融合知识迁移学习的微博社团检测模型构建

刘宇廷,倪颖杰
(江南计算技术研究所,江苏 无锡 214083)

摘 要:传统社团检测算法大多基于网络拓扑结构,没有充分利用网络节点的标签等信息,所以无法合理地解释得到的社团结构。微博、Facebook、Twitter 等社交媒体网络增长迅速,用户标签通常不完整,应用传统机器学习模型补全标签通常需要大量训练样本,这种模式需要人工标注训练数据,时间周期长、泛化能力差。将迁移学习理论应用到这类任务中,可以避免人工标注损耗、缩短训练时间,所以针对新浪微博数据特点,提出一种融合知识迁移学习的微博社团结构检测模型(community structure inference model with knowledge transfer learning,KTL-CSIM)。社团结构检测模型基于度数相关的随机块模型,建立基于拓扑结构与节点信息的似然概率模型。文本向量化模型基于知识迁移模型将源领域知识迁移到目标领域微博数据上,得到目标领域文本向量。这种方法不需要人工标注数据,有效减少了模型训练时间,提高了泛化能力。

关键词:迁移学习;机器学习;社交网络;社团检测
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2018)09-0011-05
doi:10.3969/j.issn.1673-629X.2018.09.003

Construction of Weibo Community Detection Model with Knowledge Transfer Learning

LIU Yu-ting,NI Ying-jie
(Jiangnan Institute of Computing Technology,Wuxi 214083,China)

Abstract:Most of the traditional community detection algorithms are based on the network topological structure,and do not make full use of node information,so it is impossible to interpret the community structure reasonable. Social media network like Sina Weibo,Facebook and Twitter grow rapidly,but the user tag is usually incomplete. The application of traditional machine learning model to complete the tag usually requires a large number of training samples,which requires manual labeling of training data with a long time period and poor generalization. And applying the theory of transfer learning to such tasks can not only avoid the manual tagging of data,but also shorten the training time. Therefore,based on the characteristics of Sina Weibo data,we propose a new method of community structure inference model with knowledge transfer learning (KTL-CSIM). The model is based on degree correlated stochastic block model. Likelihood probability model is build upon network topology and node information. Word embedding model based on knowledge transfer learning model transfers the knowledge from the source domain to the target domain. This method does not require manual tagging data,which effectively reduces the training time and improves the generalization as well.

Key words:transfer learning;machine learning;social networks;community detection

0 引 言

传统的社团划分方法是一种无监督的划分方法,仅仅根据网络的拓扑结构信息,即节点和节点间的连接关系来发现网络中的社团结构[1]。现实中,有些网络可能是不精确的或者不完全的,而这些传统方法划分社团的准确度依赖于网络的精确度,因此当网络中含有噪音时会迅速降低它们发现社团的准确度。在实际应用中,往往能获取结构信息之外的数据,例如科学家合作网络中已知某个科学家是属于某个小团体,蛋白质与蛋白质相互作用网络中已知某个蛋白质属于某个功能模块等等。这些信息可以直接或间接地获取部分关于社团结构的先验信息,包括节点标签和节点对约束条件等。

传统社团检测方法可以划分为基于启发式函数算法和基于统计模型的方法。基于启发式算法的核心是定义一个目标函数如模块度等,这些算法通常衡量网

收稿日期:2017-09-07 **修回日期**:2018-01-12 **网络出版时间**:2018-05-16
基金项目:国家自然科学基金(91430214);国家"核高基"重大专项项目(2015ZX01040-201)
作者简介:刘宇廷(1987-),男,硕士生,研究方向为机器学习;倪颖杰,博士,硕导,研究方向为机器学习。

络分割优良程度,目标函数的功能是最终找到一个最优的分割方案。基于模型的方法由构建一个统计模型开始,生成用来研究的网络,然后提出一个统计检测工具来学习得到潜在的社团。当前比较流行的模型包括随机块模型、度数相关的随机块模型和混合成员的随机块模型。近年来,许多学者也在研究基于这些模型进行社团检测的理论性能。例如,在随机块模型上使用基于相似度的方法[2],使用谱聚类的方法等等。Newman 提出一种基于随机块模型的统计检测方法[1],通过定义节点元数据归属不同社团的边缘概率分布,得到基于网络拓扑与节点元数据的似然概率模型,通过 EM 算法求解方程得到最优解。该方法能够自动判断节点元数据是否有利于社团划分,但是也存在明显的缺点。算法中使用的节点元数据为独立属性(如性别、种族、年级)或组合属性,这些属性单一,对社交网络这种复杂网络的划分其实作用并不明显;EM 算法在求解最优解时容易产生局部最优解。所以,Newman 等在实际应用中取多次运行的最大值作为最优解,这种方式并没有完全解决如何得到全局最优解的问题。

在传统的机器学习框架下,学习的任务就是在给定充分训练数据的基础上学习一个分类或聚类模型,然后利用这个学习到的模型对测试网络进行分类或聚类与预测。然而,对于社交媒体中新出现的数据,训练样本标签非常稀缺。随着社交网络的规模不断扩大,给训练数据打标签工作将会耗费大量的人力与物力。而没有大量的标注数据,会使很多与学习相关的研究与应用无法开展。其次,传统的机器学习假设训练数据与测试数据服从相同的数据分布。然而,在许多情况下,这种同分布假设并不满足。如线下的训练数据与线上预测数据存在差异,这往往需要去重新标注大量的训练数据以满足训练的需要。从另一个角度看,如果有大量的、在不同分布下的训练数据,完全丢弃这些数据也是非常浪费的。所以,为了解决这些问题,文中提出一种基于文本知识迁移的模型,能够利用新闻数据训练的向量模型迁移到微博领域进行社团检测,以减轻文本向量化模型训练带来的资源消耗。

1 相关工作

1.1 社团结构检测模型

近年来,越来越多的研究人员对不同类型模型在社团检测方面的理论性能进行了研究。对于随机块模型,提出了许多基于似然概率的方法,包括最大似然函数[2]、子集似然函数[3]、伪似然函数[4]和变形推导[5]等。另有一些方法,如基于谱分析[6-8],通过谱聚类进行块模型的社团检测,谱聚类和张量谱分析方法用来检测混合成员模型中重叠社团结构。除此之外,一部分研究基于构建凸函数用来保证社团检测的可靠性[9-11],同时还有基于极大极小值框架的社团检测方法[12]。

针对传统算法无法利用节点元数据的问题,许多学者进行了尝试与探索。基于启发函数的算法[13-15],M. E. J. Newman 提出一种基于节点元数据的社团检测算法,能够自动判断元数据能否帮助网络划分,使得到的网络划分评价分数更高[1];Leto Peel 等提出在使用元数据作为网络分割指标时应该小心应对,并以 Karate 俱乐部作为例子说明元数据与社团之间的关系[16]。

Newman 在文献[1]中提出了一种基于随机块模型的方法,建立了一个统计概率模型,能够自动判断节点元数据是否有利于社团划分。其主要思想是:在随机块模型上生成的网络中定义一个邻接矩阵 $\boldsymbol{A}$,在给定参数与元数据的情况下,由模型生成的网络的似然概率为:

$$P(\boldsymbol{A}\mid\boldsymbol{\Theta},\boldsymbol{\Gamma},x)=\sum_{s}P(\boldsymbol{A}\mid\boldsymbol{\Theta},s)P(s\mid\boldsymbol{\Gamma},x)=\sum_{s}\prod_{u<v}p_{uv}^{a_{uv}}(1-p_{uv})1-a_{uv}\prod_{u}\gamma_{s_u},x_u \tag{1}$$

其中,$\boldsymbol{\Theta}$ 为 $k*k$ 矩阵,元素是 θ_{st},为算法中的混合参数;$\boldsymbol{\Gamma}$ 为元数据分配到不同社团的矩阵,元素是 γ_{s_u}。

通过对方程 1 应用 Jensen 不等式,可以转而计算社团分配的完全分布 $q(s)$。

$$q(s)=\frac{P(\boldsymbol{A}\mid\boldsymbol{\Theta},s)P(s\mid\boldsymbol{\Gamma},x)}{\sum_{s}P(\boldsymbol{A}\mid\boldsymbol{\Theta},s)P(s\mid\boldsymbol{\Gamma},x)}=P(s\mid\boldsymbol{A},\boldsymbol{\Theta},\boldsymbol{\Gamma},x) \tag{2}$$

通常 EM 算法分为两个部分,即 E(期望)阶段和 M(极大值)阶段。由于在这个算法中求完全分布 $q(s)$ 与网络大小有直接关系,所以 Newman 使用 BP(置信传播)算法进行求解,目标是得到最优的参数 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Gamma}$。

1.2 迁移学习模型

当前,迁移学习算法以及相关理论研究受到了广泛关注。迁移学习是运用已有的知识对不同但相关领域问题进行求解的一种新的机器学习方法。它放宽了传统机器学习中的两个基本假设:用于学习的训练样本与新的测试样本满足独立同分布的条件;必须有足够可利用的训练样本才能学习得到一个好的分类模型。按照源领域和目标领域中是否有标签样本,迁移学习划分为 3 类[17]:目标领域中有少量标注样本的归纳迁移学习[18-20]、只有源领域中有标注样本的直推式

迁移学习以及源领域和目标领域都没有标注样本的无监督迁移学习。针对源领域中有大量已标注样本,而目标领域中有少量已标注样本,典型的模型有文献[21-23]。Harel 和 Mannor 采用谱方法并且结合类信息将源数据和目标数据建立对应关系,对源领域的数据进行重新表示,然后训练分类器,以便对目标领域中的数据进行分类。Wang 等[24]结合类别信息和特征信息建立了两个相似度矩阵和一个不相似度矩阵,然后利用谱分解的方法学到源领域和目标领域到一个新的低维空间的映射关系。新的空间里面保持了数据点原有的类别和特征信息。所有数据点通过新的表示之后,可以进行传统机器学习的方法完成聚类任务。Duan 等[25]借用支持向量机的思想,结合类信息学习,把源领域和目标领域中的数据都映射到一个新的低维空间,然后,结合各自空间的特征组成新的特征表示。

微博数据高维、低质体现在网络中用户发布的信息通常不规范,口语化,包含错别字和新生词。这给特征词的抽取带来了很大的困难,所以对于这些文本信息分析的模型需要仔细研究与设计。当前,研究的热点方向是一些突发事件。比如,赵华等基于新浪微博数据研究 H7N9 传染病话题模型[26],Ting Hua 等基于 Twitter 数据分析佛吉尼亚枪击事件[27]。这些研究通过迁移学习理论学习用户的标签,应用分类或者聚类方法得到结果。Ting Hua 在分析目标主题(如疾病爆发、犯罪事件或者大众情感)时,提出一种半监督目标兴趣事件检测 STED 方法,通过建立标签提取模型,将新闻标签迁移到 Twitter,产生初始标签数据,标签传播模型产生特征扩展标签数据,图分区模型将一个 tweets 分为多个子集,再使用 SVM 分类模型识别与目标主题相关的 tweets。

针对新浪微博用户数据的特点,文中改进了 Newman 提出的社团结构检测模型,应用知识迁移学习构造用户标签特征向量,提出一种融合知识迁移的社团结构检测模型。该模型基于新浪微博用户数据集,通过分词模型和标签生成模型对数据进行预处理,提取用户标签,使用新闻数据集上训练的模型作为源领域模型,通过应用知识迁移学习模型,构造微博用户标签特征向量,最后使用社团结构检测模型检测社团结构。文中方法进行的改进如下:改进 Newman 的社团结构检测模型;构建由源领域向量目标领域迁移的学习模型;通过 tf * idf 模型扩展用户特征向量。

2　框架与方法

2.1　基于随机块模型的社团结构检测模型构建

随机块模型在给定参数与节点元数据的条件下,可定义为有无向网络 $G(V,E,X)$,其中 V 为节点集合,标签 $u \in \mathbb{N}_n$,E 为节点间边的集合,X 为节点信息集合,$x_u \in X$,由 k 个社区分割,节点 u 属于某个社区表示为 $s_u \in \mathbb{N}_k$。基准网络中节点元数据来自不同的高斯模型,可以使用高斯混合模型求解元数据 X 的边缘概率分布。针对复杂网络节点具有高维元数据的情况,可以定义:对于所有节点元数据 $X = \{x_1, x_2, \cdots, x_n\}$,假设每个节点 u 分配到社团 s 的概率基于节点 u 的元数据 x_u,其中 $x_u \in R^u$,$n \geqslant 1$,定义概率为 γ_{sx},使用高斯混合模型对社团归属建模,得到网络中每个节点社团分配的先验概率模型为:

$$P(s \mid \pi, \mu, \sigma, x) = \prod_i \gamma_{s_i, x_i} = \prod_i \frac{\pi_{s_i} N(x_i \mid \mu_{s_i}, \delta_{s_i})}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \delta_j)} \tag{3}$$

由随机块模型可知,基于网络拓扑结构的先验概率模型为:

$$P(\boldsymbol{A} \mid \boldsymbol{\Theta}, s) = \prod_{u<v} P_{uv}^{a_{uv}} (1 - p_{uv})^{1-a_{uv}} \tag{4}$$

基于这两个模型生成网络的似然概率为:

$$P(A \mid \boldsymbol{\Theta}, \pi, \mu, \sigma, x) = \sum_s P(A \mid \boldsymbol{\Theta}, s) P(s \mid \pi, \mu, \sigma, x) \tag{5}$$

通常可以使用 EM 算法进行求解,得到:

$$\log P(\boldsymbol{A} \mid \boldsymbol{\Theta}, \pi, \mu, \sigma, x) = \sum_s^K \log P(\boldsymbol{A} \mid \boldsymbol{\Theta}, s) P(s \mid \pi, \mu, \sigma, x) \geqslant \sum_s^K \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(\boldsymbol{A} \mid \boldsymbol{\Theta}, s) P(s \mid \pi, \mu, \sigma, x)}{Q_i(z^{(i)})} \tag{6}$$

所以,EM 算法的步骤如下:

循环重复直到收敛 {

(E 步)对于每一个 i,计算

$$Q_i(z) = \frac{P(\boldsymbol{A} \mid \boldsymbol{\Theta}, s) P(s \mid \pi, \mu, \sigma, x)}{\sum_s P(\boldsymbol{A} \mid \boldsymbol{\Theta}, s) P(s \mid \pi, \mu, \sigma, x)}$$

(M 步)更新 $\boldsymbol{\Theta}, \pi, \mu, \sigma, x$

}

2.2　知识迁移学习模型构建

领域定义:领域 D 由特征空间 χ 和边缘概率分布 $p(x)$ 组成,其中 $X = \{x_1, x_2, \cdots, x_n\} \in \chi$。$\chi$ 是所有词向量的空间,x_i 是第 i 个词向量,对应于一些文本,X 是一个特定的学习样本。一般地,如果两个领域不同,则它们或许具有不同的特征空间或不同的边缘概率分布。

对于给定的领域 D,一个分类任务 T 由类别空间 y 和目标预测函数 $f(x)$ 组成。目标预测函数无法通过观察得到,需要通过学习训练数据获得,训练数据由 $\{x_i, y_i\}$ 组成,其中 $x_i \in x, y_i \in y$。

迁移学习定义:给定源领域 D_s 和学习任务 T_s,目标领域 D_T 和学习任务 T_T,迁移学习旨在使用 D_s 和 T_s 中的知识,帮助提高目标领域 D_T 中预测函数 $f_T(x)$ 的学习,其中 $D_S \neq D_T$,或者 $T_S \neq T_T$。

任务是通过微博文本特征向量判断用户的社团信息。考虑到当前针对社团信息的分类学习尚无多见,不妨将这个问题转化为如何由一个更全的知识库来得到词语向量,这个词语向量能够反映出词语在不同句子、段落中的语义信息,所以源领域与目标领域可以不需要标注信息,从而可以将迁移学习任务定义为无监督迁移学习。

选择新闻文本数据集作为源领域 D_s,新闻文本数据量要远大于微博文本数据量,通常句式比较规范,很少出现错用词语的情况。通过学习任务 T_s 得到词语向量,能够较为真实地反映词语之间的关系。学习任务 T_s 的实现方法可参考文献[28]。

微博文本作为目标领域 D_T,通过学习任务 T_T 得到词语向量,定义迁移模型为:

$$x_{T_i} = x_{S_i} \tag{7}$$

$$a_{T_i} = \frac{1}{N}\sum_{N}\beta_i x_{T_i} \tag{8}$$

其中,$x_{S_i} \in X$ 为源领域词向量;$x_{T_i} \in X$ 为目标领域词向量;β_i 为权重调整系数;a_{T_i} 为目标领域用户 k 的加权平均向量;N 为微博用户标签数量。

3 实验及结果分析

实验数据来自新浪微博,包含 3 200 个微博用户共 237 801 条微博,319 565 条边。

通常对微博用户数据进行分词后,一般需要进行标签过滤,去掉标点符号与停用词,由于符号(@,#)与社交联系相关,所以在分词时进行了预处理,没有进行切分,部分社交联系词如图 1 所示。通过标签提取模型,共提取出 61 372 个标签。训练集中用户社团分布如图 2 所示。

@上海	@天津	@浙江卫视	@东方卫视	@格瓦拉	@腾讯
@百度云	@百思不得	@北京	@土豆网	@搜狐新闻	@爱奇艺

图 1 部分社交联系词

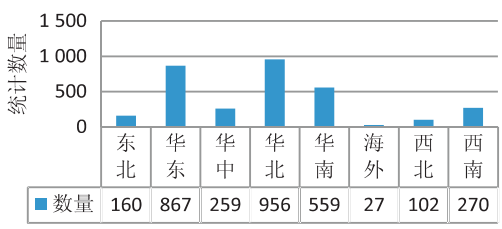

图 2 训练集用户社团分布

微博用户网络参数如表 1 所示。

表 1 基准网络参数清单

参数	数值
K	8
N	3 200
D	[300,61 672]
平均度	3.3

应用知识迁移模型可以得到微博用户文本特征向量 $w_i \in R^{300}$,应用 EM 算法求解社团检测模型最优参数得到社团分布,准确率为 30%。通过标签提取模型进一步扩展用户特征,得到微博用户文本特征向量 $w_i \in R^{61\,672}$,准确率达到 41%。

由结果可知,预测结果性能较低,主要原因是微博用户发布的文本内容分散,有价值的信息不多,所以在特征工程上,仍然缺少对问题有帮助的特征;文中使用的知识迁移模型仍然比较粗糙,需要进一步改进。

4 结束语

针对微博数据快速变化、人工标注费时费力的特点,提出一种利用已有的新闻领域模型进行知识迁移,得到微博用户的特征向量,从而进行社团检测的方法。通过改进基于块模型的社团结构检测模型,能够同时利用网络拓扑结构与节点信息进行社团结构检测,将网络节点特征扩展到多维,更符合现实世界网络情况。

该方法在应用过程中仍然存在一些问题需要解决,比如微博用户发微博通常比较口语化,经常出现新词和错字。对于这类问题,文中没有提出合理的解决方案。通常机器学习任务需要具体问题具体分析,所以针对社团检测任务,仍然需要大量的时间分析数据标签与社团的关系,文中没有提出解决标签与任务相关度的方法。在应用迁移学习理论中,非常容易出现“负迁移”的情况,文中进行知识迁移的出发点是词语在不同语句中表达含义相似,但是也存在一词多义的情况,这种情况并没有进行测试与研究。这些都是下一步需要研究的方向。

参考文献:

[1] NEWMAN M E, CLAUSET A. Structure and inference in annotated networks[J]. Nature Communications, 2016, 7: 11863.

[2] CELISSE A, DAUDIN J J, PIERRE L. Consistency of maximum-likelihood and variational estimators in the stochastic block model[J]. Electronic Journal of Statistics, 2012, 6(1): 1847-1899.

[3] BICKEL P J, CHEN A. A nonparametric view of network models and Newman-Girvan and other modularities[J]. Proceedings of the National Academy of Sciences of the United

States of America,2009,106(50):21068-21073.
[4] AMINI A A,CHEN Aiyou,BICKEL P J,et al. Pseudo-likelihood methods for community detection in large sparse networks[J]. Annals of Statistics,2013,41(4):2097-2122.
[5] BICKEL P, CHOI D, CHANG Xiangyu,et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels[J]. Annals of Statistics, 2013,41(4):1922-1943.
[6] ZHAO Yunpeng, LEVINA E, ZHU Ji. On consistency of community detection in networks under degree-corrected stochastic block models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(6):1134-1148.
[7] ROHE K,CHATTERJEE S,YU B. Spectral clustering and the high-dimensional stochastic blockmodel[J]. Annals of Statistics,2010,39(4):1878-1915.
[8] JIN Jiashun. Fast community detection by SCORE[J]. Annals of Statistics,2012,43(2):672-674.
[9] ANANDKUMAR A,GE Rong,HSU D,et al. A tensor approach to learning mixed membership community models [J]. Journal of Machine Learning Research,2013(1):2239-2312.
[10] CHEN Yudong,SANGHAVI S,XU Huan. Clustering sparse graphs[C]//Proceedings of NIPS. [s.l.]:[s.n.],2012.
[11] CAI T T,LI Xiaodong. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes[J]. Annals of Statistics,2014,43(3):5-24.
[12] GUÉDON O, VERSHYNIN R. Community detection in sparse networks via Grothendieck's inequality[J]. Probability Theory & Related Fields,2016,165(3-4):1025-1049.
[13] ZHANG A Y,ZHOU H H. Minimax rates of community detection in stochastic block models[J]. Annals of Statistics, 2015,44(5):2252-2280.
[14] RUAN Yiye, FUHRY D, PARTHASARATHY S. Efficient community detection in large networks using content and links[C]//Proceedings of the 22nd international conference on world wide web. Rio de Janeiro,Brazil:ACM,2012:1089-1098.
[15] CHANG J, BLEI D M. Hierarchical relational models for document networks[J]. Annals of Applied Statistics,2010,4(1):124-150.
[16] PEEL L, LARREMORE D B, CLAUSET A. The ground truth about metadata and community detection in networks [J]. Science Advances,2017,3(5):e1602548.
[17] 庄福振,罗　平,何　清,等. 迁移学习研究进展[J]. 软件学报,2015,26(1):26-39.
[18] WEI Fengmei,ZHANG Jianpei,YAN Chu,et al. FSFP:transfer learning from long texts to the short[J]. Applied Mathematics & Information Sciences,2014,8(4):2033-2040.
[19] 梅灿华,张玉红,胡学钢,等. 一种基于最大熵模型的加权归纳迁移学习方法[J]. 计算机研究与发展,2011,48(9):1722-1728.
[20] 庄福振,罗　平,何　清,等. 基于混合正则化的无标签领域的归纳迁移学习[J]. 科学通报,2009(11):1618-1625.
[21] 孟佳娜. 迁移学习在文本分类中的应用研究[D]. 大连:大连理工大学,2011.
[22] 毕安琪,王士同. 基于 SVC 和 SVR 约束组合的迁移学习分类算法[J]. 控制与决策,2014,29(6):1021-1026.
[23] ZHOU J T,TSANG I W. Heterogeneous domain adaptation for multiple classes[C]//AISTATS. Reykjavik,Iceland:[s.n.],2014:1273-1277.
[24] WANG Chang, MAHADEVAN S. Heterogeneous domain adaptation using manifold alignment[C]//Proceedings of the international joint conference on artificial intelligence. Barcelona,Catalonia,Spain:AAAI Press,2012:1541-1546.
[25] DUAN Lixin,XU Dong,TSANG I. Learning with augmented features for heterogeneous domain adaptation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012,36(6):1134-1148.
[26] 赵　华,章成志. 中英文突发事件话题演化对比研究—以 H7N9 微博为例[J]. 情报资料工作,2016(3):5-13.
[27] WENG J,LEE B S. Event detection in Twitter[C]//International conference on weblogs and social media. Barcelona, Catalonia,Spain:[s.n.],2011:311-312.
[28] HOLLAND P W, LASKEY K B. Stochastic blockmodels: first steps[J]. Social Networks,1983,5(2):109-137.