ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 2,Mar. 2017,pp. 363－374
DOI:10.16337/j.1004-9037.2017.02.018
© 2017 by Journal of Data Acquisition and Processing
http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于标记关系的多标记社团发现算法

李　娜[1,2]　潘志松[1]　施　蕾[1]　薛　胶[1,2]　任义强[3]

(1.解放军理工大学指挥信息系统学院,南京,210007; 2.中国电子科技集团公司第三十二研究所,上海,201808; 3.西门子电力自动化有限公司,南京,211100)

摘　要: 真实世界的对象具有多义性,具有非单一的多种标记。对于多标记的学习,现阶段的工作虽然能够利用标记间的重用评分分析多标记间的关系,但是尚不能直观挖掘出多标记的关系结构,也不能准确掌握多标记的主从关系以及多标记的重要性排名情况。而非负矩阵分解(Nonnegative matrix factorization,NMF)方法能对有关联的节点进行有效的社团划分,发掘关联节点的潜在关系,因此利用 NMF 方法对多标记关系进行社团结构分解成为有价值的研究内容。本文提出多标记社团发现算法,有效地对多标记进行挖掘,发现其中的社团结构,得到多标记的社团关系,并且能够对多标记节点的重要程度排序,分析多标记的主从结构,验证多标记关系算法的有效性,挖掘出其中隐藏的价值,这对于多标记的研究具有重要意义。

关键词: 多标记;标记关系;非负矩阵分解;社团发现

中图分类号: TP391　　**文献标志码:** A

Multi-label Community Detection Algorithm based on Multi-label Relationship

Li Na[1,2], Pan Zhisong[1], Shi Lei[1], Xue Jiao[1,2], Ren Yiqiang[3]

(1. College of Command Information Systems, PLA University of Science and Technology, Nanjing, 210007, China; 2. The 32th Research Institute of China Electronic Technology Group Corporation, Shanghai, 201808, China; 3. SIEMENS Power Automation Co, Ltd., Nanjing, 211100, China)

Abstract: The objects of the real world can be assigned multiple meaning, with a variety of non-single labels. As to multi-label learning, although the related current work may take advantage of the reuse score to analyze the relationship between multiple labels, it still can find neither the label structure nor the main labels and importance rankings. The nonnegative matrix factorization (NMF) method can divide associated nodes into societies effectively, and explore the potential relationship between them. Consequently, it is worth studying how to use NMF in multi-label community detection. Here, an algorithm is proposed for multi-label community detection, which can analysis labels effectively and discover the community structure inside, and then obtain relations community. Besides, these multi-label nodes can be sorted according to their importance scores, and then the master-slave structure of these marked nodes can be obtained and the effectiveness of this algorithm is thus verified, which helps us learn the hidden information.

Key words: multi-label; label relations; nonnegative matrix factorization; community detection

---

基金项目: 国家自然科学基金(61473149)资助项目。
收稿日期: 2015-03-12;**修订日期:** 2016-06-02

引 言

在监督学习框架中,每一个训练的样本都具有一个类别标记来表示其语义信息,学习的目标则是从训练样本中获得标记的概念,使得学习模型能够给未见过的样本预测正确的标记。然而生活中具有多种语义信息的学习对象非常常见,传统的监督学习总是假设真实世界的对象同它相对应的描述以及它的概念标记之间是完全一对一的关系。然而在物质世界中,各种事物和各类现象基本上可由多种语义标记表示,如在文本分析领域的论文分类问题中,每篇论文可同时属于多个研究领域;在多媒体理解领域如视频检索过程中,每段视频包含了多种语义类别;在生物信息学领域,一种蛋白质可以有多种功能的分类。随着多标记学习[1]的出现,利用相关标记上的有用信息互相帮助能减少学习任务的难度。同时,相关标记上的语义信息之间存在着一定联系,结构明显,随着社团工作的开展,多标记中的社团结构也逐渐引起人们的兴趣。社团发现[2](Communities detection)最早来源于 Girvan 和 Newman 提出的基于边介数的经典社团划分 GN 算法,起源于对社会网络的研究,并逐渐成为该领域特殊而新颖的研究方向。以往的工作很少将多标记学习与社团发现联系起来,而对于多标记的社团结构也没有过明确的表示。由于标记间的相关性越来越多地被应用到多标记学习领域,标记间的关系不断受到研究人员的重视。文献[3]指出,多标记分类的任务是基于训练数据学习一个分类器,从而在面对未知样本时,能够通过此分类来预测其他可能的样例,因此标记是可能联系在一起的,并且每个标记占的比重是不平均的。针对多标记的这些特征,若将社团发现方法作用于多标记关系上,其社团结构不仅能够验证其相关性,还能够得到标记所占的比重。

本文的多标记社团发现算法是在已有工作的基础上,根据文献[4]提出的多标记学习方法——多标记假设重用(Multi-label algorithm of hypothesis reuse, MAHR)方法,在学习过程中自动地挖掘出标记之间的关系并输出对标记关系的有效估计结果,经过重组得到多标记关系矩阵,在非负矩阵分解(Non-negative matrix factorization, NMF)算法基础上,根据标记关系矩阵得到标记的社团结构。该方法的目的是得到多标记社团结构,并根据社团结构来评估标记的重要性,得出标记重要性排名。文献[4]的 MAHR 算法基于 Boosting 框架,基分类器采用 Boosting 框架中最常见的决策树桩(Decision Stump)算法来训练,通过假设重用的方式来得到标记的促进关系与互斥关系评分,组合得到的矩阵作为标记关系矩阵。在多标记社团发现算法中,用迭代更新规则对分解矩阵进行优化求解,最终得到标记的社团关系矩阵。通过近几年的发展,在多标记学习和社团发现工作上已经积累了一些优秀的思想和成果,本文对其进行全面的分析与融合,为今后的发展提供帮助和借鉴。

1 多标记社团发现背景

1.1 多标记学习

1.1.1 多标记学习概念

随着信息社会的高度发展,各种海量数据以及信息触手可及。在物质世界中,各种事物和各类现象基本上可由具备多种属性的海量数据表示。在多标记学习框架下,每个样本由一个特征向量表示,但此样本可能同时隶属于多个类别标记,这样做的目的就是通过学习给定的多标记训练集从而能够有效预测未知样本的类别标记。然而,在传统的监督学习框架下,待学习样本是具有明确的而单一的类别标记,即每个样本只隶属于一个标记,因此难以处理真实世界的多义性对象。针对上述问题,自 20 世纪 90 年代末起,机器学习领域的研究者们不断关注如何对多标记对象有效建模,多标记学习[5-7]框架应运而生,并成为一个具有挑战性的课题。

目前已提出的多标记学习算法大致可以分成两大类[4,8,9]:问题转化方法(Problem transformation

methods,PTM)和算法适应方法(Algorithm adaptation methods,AAM)。PTM 将多标记学习问题转化为其他已知的学习问题,例如两类问题、多类问题和标记排序问题等。AAM 是直接设计多标记学习算法处理多标记数据,即改编一些著名的算法来直接处理多标记数据[10]。无论是 PTM 或是 AAM,多标记学习算法都面临着一个相同且关键的问题,即如何利用标记之间的关系[4],这对于解决维度灾难有重要的意义。随着对机器学习理论和应用的研究日益深入,多标记学习近年来已经逐渐成为机器学习和模式识别领域的热门课题,大量的公开数据集以及专门针对多标记学习的工具箱为多标记学习的深入研究提供了有力支持。

1.1.2　标记间相互关系

多标记学习区分于传统监督学习的最主要特点就是不同标记之间存在相互关系。根据标记之间的关系就能使得其中有价值信息互相帮助,提高学习效率。比如标记之间的促进和互斥关系能够有效避免不必要的训练和测试。另外在样本不足情况下,标记也可以利用其他具有足够样本的相关标记上的信息来帮助自身学习。因为,多个标记之所以同时与一个样本关联,意味着这些标记之间本身也存在着一定的相关性,于是如何利用标记关系来帮助学习是多标记学习的关键任务。

文献[4]针对以往多标记学习方法需要事先获得标记关系,在缺乏外界知识源时易过拟合的缺陷,提出了一种不需事先获得标记关系就能有效学习并能产生标记关系估计结果的 MAHR 算法。该方法通过自动重用不同标记的分类模型,通过重用权重计算出标记之间的相关值,不仅可产生强泛化能力的多标记学习器,还能输出对标记关系的估计结果,提高多标记学习系统的泛化性能[10]。

MAHR 算法基于 Boosting 框架,通过假设重用的方式来实现。它为每一个标记训练一个 Boosting 分类器,在该算法每一轮训练中,某个标记上的基分类器是由该标记上训练好的模型及其他标记上已经训练好的模型加权求和得到。在此过程中,各个标记所对应的权重可通过最小化训练集上的损失函数(Hamming loss)自动确定,因此有帮助信息标记上的模型就会被赋予较大的权重,从而能够自动地挖掘出标记之间的关系[4,9]。

1.2　社团发现

1.2.1　社团发现背景

社团发现是数据挖掘领域另一热门课题,即在网络中找出一些“稠密”的子图。在大规模网络中,所有节点之间的相互作用和相互连接不是无规律和随机的,而是形成许多社团结构,社团结构是复杂网络的基本特征之一[11]。随着互联网的迅速发展,各种社会网络成为人们生活中不可或缺的一部分。目前社会网络研究的难点主要是最有价值的信息往往隐藏在海量数据之中,通过对这些数据的分析与挖掘,可以增进对社会网络上消息传播的认识。网络的社团结构是指相互之间具有较大相似性而与网络中的其他部分有着很大不同的节点的群[12,13]。随着对各种网络关系研究的深入,很多实际网络都具有社团结构。

1.2.2　多标记中社团结构

在多标记的社团发现方面,本文基于 NMF 方法对多标记进行社团划分,从而使标记间的关联更加明确,结构一目了然,主从关系亦清晰可见,并且对于数据中隐藏含义的分析有突出贡献,并且能起到化繁为简的作用[14]。随着多标记数据集中标记数量的不断增多,多标记关系也越来越广泛地应用到多标记学习框架当中去,并且有效地提高了算法分类的精度与效率。由于多标记数据集中的标记节点众多,结构较单标记数据更为复杂,而且标记间的关系有正有负,标记间的正作用一般代表友好、信任等信息,而负作用则常常应用于否认、排斥等方面。正是由于标记间的这种作用,其正相关与负相关的信息才有了明显的区分,比如在图像分类中沙漠与海洋一般不会同时出现。然而标记间的相关性千丝万缕,文本分类中一则正面的新闻可能牵涉到一个负面消息,此时寻找标记关系的准确结构,将其中的关系直观准

确展现出来就显得尤为必要,因此需要一种方法针对多标记数据集的网络结构与特征进行分析。近来随着对海量数据分析不断深入,探索数据中的社团结构逐渐成为研究热点。在复杂网络的社团内部,节点之间的联系紧密,社团之间的联系相对比较稀疏[15],可以表示包含大量个体和个体之间相互作用的系统。然而社团结构在多标记数据集中的应用还不是很广泛,由于多标记学习中现有算法已经越来越侧重于对标记间相互关系的利用,而多标记之间的相关性亦可以表示成由标记节点形成的网络结构,因此,对于多标记的节点结构与特征分析就有了突破口,从而能够利用社团发现算法分析标记之间的隐含意义。

本文在多标记假设重用方法基础上,将大量多标记数据集样本应用该方法有效估计标记关系,并得到标记的评分关系矩阵来量化标记关系的估计结果,对标记关系进行分析与研究,并将标记关系重组,根据此结果利用基于 NMF 的多标记社团发现方法得到数据集样本中所有标记间的社团结构,度量出标记的重要程度并作重要性排序,最后验证其真实性和有效性。

2 多标记社团发现模型和节点特征分析

由于样本标记数目较大,并且标记间的关联不明显,所以目前已经有许多工作尝试挖掘和利用标记之间的关系。对于标记关系矩阵,它表示各个节点之间的相互作用关系,且这种节点间的相互作用并不是对称的[4,9]。现有的社团发现算法往往针对的是节点间有联系的 0-1 邻接矩阵,或是加权的邻接矩阵,通过计算节点的度来获得节点社团关系与重要程度。然而,对于多标记的关系矩阵,标记间存在着正负作用,这种作用就用节点的权重来表示,且这种作用并不是完全相互的。同时,多标记节点的重要性不仅与度的大小有关,而且与该节点相邻的节点特征有很大关联。本文尝试利用基于 NMF 的多标记社团发现算法挖掘多标记关系矩阵并揭示多标记的社团结构,这对于了解标记结构、分析标记关系、理解标记含义、发现标记中隐藏的规律和预测数据集的标记都尤为重要,而揭示社团结构的方法就是利用标记中所已知的特性和信息,将看似无规律的标记划分出隐藏在其中的结构。文献[15]中提出了针对有向加权节点的社团划分方法,定义节点贡献比与节点重要性排名两个节点特征指标,有效地反映标记节点的特征。本文认为,根据已知的标记关系矩阵所划分出的多标记社团结构对于多标记关系的研究更具针对性与明确性。在本文中所提到的多标记社团结构是指根据多标记社团发现算法按照某种规则划分出多标记社团关系矩阵,该结构对所给出的标记关系有着特殊意义。

2.1 多标记社团发现模型

定义 1 设多标记数据集

$$D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\} \tag{1}$$

式中:n 表示多标记数据集样本维数;$\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{im}]$表示第 i 个示例;$\boldsymbol{y}_i=[y_{i1},y_{i2},\cdots y_{im}]$为样本 $\boldsymbol{x}_i$ 的标记向量,其中 $y_{ij}\in\{-1,+1\}$,$y_{ij}=+1$ 表示 $\boldsymbol{x}_i$ 具有标记 $\boldsymbol{y}_j$,$y_{ij}=-1$ 表示 $\boldsymbol{x}_i$ 不具有标记 $\boldsymbol{y}_j$。

算法 1 多标记社团发现算法

(1) 输入:多标记训练数据集 $D^*=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_1^n$,基分类器算法 Γ,重用函数 R,训练轮数 T,多标记社团划分个数 k,迭代次数 N。

(2) 输出:多标记社团关系矩阵 $\boldsymbol{Com}$。

(3) 初始化基分类器权重 $D(i)=\dfrac{1}{n}$与多标记社团关系矩阵 $\boldsymbol{Com}_0$;训练模型 $Q_{1,l}=0$;$t=1$;定义 $\boldsymbol{H}$ 为非负矩阵分解后的节点所属社团矩阵,同时对 $\boldsymbol{Com}_0$ 的每一列数据归一化;其中 $\boldsymbol{H}$ 为非负矩阵。

(4) while $t<T$ do

利用基分类器 Γ,从数据集 D^* 中为标记 l 训练一个模型 $\hat{f}_{t,l}$;

利用重用函数 R,从 $\hat{f}_{t,l}$和全部模型候选集合 $F_{t,l}$中共同得到 y_l 的基分类器 $f_{t,l}$,并计算分类错

误率 ε 与 $f_{t,l}$ 的权重 $\beta_{t,l}$;

更新训练数据的样本权重:$D_{t+1,l}(i)=(1+\exp(y_{i,l}\sum_{j=1}^{t}\beta_{j,l}f_{j,l}(\boldsymbol{x}_i)))^{-1}$;

输出综合模型 $Z_l(\boldsymbol{x})$ 及候选集以及模型候选集 $F_{t+1,l}$;其中 $Z_l(\boldsymbol{x})=\sum_{t=1}^{T}\beta_{t,l}f_{t,l}(\boldsymbol{x})$。

(5) end while(步骤 4 详见文献[4])。

(6) 根据 $Z_l(\boldsymbol{x})$ 分别得出多标记间正作用与负作用评分 $\boldsymbol{S}_{\text{plus}}$,$\boldsymbol{S}_{\text{minus}}$。

(7) 根据评分得到关系矩阵规范化并重组,得到多标记关系矩阵 $\boldsymbol{G}$,$\boldsymbol{G}=\begin{bmatrix}\boldsymbol{R}_{\text{plus}} & \boldsymbol{R}_{\text{minus}}\\ \boldsymbol{R}_{\text{minus}} & \boldsymbol{R}_{\text{plus}}\end{bmatrix}_{2n\times 2n}$

(8) for $i=1$ to N do

更新 $\boldsymbol{H}$ 矩阵行向量:$\boldsymbol{H}_{i,j}=\dfrac{\boldsymbol{H}_{i,j}*(\boldsymbol{Com}'*\boldsymbol{G})_{i,j}}{(\boldsymbol{Com}'*\boldsymbol{Com}*\boldsymbol{H})_{i,j}}$

更新 $\boldsymbol{Com}$ 的列向量:$\boldsymbol{Com}_{n,j}=\dfrac{\boldsymbol{Com}_{n,j}*(\boldsymbol{G}*\boldsymbol{H}')_{n,j}}{\boldsymbol{Com}*\boldsymbol{H}*\boldsymbol{H}'_{n,j}}$

重新对 $\boldsymbol{Com}$ 进行列归一化

(9) end

定义 2　模型候选集合[4,9]

$$F_{t+1,l}=\{Z_{t,k}\mid k\neq l\}\cup\{-Z_{t,k}\mid k\neq l\}\tag{2}$$

式中:$Z_{t,k}$ 为第 t 轮更新标记 y_k 上的综合模型[4]。为了优化的方便,引入 $-Z_{t,k}$,表示一个输出与 $Z_{t,k}$ 正好相反的模型。最后,将由 T 个基分类器绑定得到的综合模型 $Z_{T,k}$ 作为标记 y_l 的分类器。

定义 3　重用函数(Reuse function)R[4,9]

$$R_{\omega}(\hat{f},F)=\omega(\hat{f})\cdot\hat{f}+\sum_{Z\in F}\omega(Z)\cdot Z\tag{3}$$

式中:每个模型 Z 拥有一个需要优化的权重向量 $\boldsymbol{\omega}$,表示对此模型赋予的权重。而 $\omega(Z)$ 表示向量 $\boldsymbol{\omega}$ 中与 Z 对应的元素。变量的优化可用最小化多标记学习的损失函数来实现。对于 $\boldsymbol{\omega}$ 的优化,文献[4]选择最小化损失函数,因此按照定义 3 的公式来实现。式(3)中 R 将 $\hat{f}$ 和 F 中模型绑定并生成一个新分类器。在文献[4]中,将 R 定义为线性加权的形式,使得算法获得良好的性能。

2.1.1　标记相关性

由于算法 1 在标记之间重用彼此已经得到的模型,因此根据文献[4]可知,可以用重用的多少来衡量标记之间关系的强弱,并将重用评分矩阵作为标记的关系矩阵。如果两个标记是独立的,那么不同标记上模型的预测则可能产生较大的错误率,因此算法会自动地赋予该模型一个很小的权重。相反,如果标记之间是相关的,各自的模型提供的信息对于预测其他的标记具有重要作用,那么算法就会赋予其较大的权重。于是通过标记重用的权重 $\boldsymbol{\omega}$ 来度量标记关系。两个标记间的作用有正有负,在更新权重 $\boldsymbol{\omega}$ 被约束为正的情况下,用 $\omega_{t,i}(-Z_{*,j})$ 来反映 y_j 对 y_i 的负作用。当两个标记的负作用大于正作用时,标记关系就表现为负[4]。而对于 NMF 来说,要求输入的关系矩阵为非负矩阵,因此将重用评分 $S(i,j)$ 分解为 $S_{\text{plus}}(i,j)$ 和 $S_{\text{minus}}(i,j)$,分别表示标记间的正作用和负作用。具体定义 y_j 到标记 y_i 的重用评分 $S(i,j)$ 如式(4)所示。

定义 4　重用评分

$$S(i,j)=S_{\text{plus}}(i,j)-S_{\text{minus}}(i,j)\tag{4}$$

其中

$$S_{\text{plus}}(i,j)=\sum_{t=2}^{T}\beta_{t,i}(\omega_{t,i}(Z_{t-1,j}))\tag{5}$$

$$S_{\text{minus}}(i,j)=\sum_{t=2}^{T}\beta_{t,i}(\omega_{t,i}(-Z_{t-1,j})) \tag{6}$$

同时，$S_{\text{plus}}(i,j)$和$S_{\text{minus}}(i,j)$分别为正值，且由S_{plus}和S_{minus}得到的关系矩阵$\boldsymbol{R}_{\text{plus}}$和$\boldsymbol{R}_{\text{minus}}$均为$n\times n$的方阵($n$为标记个数)。

设标记关系矩阵

$$\boldsymbol{G}=\begin{bmatrix}\boldsymbol{R}_{\text{plus}} & \boldsymbol{R}_{\text{minus}}\\ \boldsymbol{R}_{\text{minus}} & \boldsymbol{R}_{\text{plus}}\end{bmatrix}_{2n\times 2n} \tag{7}$$

式中：$\boldsymbol{G}$中的元素被约束为非负值，且$\boldsymbol{G}$为$2n\times 2n$的方阵。这样的方式既能保留所有标记之间的相互作用，又能保证关系矩阵为非负矩阵，并且没有降低关系矩阵的维数。根据算法得出的标记关系是非对称的，这在实际工作中有较大意义。

2.1.2 多标记中的非负矩阵分解

NMF是一种聚类和降维技术，具体定义如式(8)所示。

定义 5 非负矩阵分解形式

$$\boldsymbol{G}=\boldsymbol{H}*\boldsymbol{Com} \tag{8}$$

式中：矩阵$\boldsymbol{G}$,$\boldsymbol{Com}$,$\boldsymbol{H}$分别表示3个非负矩阵。$\boldsymbol{G}$中的列向量都可以用$\boldsymbol{H}$中的行向量乘以矩阵$\boldsymbol{Com}$来表示，其含义是多标记关系矩阵$\boldsymbol{G}$可用维数更少的基向量$\boldsymbol{Com}$来表示。从物理意义上讲，由于$\boldsymbol{H}$是$n\times k$维，$\boldsymbol{Com}$是$k\times n$维，可利用$\boldsymbol{H}$将原来矩阵$\boldsymbol{G}$中的n个节点聚成k个类，$\boldsymbol{H}$中每行的值表示节点对k个社团的隶属度[15,16]。

算法1可求解出多标记社团矩阵$\boldsymbol{Com}$，其中k作为输入量，表示定义社团个数，矩阵$\boldsymbol{Com}(0<j<k)$表示一个二维的社团数量为j的社团，行对应标记结点类别，列对应社团类别，矩阵$\boldsymbol{Com}$每一行最大的元素所对应的列则可标识为对应标记所在的社团类别，由此可以设计出用户归属社区的划分算法。

2.2 多标记社团节点特征分析

在社会网络中，社团中的结点由于丰富的交互性而相互扮演者不同的角色。通常，只有几个节点是整个网络的中心节点，而其他节点对网络的影响并不大。对于社交网络，顶点的度可以部分反映节点的重要性，但它不能找出社团中节点特性的差异。实际工作中，节点的重要性不仅取决于它的度，也依赖于它的邻接节点以及“朋友”关系。此外，由于网络中节点众多，这就更有必要研究社团中节点的特性。通过算法2，在计算得到节点所属社团的矩阵$\boldsymbol{Com}$中，可发现多标记节点的某些属性。

算法 2 节点特征算法流程示意图如图1所示。

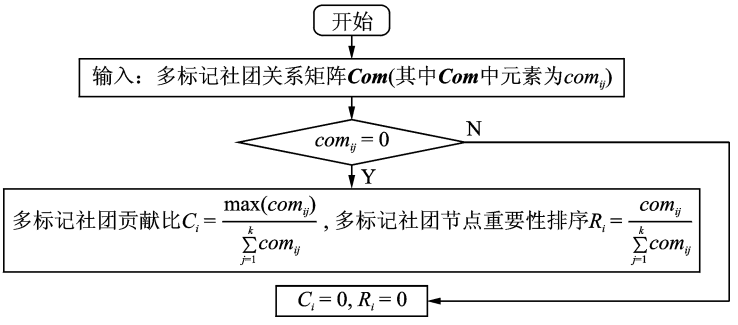


图1 多标记节点特征算法流程

Fig. 1 Process of multi-label nodes characteristics algorithm

对于网络节点社团关系矩阵 ***Com***,多标记社团的节点贡献比描述了节点对于构成各个社团所提供的支持力度。C_i 越大,说明多标记社团节点活动集中在某社团,反之则说明没有明显的社团特征。异常节点的活动相对稳定,不会集中在某一个固定的社团内,与其他节点尚未形成明显的社团特征,其 C_i 值就会比较小。而多标记社团节点重要程度排序 R_i 就是分别计算每个节点的贡献比[15],进而根据大小排序。

事实上,在社团工作中,并非所有的节点都同等重要,对于社团的建立只有少数节点发挥主要作用。传统的社交网络利用节点度的大小及中心性等概念来判别重要节点。根据节点重要性排序(R_i)可以找到社团中节点即标记的重要性。根据 R_i 结果,第一个节点在社区具有最大的 C_i,是中心节点,具有较低的 C_i 节点是边界节点,并且最低的被定义为在社区最远的节点[15]。中心节点在社团中起重要作用,而边界节点对于社团影响较小。在社团结构中,有时可以只考虑中心节点,有的边界节点就可以忽略。在多标记研究中,一些边界的标记对于多标记网络的影响微乎其微,诸如邮件标记中的无用邮件,对于分析邮件内容没有意义,因此可根据 R_i 网络评估节点的重要性。

3　实验分析

3.1　实验设置

通过算法 1 首先得到多标记的评分矩阵,其中基分类器使用决策树桩(Decision stump),训练轮数 $\boldsymbol{T}$ 为 $2*n$(n 为特征维数),进而得到多标记的社团结构矩阵。

本实验中共使用 5 个数据集,包括图像分类数据集 Image[17],Scene[18],基因功能预测数据集 Yeast[19],邮件分析数据集 Enron[20] 和文献管理数据集 BibTex[21],表 1 给出了这些数据集的统计信息,包括样本个数,标记个数,特征维度和平均每个样本的相关标记个数。其中 Bibtex 数据集在多标记关系得到过程中已经预先划分好训练集和测试集,对于其他数据集,随机选择 1 500 个样本作为多标记关系得到过程中的训练集,剩下的样本作为测试集,并将数据的随机划分重复 30 次。最终利用多标记社团发现算法挖掘 5 个数据集的多标记关系矩阵的社团结构,从而得到标记的社团关系并图示。

表 1　实验数据集

Tab. 1　Experimental data sets

数据集	样本数量	特征维数	标记数量
Image	2 000	135	5
Scene	2 407	294	6
Yeast	2 417	103	14
Enron	1 702	1 001	53
BibTex	7 395	1 836	159

3.2　多标记社团结构验证

本节列举出算法 1 中的多标记关系矩阵,并展示出多标记的社团关系图。由于数据集标记数目过多,这里只列出两个标记适中的数据集的评分矩阵。

算法 1 在 Scene 数据集上输出的标记之间关系矩阵 $\boldsymbol{G}$,其中 $\boldsymbol{G}$ 表示为正作用评分和负作用评分组合的 $2n\times 2n$ 方阵,所得的 Scene 多标记关系矩阵和社团关系如表 2 所示。

通过得到的多标记关系矩阵得知,每个社团中除了包含正作用的标记元素外,还包含着负作用的标记元素。由于标记元素是不变的,变化的只是标记之间的关系是正作用还是负作用,因此在标记社团结构中,可以同时考虑正作用和负作用的标记关系。根据文献[22],可将正作用的标记看作是朋友与信任

表 2 Scene 数据集多标记关系矩阵

Tab. 2 Multi-label relationship matrix of Scene data set

	标记	正作用						负作用					
		Beach	Sunset	Fall foliage	Field	Mountain	Urban	Beach	Sunset	Fall foliage	Field	Mountain	Urban
正作用	Beach	1.00	0.00	0.02	0.00	0.20	0.17	0.00	0.00	0.00	0.00	0.00	0.00
	Sunset	0.04	1.00	0.00	0.00	0.12	0.10	0.00	0.00	0.08	0.00	0.00	0.00
	Fall foliage	0.00	0.00	1.00	0.00	0.22	0.19	0.00	0.04	0.00	0.00	0.00	0.00
	Field	0.08	0.00	0.04	1.00	0.24	0.20	0.00	0.00	0.00	0.00	0.00	0.00
	Mountain	0.05	0.00	0.02	0.00	1.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00
	Urban	0.06	0.00	0.02	0.00	0.28	1.00	0.00	0.00	0.00	0.00	0.00	0.00
负作用	Beach	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.02	0.00	0.20	0.17
	Sunset	0.00	0.00	0.08	0.00	0.00	0.00	0.04	1.00	0.00	0.00	0.12	0.10
	Fall foliage	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.22	0.19
	Field	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.04	1.00	0.24	0.20
	Mountain	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.02	0.00	1.00	0.26
	Urban	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.02	0.00	0.28	1.00

的促进关系，而将负作用的标记看作是敌人与不信任的排斥关系。如果在一个社团中既有正作用矩阵中的标记，又存在负作用矩阵中的标记，那么可以认为负作用标记就是正作用标记的敌人，即负作用标记为正作用标记的排斥节点，而社团发现就可以看成是在同一个社团结构中找出共同的排斥标记，有同一个排斥标记的正作用标记便可以结盟形成一个社团（以社团关系矩阵最终得到的正负作用标记个数为准，正作用标记个数多，则正作用标记形成社团，负作用标记个数多，则负作用形成一个社团。一般社团中的排斥标记即负作用标记是少数），因此多标记之间相关性并不是对称的，而且标记之间具有一定的负相关性。如 Sunset（日落）与 Fall foliage（秋叶景色），因为有 sunset 的景色一般以平原居多，而不是树木多的 Fall foliage。因此若把上述标记分为两类，则得到的多标记社团划分结果如表 3 所示。

由表 3 看出，每一个社团中均包含正作用的标记和负作用的标记，如果两个标记一定不相关，那么不会在同一个社团中。如 Beach 与 mountain，有 beach 的图片信息中一般不会出现 Mountain，反而是出现 Sunset 较多，因此 Beach，Sunset 同相反作用的 Mountain 等划分为同一社团，以此也能说明 Beach 与 Sunset 相关性强，mountain，Fall foliage 相关性强，取两个社团中都是正作用的标记，于是得到最终的多标记社团划分如表 4 所示。

Scene 数据集包含 6 个标记：Beach，Sunset，Fall foliage，Field，Mountain 和 Urban。最终通过多标记关系矩阵作出的社团关系图如图 2 所示（其中横线上的数据代表归一化统一处理后的标记与标记之间相互作用的大小）。社团 1{Beach，Sunset}，社团 2{Fall foliage，Field，Mountain，Urban}，在社团 1 中，Beach 与 Sunset 有很强的相关性，而 Beach 与 Fall foliage，Field，Mountain 等一般不会同时在一幅图中出现，甚至有负作用。同时得到两个社团中最重要的标记 Beach 和 Fall foliage，说明两个标记对社团中其他标记提供的帮助最多。在 Image 数据集中这种关系也能够很好地展现，表 5 为将 Image 数据集中的标记划分为 3 个社团的结果。

表 3 Scene 数据集多标记社团结构的初步分析

Tab. 3 Preliminary analysis of multi-label community structure of Scene data set

所属社团	包含标记	
	正作用	负作用
1	Beach，Sunset	Fall foliage，Field，Mountain，Urban
2	Fall foliage，Field，Mountain，Urban	Sunset，Beach

表 4　Scene 社团包含标记及符号

Tab. 4　Labels and signs in Scene community

所属社团	包含标记	表示符号
1	Beach,Sunset	●
2	Fall foliage,Field,Mountain,Urban	■

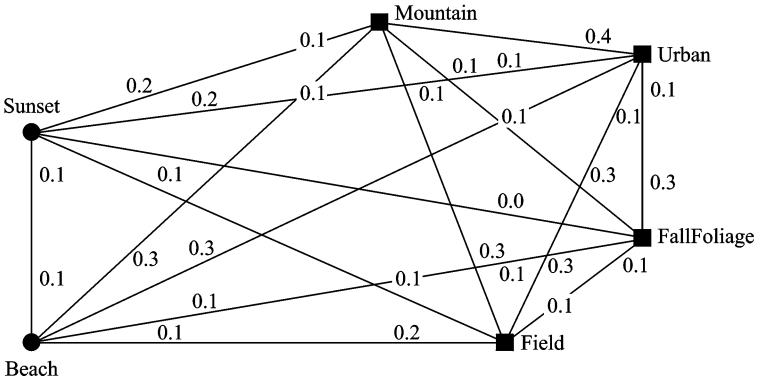


图 2　Scene 数据集的多标记社团结构

Fig. 2　Multi-label community structure of Scene data set

表 5　Image 数据集多标记社团结构的初步分析

Tab. 5　Preliminary analysis of multi-label community structure of Image data set

所属社团	包含标记	
	正作用	负作用
1	Mountain,Tree	Sunset,Desert
2	Desert	Sea
3	Sea,Sunset	Mountain,Tree

在 Image 数据集中,将多标记划分为 3 个社团,可以看到经过划分后,社团 1 中正作用的 Mountain,Tree 和负作用的 Sunset 属于一个社团,也就是在正常情况下,Mountain,Tree 和 Sunset 没有太大的相关性。同理,社团 2 中 Desert 和 Sea 的关系,以及社团 3 中 Sea,Sunset 与 Mountain 和 Tree 的关系正好印证了这一结果。同样取 3 个社团中均为正作用的标记形成最终的社团结构,结果如表 6 所示。多标记社团结构如图 3 所示(图中联系为去除冗余联系,将标记间关系二值化结果)。

Enron 数据集由安然公司员工的电子邮件组成,有 53 个类标。由于Enron数据集中标记的负作用不明显,因此可单纯分

表 6　Image 社团包含标记及符号

Tab. 6　Labels and signs in Image community

所属社团	包含标记	表示符号
1	Mountain,Tree	●
2	Desert	■
3	Sea,Sunset	▲

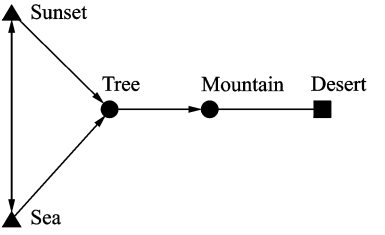


图 3　Image 数据集的多标记社团结构

Fig. 3　Multi-label community structure of Image data set

析标记间的正相关性。同时，由于 Enron 数据集标记较多，只展示排名位于前 5 的标记其重要性大小（表 7）。其相应的 Enron 社团包括标记及符号如表 8 所示（受表格大小限制，本表只展示标记序号而非标记含义）。

表 7 Enron 数据集多标记社团结构及节点重要性排名（重要性位于前 5 的标记）

Tab. 7 Multi-label community structure and importance rank (top five nodes shown) of Enron data set

所属社团	标记数量	No. 1	No. 2	No. 3	No. 4	No. 5
1	17	0.016 6	0.016 1	0.014 7	0.013 7	0.013 3
2	12	0.014 7	0.014 7	0.012 8	0.011 7	0.009 6
3	12	0.017 6	0.017 2	0.016 7	0.013 8	0.012 4
4	12	0.018 9	0.017 4	0.017 0	0.014 9	0.012 4

表 8 Enron 社团包含标记及符号（重要性位于前 5 的标记）

Tab. 8 Labels and signs in Enron community (top five nodes shown)

所属社团	包含标记					表示符号
1	17	52	27	44	21	◆
2	19	50	39	33	15	■
3	20	7	4	1	51	▲
4	45	8	11	36	49	●

由表 7，8 看出，在本数据集 53 个标记中，将其划分为 4 个社团，根据重要性排名得到属于最大的社团贡献比的标记是处于中心位置的标记，4 个社团的重要标记分别为：17-鄙视情绪；19-关注；20-感激之情；45-保密，其中，45-保密具有最高的重要性排名得分。这 4 个重要标记分别代表了 4 个社团的主要信息。其中社团 1 包括不喜欢、钦佩和公司形象等标记，代表了员工的某种情绪；社团 2 包含了关注、支持、公司业务战略等，体现了员工对公司的某种态度；社团 3 包括感激、谈话纪要、转发的邮件和友情等，体现了公司内部员工之间的联系及良好的关系；社团 4 包括保密、文档、竞争和通讯等，表现出公司最重要、最有价值和最需要谨慎对待的信息。使用表 8 中对应的图形符号表示相应社团的分类，为区分标记的重要性程度，用较大的图形符号表示社团中最重要的标记，最终得到的社团关系矩阵图如图 4 所示。

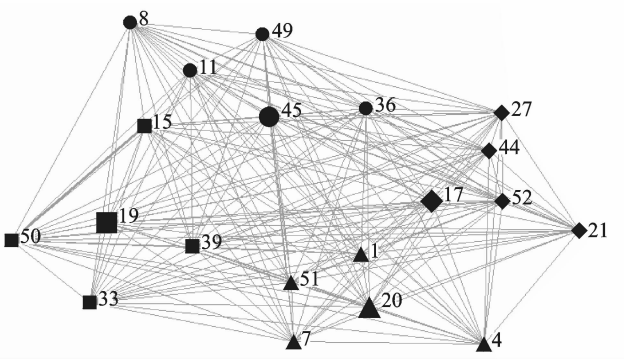


图 4 Enron 数据集的多标记社团结构

Fig. 4 Multi-label community structure of Enron data set

本文已经通过该算法得到多标记关系矩阵,通过关系矩阵可以得到标记间两两相互作用的强弱,然而,仅有相互作用的关系矩阵对于多标记关联的研究与多标记潜在含义的理解没有太大的帮助。通过本文算法得到的上述结果,能很明显知道每一封邮件所属类别与情感倾向,并区分出哪些是重要邮件,哪些是无用邮件,可以根据多标记社团发现结果对与公司业务相关的邮件进行仔细甄别,从而发现公司存在的安全漏洞与业务疏漏,对与情绪有关的邮件进行分析可发现员工的态度,最终对公司业务等奖评绩效做出调整。同时根据中心标记也能挖掘出公司近期的运作情况,员工的态度是消极还是积极性占主要地位。

通过上述的结果可以得知,利用多标记社团发现方法不仅能够得到多标记的关系,同时根据标记关系将多标记进行社团划分,经挖掘将有紧密内在关联的标记划分到同一社团,并且得到社团的贡献率以及多标记的重要性排名情况。通过得到上述结果,多标记的主从关系得以明确的展现,单个标记的重要程度也得以体现。相较于文献[4]通过 MAHR 算法单纯得到的多标记关系,本文最终得到的多标记社团结构表示的多标记关系更直观,社团中的标记关系更强烈,表示的含义更丰富,对于事例的研究更有价值。对于剩余数据集的实验结果,虽然实验得到了多标记社团结构以及标记节点特征,但是现阶段上述数据标记的具体含义仍旧未知,因此对于社团划分的结果不能够盲目分析其准确的含义,但是通过上述理论与实验分析,得到的结果可为以后的研究提供依据。

4　结束语

本文的多标记社团发现算法是在文献[4]中 MAHR 算法基础上,通过重用相关的标记上已经训练好的模型来帮助自身的学习,并通过最小化每个标记上的损失来自动优化重用的权重来自动发掘标记关系,将标记关系重组,得到多标记的关系矩阵,并基于 NMF 算法通过迭代方式来挖掘标记关系的社团结构,得到多标记的社团关系矩阵并分析节点重要性排名,得到节点的主从关系。这样的方式实现了多标记学习与社团发现工作的融合,同时得到了多标记的社团结构,为多标记关系的研究分析提供了有效的支持,并以此提高学习效率。通过大量的多标记样本实验,本文也证实了多标记社团关系的合理性与有效性。但是由于多标记数据集信息的局限性,实验部分还缺少更多的多标记数据进行验证,下阶段将寻找更多不同领域的数据进行社团分析,将本文的实验结果应用于更广泛的数据挖掘领域。

参考文献:

[1] Zhang Minling, Zhou Zhihua. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(8):1819-1837.

[2] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12):7821-7826.

[3] Gibaja E. A tutorial on multi-label learning[J]. ACM Computing Surveys, 2015, 47(3):1-38.

[4] Huang Shengjun, Yu Yang, Zhou Zhihua. Multi-label hypothesis reuse[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2012: 525-533.

[5] McCallum A. Multi-label text classification with a mixture model trained by EM[C]// AAAI'99 Workshop on Text Learning. Palo Alto, USA: AAAI, 1999: 1-7.

[6] Schapire R E, Singer Y. BoosTexter: A boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2/3): 135-168.

[7] 张敏灵. 偏标记学习研究综述[J]. 数据采集与处理, 2015, 30(1):77-87.
Zhang Minling. Research on partial label learning[J]. Journal of Data Acquisition and Processing, 2015, 30(1):77-87.

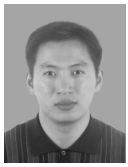
[8] Zhou Zhihua. Exploitation of label relationship in multi-label learning[C]// Proceedings of the 2012 IEEE International Conference on Granular Computing (GrC-2012). New Orleans, LA, USA: IEEE Computer Society, 2012:19.

[9] 黄圣君. 多标记学习中标记关系利用的研究[D]. 南京:南京大学,2014.
Huang Shengjun. Research on label relationship exploitation in multi-label learning[D]. Nanjing: Nanjing University, 2014.

[10] 何志芬，杨明，刘会东. 多标记分类和标记相关性的联合学习[J]. 软件学报，2014,25(9):1967-1981.
He Zhifen, Yang Ming, Liu Huidong. Joint learning of multi-label classification and label correlations[J]. Journal of Software, 2014, 25(9):1967-1981.
[11] Newman M E J. Detecting community structure in networks[J]. The European Physical Journal B, 2004,38(2):321-330.
[12] 汪小帆. 复杂网络理论及其应用[M]. 北京:清华大学出版社，2006.
Wang Xiaofan. Complex network theory and its application[M]. Beijing: Tsinghua University Press, 2006.
[13] 汪小帆，刘亚冰. 复杂网络中的社团结构算法综述[J]. 电子科技大学学报，2009，38(5)：537-543.
Wang Xiaofan, Liu Yabing. Overview of algorithms for detecting community structure in complex networks[J]. Journal of University of Electronic Science and Technology of China, 2009, 38(5): 537-543.
[14] 李乐，章毓晋. 非负矩阵分解算法综述[J]. 电子学报，2008，36(4)：737-743.
Li Le, Zhang Yujin. A survey on algorithms of non-negative matrix factorization[J]. Acta Electronica Sinica, 2008, 36(4): 737-743.
[15] Li Guopeng, Pan Zhisong, Xiao Bo, et al. Community discovery and importance analysis in social network[J]. Intelligent Data Analysis, 2014, 18(3):495-510.
[16] 张梁梁，潘志松，李国鹏，等. 基于小波去噪的有向加权社团发现研究[J]. 数据采集与处理，2014，29(5):833-839.
Zhang Liangliang, Pan Zhisong, Li Guopeng, et al. Weighted directed community detection method based on wavelet denoising[J]. Journal of Data Acquisition and Processing, 2014,29(5):833-839.
[17] Zhang Minling, Zhou Zhihua. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7):2038-2048.
[18] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9):1757-1771.
[19] Elisseeff A, Weston J. A kernel method for multi-labeled classification[J]. Advances in Neural Information Processing Systems, 2001,14(2):681-687.
[20] Klimt B, Yang Y. Introducing the Enron corpus[C]// CEAS 2004-First Conference on Email and Anti-Spam, Mountain View. California, USA:[s. n.], 2004: 30-31.
[21] Katakis I, Tsoumakas G, Vlahavas I. Multilabel text classification for automated tag suggestion[J]. ECML PKDD Discovery Challenge, 2008, 75(1):83-91.
[22] Fortunato S, Barthélemy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41.

作者简介：

李娜(1990-)，女，硕士研究生，研究方向：模式识别与机器学习，E-mail：idafighting@163.com。



潘志松(1973-)，男，教授，博士生导师，研究方向：模式识别与机器学习。

施蕾(1981-)，女，博士，讲师，研究方向：计算机软件理论、数据挖掘。

薛胶(1990-)，男，硕士研究生，研究方向：模式识别与机器学习。

任义强(1987-)，男，工程师，研究方向：电力继电保护及SCADA。