 等人提出了基于训练数据的阈值学习方法[19]. 但是，这些阈值学习的方法由于没有与具体的分类方法相结合，往往难以取得好的效果.

本文基于 MLRW 算法，给出一种新的阈值学习方法. 具体地，首先对训练集合 D 进行随机采样，生成采样集合 D'. 对 D' 中的每一个数据 x_i，以 x_i 对应的顶点为输入应用 MLRW 算法，由式(18)可以得到一个 q 维的概率分布向量，记作 $\boldsymbol{P}_i$. 而后我们使用这 $|D'|$ 个向量通过如下操作得到一个 q 维的接受阈值(accept threshold)向量和一个 q 维的拒绝阈值(reject threshold)向量，分别记作 $\boldsymbol{P}_a$、$\boldsymbol{P}_r$，如

$$\boldsymbol{P}_a(j)=avg\{\boldsymbol{P}_i(j) | x_i \in D', \lambda_j \in Y_i\} \quad (20)$$

$$\boldsymbol{P}_r(j)=avg\{\boldsymbol{P}_i(j) | x_i \in D', \lambda_j \notin Y_i\} \quad (21)$$

其中，$\boldsymbol{P}_a(j)$ 表示向量 $\boldsymbol{P}_a$ 的第 j 个元素，其它类同. 最终的阈值向量为这两个阈值的平均：

$$\boldsymbol{P}_T=avg\{\boldsymbol{P}_a,\boldsymbol{P}_r\} \quad (22)$$

当输入一个未分类数据时，首先通过算法 1 得到 x 具有每个标签的概率，而后与阈值向量 $\boldsymbol{P}_T$ 比较，进而确定每个标签的有无.

5　算法讨论

5.1　收敛性

定理 1.　MLRW 算法是收敛的.

证明.

(1) 因为向量 $\boldsymbol{d}$ 中不包含非零元素，并且 $0<\alpha<1$，所以从任意点开始，随机跳跃到图 G_k 中的任意点都是可能的，故邻接矩阵 $\boldsymbol{P}$ 是不可规约的(irreducible).

(2) 当随机游走过程遍历到某一顶点后，再次遍历到这个顶点所需的步数是不确定的，故整个随机游走过程是一个非周期的过程(aperiodic).

(3) 显然，当图中任意一个顶点被遍历后，都可能在有限步数内再次遍历这个顶点，且再次遍历之前经过的步数是不完全相同的(positive recurrent).

由以上 3 点，可以得出 MLRW 算法是各态历经的(ergodic)[20]，故此算法是收敛的. 即存在向量 $\boldsymbol{\pi}$，满足式(7).　　证毕.

5.2　复杂度分析

定理 2.　MLRW 算法的时间复杂度为 $O(q\log m)$，空间复杂度为 $O(q|E|)$. 其中，q 表示标签集合 Y 的大小，E 表示由训练集合 D 导出的随机游走图 G 中边的集合，m 是训练集的大小.

证明.　算法 1 的第 3～6 行循环的复杂度由算法第 6 行随机游走的迭代次数决定. 根据文献[21]，MLRW 算法中随机游走的迭代次数为 $O(\log|E|)=O(\log m)$. $|Y|=q$，故 MLRW 算法的时间复杂度为 $O(q\log m)$.

算法 1 中需要存储的变量为转移矩阵 $\boldsymbol{P}$、随机跳转向量 $\boldsymbol{d}$、每次迭代后的概率分布向量 $\boldsymbol{s}$、数值 α 和 λ_k. 其中，转移矩阵 $\boldsymbol{P}$ 的大小等于图 G_k 中边的数目. 故算法的空间复杂度为 $O((m+m+|E|)q+1+1)$，由于图 G_k 连通，所以总的空间复杂度为 $O(q|E|)$. 证毕.

6　实　验

6.1　数据集和度量标准

本文采用 yeast 数据集①，该数据集是对啤酒酵

① 数据集可在 http://mulan.sourceforge.net/datasets.html 下载

母菌细胞基因表达的研究结果. 经过微阵列实验(microarray experiments),大量的基因片段(大约3300个)被按照功能进行分类,其中的2417条数据构成了yeast数据集[22],其统计数据如表2所示. 其中标签密度等于标签集的大小 q 除以标签集的势 c,表示每个标签出现的平均概率.

表2 数据集统计信息

数据集名称	训练集大小 m	测试集大小	特征空间维度 d	标签集大小 q	标签密度	标签集的势 c
yeast	1500	917	103	14	0.30	4.20

本文使用平均精度(*Avg-Precision*)、(*One-Error*)、结果覆盖长度(*Coverage*)等指标对实验结果进行度量,它们的定义分别为[23]

$$One\text{-}Error=\frac{1}{m}\sum_{i=1}^{m}I(\arg\min_{\lambda\in Y}r_i(\lambda)\notin Y_i) \quad (23)$$

$$Coverage=\frac{1}{m}\sum_{i=1}^{m}\max_{\lambda\in Y_i}r_i(\lambda)-1 \quad (24)$$

$$Avg\text{-}Precison=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y|}\sum_{\lambda\in Y_i}\frac{\{\lambda'\in Y_i,r_i(\lambda')\leq r_i(\lambda)\}}{r_i(\lambda)} \quad (25)$$

其中 $r_i(\lambda)$ 表示标签 λ 的排名.

本文实验环境为Intel Core 2.33GHz的CPU,4GB内存,1.5TB硬盘的PC机. 操作系统为Ubuntu 9.10,Java版本Sun JDK 1.6.0.

6.2 实验结果

6.2.1 对比实验

我们基于MuLan① 实现了MLRW算法. MuLan是一个基于Weka② 的开源项目,它实现了一些最近提出的多标签分类和排序算法.

实验中采用的对比算法有Homer、BR(Binary Relevance)、CLR(Calibrated Label Ranking)、MLkNN(Multi-Label k-nearest neighbor)、RAkEL(Random-k Labelsets)、LP(Label Powerset)等(见第2节). 其中Homer算法中Cluster的数量为3,MLkNN中 $k=10$,其它均采用默认参数. BR、CLR、IncludeLabels、RAkEL、LP算法的基础分类器(base classifier)采用SVM分类器,该SVM分类器采用线性核函数,常数 c 的值为1. Homer分类算法采用CalibratedLabel分类器作为基础分类器.

在对比实验中,我们将原有数据集中的训练集和测试集混合,随机重新采样排序,然后用10-fold交叉验证(cross validation)的方法对结果进行验证. 对以上实验重复进行10次,取其平均值. MLRW算法中剪枝粒度 k 设定为100.

如表3所示,MLRW算法可以达到较好的平均精度和较小的误差. MLRW算法的平均精度与MLkNN算法几乎相同,但MLRW算法的结果覆盖长度比较小,也就是说,使用MLRW算法可以用较小的误差找全所有的正确标签集合. 而较低的 *One-Error* 值(相比MLkNN领先5.6%),则说明MLRW算法给出的排名最靠前的标签(top-one related label)不在该数据实际标签集合中的概率较低. 这在信息检索应用中非常重要,因为大多数用户往往只关心排名靠前的检索结果[24].

表3 各算法实验结果对比

	One-Error	*Coverage*	*Avg-Precison*
Homer	0.2501	8.2302	0.6955
BR	0.3664	7.5070	0.6613
CLR	0.2597	6.7971	0.7097
MLkNN	0.2844	7.4143	0.7284
LP	0.5267	9.5965	0.5633
RAkEL	0.2911	7.6543	0.7096
MLRW	0.2451	7.4100	0.7069

6.2.2 剪枝粒度对实验结果的影响

通过实验我们发现,剪枝粒度越大(k 值越小),随机游走过程中,达到稳定前迭代的次数就越多. 从图4中可以看出,随着 k 值的增大. 迭代数目明显减小. 这是因为,剪枝粒度的减小,图 G 中边的数量增加,图的连通性增强,邻接矩阵 $\boldsymbol{P}$ 中每一列的方差减小,遍历到每点的概率趋于平均,因此收敛速度加快.

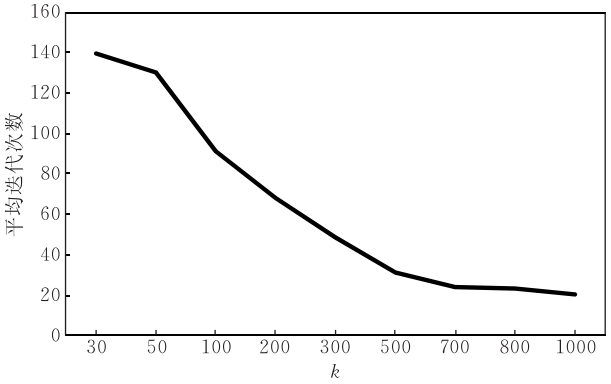

图4 剪枝粒度对迭代次数的影响

通过对比剪枝粒度对算法精度的影响,我们发现,改变剪枝粒度的大小(k 值),对实验精度的影响极小,算法的平均精度维持在69.75%±0.5%(如图5所示). 可以认为,剪枝的粒度只会改变算法的收敛速度,不会对算法的精度造成大幅度影响.

① 源码和文档可在 http://mulan.sourceforge.net/下载
② 源码和文档可在 http://www.cs.waikato.ac.nz/~ml/weka/下载

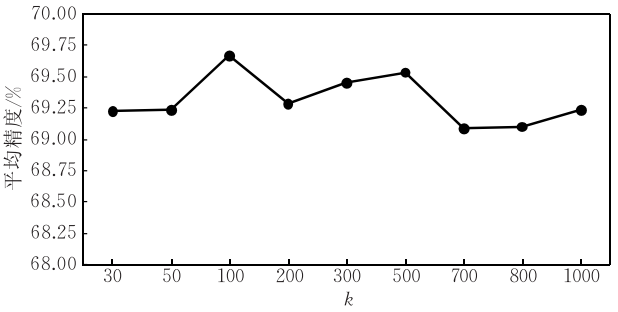

图 5　剪枝粒度对算法精度的影响

6.2.3　α 取值对实验精度的影响

由式(6)得知，α 表示游走过程中随机跳跃到某个点进行遍历的概率. α 值越大，则随机跳跃到其它点的概率就越大.

当取 $k=100$ 时，α 变化对算法精度的影响如图 6 所示. 从中可以看出，α 取值的变化对算法精度的影响比较小(不超过 $\pm 3\%$). 当算法进行迭代时，尽管随 α 的变化，实验结果也会有所变化，但是由于该数据集中标签集的势较高($c=4.20$)，图 G 的连通性较强，因此，α 取值的变化对收敛后的概率分布影响不大.

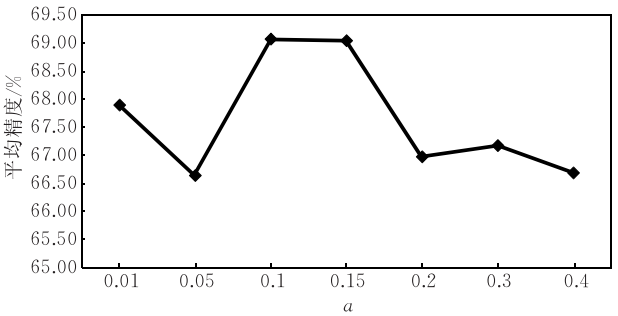

图 6　α 取值对算法精度的影响

7　结　论

本文介绍了一种基于随机游走模型的多标签分类算法 MLRW，能够有效解决多标签分类和排序问题. 今后将考虑标签数量和分布对分类结果的影响. 同时，在今后的工作中也将考虑进一步提高算法的效率.

参　考　文　献

[1] Streich A, Buhmann J. Classification of multi-labeled data: A generative approach//Proceedings of the ECML/PKDD. Antwerp, Belgium, 2008, 2: 390-405

[2] Lewis D, Yang Y, Rose T, Li F. RCV1: A new benchmark collection for text categorization research. The Journal of Machine Learning Research, 2004, 5: 361-397

[3] Veloso A, Meira Jr W, Zaki M. Calibrated lazy associative classification//Proceedings of the 23rd Brazilian Symposium on Databases. Brazil, 2008: 135-149

[4] Snoek C, Worring M, Gemert J V, Geusebroek J, Smeulders A. The challenge problem for automated detection of 101 semantic concepts in multimedia//Proceedings of the ACM Multimedia. Santa Barbara, CA, USA, 2006: 421-430

[5] Tsoumakas G. Multi-label classification. International Journal of Data Warehousing & Mining, 2007, 3(3): 1-13

[6] Shen X, Boutell M, Luo J, Brown C. Multi-label machine learning and its application to semantic scene classification//Proceedings of the 2004 International Symposium on Electronic Imaging. San Jose, California, USA, 2004: 18-22

[7] Hullermeier E, Furnkranz J, Cheng W, Brinker K. Label ranking by learning pairwise preferences. Artificial Intelligence, 2008, 172(16): 1897-1916

[8] Read J. A pruned problem transformation method for multi-label classification//Proceedings of the New Zealand Computer Science Research Student Conference. New Zealand, 2008: 143-150

[9] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification//Proceedings of the ECML. Warsaw, Poland, 2007: 406-417

[10] Schapire R, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000, 39(2): 135-168

[11] Zhang M, Zhou Z. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351

[12] Zhang M, Zhou Z. A k-nearest neighbor based algorithm for multi-label classification//Proceedings of the IEEE International Conference on Granular Computing. Beijing, China, 2005, 2: 718-721

[13] Clare A, King R. Knowledge discovery in multi-label phenotype data//Proceedings of the ECML/KDD. Freiburg, Germany, 2001: 42-53

[14] Tsoumakas G, Dimou A, Spyromitros E, Mezaris V, Kompatsiaris I, Vlahavas I. Correlation-based pruning of stacked binary relevance models for multi-label learning//Proceedings of the ECML/PKDD. Slovenia, 2009: 101

[15] Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web//Proceedings of the ASIS. Orlando, FL, 1998: 161-172

[16] Zhang L, Wu J, Zhuang Y, Zhang Y, Yang C. Review-oriented metadata enrichment: A case study//Proceedings of JCDL. Austin, TX, USA, 2009: 173-182

[17] Zhu X, Goldberg A, Van Gael J, Andrzejewski D. Improving diversity in ranking using absorbing random walks//Proceedings of the NAACL HLT. Rochester, New York, USA, 2007: 97-104

[18] Fan R, Lin C. A study on threshold selection for multi-label classification. Department of Computer Science, National Taiwan University, Taiwan, China, 2007

[19] Tang L, Rajan S, Narayanan V. Large scale multi-label classification via metalabeler//Proceedings of the WWW. Madrid, Spain, 2009: 211-220

[20] Lovász L. Random walks on graphs: A survey//Miklos D et al eds. Combinatorics, Paul Erdos is Eighty. USA: Janos Bolyai Mathematical Society, 1993

[21] Grimmett G, Stirzaker D. Probability and Random Processes. USA: Oxford University Press, 2001

[22] Elisseeff A, Weston J. Kernel methods for multi-labelled classification and categorical regression problems//Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press, 2002: 681-687

[23] Godbole S, Sarawagi S. Discriminative methods for multi-labeled classification//Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2004, 3056: 22-30

[24] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. Pearson Education, 1999: 389-392

ZHENG Wei, born in 1986, M. S. candidate. His current research interests include multi-label classification and clustering and music information retrieval.

WANG Chao-Kun, born in 1976, Ph. D., lecturer. His research interests include music data management and cloud computing.

LIU Zhang, born in 1985, Ph. D. candidate. His research interests include music data management and unstructured data management.

WANG Jian-Min, born in 1968, Ph. D., professor, Ph. D. supervisor. His research interests include data management and information system, unstructured data management in cloud environment, service processing and lifecycle management, digital rights management and security system and database testing.

Background

The traditional single-class and multi-class classification problems have been well studied in recent years. Because a data point can only be assigned to one category, both of these two kinds of problems are called the Single-Label Classification (SLC) problem, collectively. But in many practical applications, a data point can be assigned to more than one category. This kind of problem is called the Multi-Label Classification (MLC) problem. The existing of label relevant makes the MLC problems quite different from the traditional SLC problems. As a result, the classic SLC solutions cannot be directly applied to MLC problems. Generally, there are two kinds of MLC solutions. The Problem Transformation based (PT) one and the Algorithm Adaptation (AA) based one.

This paper proposes a novel multi-label classification algorithm based on the random walk model, called MLRW. Firstly, a multi-label random walk graphic system is built on the training set, by which we aim to map the multi-labeled data points into graphics. And then, the random walk model is applied on these graphs. After that, a probability distribution among all labels is obtained. Up to now, the MLRW provides an effective ranking solution to the MLC problems, but it is expected to learn a threshold to transform the ranking solution to a classification one in which the authors decide whether the unlabeled data contains each label. Thus a threshold learning algorithm is proposed based on the MLRW algorithm. With this threshold and aforementioned ranked result, MLRW provides a complete solution to the MLC problem. Experiments show that the MLRW algorithm produces an effective solution to the MLC problems.

This research is supported by the National Natural Science Foundation of China under grant No. 60803016, the National High Technology Research and Development Program (863 Program) of China under grant No. 2008AA042301 and the National Basic Research Program (973 Program) of China under grant No. 2007CB310802.