一种改进的无偏节点标签预测方法研究

俞 刚[1] 张泉方[2]
(浙江大学软件学院 杭州 310027)[1] (浙江大学计算机科学与技术学院 杭州 310027)[2]

摘 要 在社会网络中,用户的位置和属性以及图片的标签预测等都具有广泛的应用前景。为了提高标签预测的性能,提出了一种改进的无偏节点标签预测算法。首先,对社会网络中的标签预测问题进行了形式化描述。其次,基于所有观察数据的训练目标的联合概率最大化与以这些数据为条件的单变量边缘预测值的不匹配现象,提出了一种改进的图模型训练方法。最后,通过对置信度的无偏估计,基于子图方法提出一种不包含额外标签数据的无偏算法用于模型的训练。在Twitter和Pokec数据集上的实验表明,提出的算法与相关的标签预测算法相比,其准确性和运行效率都得到了明显的提升。

关键词 社会网络,标签预测,无偏估计,图模型
中图法分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.050

Improved Unbiased Node Label Prediction Algorithm

YU Gang[1] ZHANG Quan-fang[2]
(School of Software Technology, Zhejiang University, Hangzhou 310027, China)[1]
(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)[2]

Abstract In social networks, predictions of attributes and locations of users and labels of images are extensively applied in many fields. In order to improve the performance of label prediction, this paper proposed an improved unbiased node label prediction algorithm. Firstly, we formalized the label prediction problem in social networks. Secondly, based on the mismatch of the maximization of joint likelihood of training objective under all observed labels and the single variable marginal prediction scores conditioned by the observed labels, we proposed an improved graphical model training algorithm. Finally, according to the unbiased estimation of confidence, we proposed a training model not including additional labels based on sub-graph method. Experiments on the Twitter and Pokec datasets show that, compared with related works, the proposed algorithm has better accuracy and execution efficiency while predicting labels.

Keywords Social networks, Labeling prediction, Unbiased estimation, Graphical model

在社会网络服务中,预测用户的标签具有非常广泛的应用。在社会网络的位置服务中,大约1%~2%的用户通过签到等形式对发言或评论加上了地理标签,通过这些位置信息可以预测其他用户的位置信息,从而提供与位置相关的服务[1]。通过位置预测,可以向用户提供推荐、广告、自动语言选择等基于位置的服务[2]。此外,社会网络中的部分用户提供了年龄、性别、教育背景以及兴趣爱好等画像信息,通过已有的用户画像属性可以对未知的用户属性进行预测[3]。

数据的稀疏性是所有标签预测算法面临的挑战之一。对于大多数用户来说,并没有足够的信息来推导出这些用户的相应标签。例如在Twitter中,位置预测的准确率往往在50%~60%之间[4,5],这意味着几乎一半的位置预测是错误的,而根据错误预测的位置提供的相应推荐、广告等服务也是无意义的。本文的目的是针对那些可以进行标签预测的用户,并提高这些用户标签预测的准确度。

标签预测或选择问题在社会网络中具有广泛的应用,并且随着应用领域的不同,标签预测往往表现出很高的多样性特征[6,7]。例如,社会网络中用户的连通性差异、预测标签的同质性程度以及已知标签的比例都影响着标签预测的准确性。节点标签预测算法需要将这种多样性用正确的概率来表示被预测的标签。研究人员对部分节点加标注的图进行标签预测的问题进行了广泛的研究,采用的方法包括迭代式分类[8]、随机游走[9]、张量分解[10]、标签扩张[11]、社团[12]以及共同引用[13]等。这些方法主要关注标签预测的整体准确性,而本文则主要关注那些可以预测的用户并提高这些用户的预测准确性。

本文提出了一种改进的无偏节点标签预测算法。基于所有观察数据的训练目标的联合概率最大化与以这些数据为条件的单变量边缘预测值的不匹配现象,提出了一种改进的图模型训练方法。然后,通过对置信度的无偏估计,基于子图方法提出一种不包含额外标签数据的无偏算法用于模型的训练。

1 问题描述

给定社会网络$G=(V,E)$,其中节点集合V含有的元素

到稿日期:2014-12-02 返修日期:2015-01-31 本文受国家自然科学基金(61170306),浙江省卫生厅项目(2012KYA123)资助。
俞 刚(1979—),男,硕士,工程师,主要研究方向为算法设计与网络安全、医学信息学;张泉方(1963—),男,硕士,副教授,主要研究方向为操作系统、计算机网络、安全及应用。

个数为 n，边的集合为 E。节点的子集 $O\subset\{1,\cdots,n\}$ 中每个节点的标签是已知的，集合 $R=V-O$ 中的节点的标签是未知的，令 $O=\{1,\cdots,k\}$。Y 为标签的集合，并且 $\forall u\in O$，其标签 $c_u\in Y$。令向量 $c=(c_1,\cdots c_k)$，标签的个数为 $|Y|$。

G 中的每个节点和边都有一个特征集合，该特征集合可能是用户提供的相关信息，也可能是 G 的结构属性。当节点 u 含有标签 l 时，用 $f(l,u)$ 来表示节点 u 的特征向量；当节点 u 含有标签 l 并且 v 含有标签 l' 时，边 (u,v) 的特征向量为 $f(l,l',u,v)$。

本文的研究考虑如下情况：节点的特征向量是空的，边的特征向量仅仅依赖于 l 和 l' 是否是相等的。本文的目的是对 R 中的每个用户 u 推导出一个标签 c_u 及正确预测的概率 $p_u(c_u)$。该预测任务主要分为两步，首先根据已有的标签对模型的参数进行训练，然后部署学习得到的模型对 $\forall u\in R$ 预测 (c_u,p_u)。求解该问题的算法框架如下所示。

算法 1 EM 算法

输入：G,f,O,c;

1. 初始化参数 w;
2. While w 没有收敛 do
3. E-step：调用推理函数 infer(G,w,c)预测 c_u 和 $p_u(c_u)$;
4. M-step：对 $\forall u\in R$ 和 $\forall u\in O$，根据 c_u 和向量 p_u 重新估计参数 w;
5. End while
6. 部署算法：调用推理函数 infer(G,w,c)对 $\forall u\in R$ 预测 c_u 和 $p_u(c_u)$;

2 无偏标签预测

本文首先对图模型进行改进，使所有观察数据的联合分布的训练目标最大化与在所有观察数据条件下预测得到的单变量边缘值相一致，然后对标签预测的置信度估计问题进行分析，提出了一种无偏标签集合采集方法用于置信度估计的训练。

2.1 改进的图模型训练方法

在标签预测中，所有观察数据的训练目标的联合概率最大化 $Pr(x_O=c\mid w)$ 与以这些数据为条件的单变量边缘预测值 $Pr(x_u=c_u\mid x_O=c)$ 是不匹配的。为了解决这种不匹配，本文通过最大化条件似然性来进行目标函数的训练。对于每个观测到的标签 c_i，本文最大化以所有观测到的标签为条件的条件概率，其改进后的训练目标函数为：

$$\max_w \log\prod_{i\in O}Pr(x_i=c_i\mid x_{O-i}=c_{-i},w)$$
$$=\max_w\sum_{i\in O}\log\sum_{x_R}e^{w\cdot F(c,x_R)}-\log\sum_{x_{R\cup\{i\}}}e^{w\cdot F(c_{-i},x_{R\cup\{i\}})}\quad(1)$$

其中，条件 $x_{O-i}=c_{-i}$ 表明除了第 i 个节点外所有的观测标签都是固定的。

下面详细描述如何高效地求解该目标函数。该函数是一个非凸函数，但是与联合似然性一样，函数的特征仅仅由第一项条件决定。本文应用第 1 节的 EM 算法框架来进行局部最优化的求解。

E-step：同上述算法。

M-step：由于式(1)中的条件含有 $x_{O-i}=c_{-i}$，因此该步骤是不同的。本文需要对每个观察到的节点求解期望条件似然性：

$$\sum_i w\cdot E_\mu(F(c,x_R))-\log\sum_{x_{R\cup\{i\}}}e^{w\cdot F(c_{-i},x_{R\cup\{i\}})}\quad(2)$$

其中 $E_\mu(F(c,x_R))$ 的定义为：

$$E_\mu(F(c,x_R))=\sum_{(i,j)\in E,l,l'}\mu_{ij}(l,l')f(l,l',i,j)+\sum_{i\in V,l}\mu_i(l)f(l,i)\quad(3)$$

其中，第二项为分割函数，用 $\log Z(w\mid c_{-i})$ 来表示。为了计算 $\log Z(w\mid c_{-i})$，由于条件变量对于每个 i 都不同，需要对每个观测到的节点分别进行推导。对于不同的 i 值，不能简单地重复利用已有的计算，因此在每次优化循环中需要对整个图 G 推导计算 k 次。为了减小该计算代价，本文提出了一种近似方法。

在计算 $Pr(x_i=c_i\mid x_{O-i}=c_{-i},w)$(简写为 $Pr(c_i\mid c_{-i})$)时，从图 G 中选取一个小的子图 G_i，并且根据用户指定的置信度 ε 应用子图 G_i 的特征计算 $Pr(c_i\mid c_{-i})$ 的近似概率值。令 R_i 表示 R 在子图 G_i 上的子集，$F_{G_i}(x)$ 表示 G_i 上的节点和边的特征之和，改进后的 M-step 可改写为：

$$\sum_i w\cdot E_\mu(F_{G_i}(c,x_R))-\log\sum_{x_{R_i\cup\{i\}}}e^{w\cdot F_{G_i}(c_{-i},x_{R\cup\{i\}})}\quad(4)$$

对于每个 i，由于采用子图 G_i 对 G 进行简化，因此在 M-step 中计算联合概率时计算量明显降低。然而在原始的联合概率分布中，对于任意的 G 计算 $\log Z(w\mid c_{-i})$ 是不可能的。在对参数 w 进行学习后，本文对 $\forall u\in R$ 应用推导方法计算 c_u 和 $\mu_u(c_u)$。

2.2 置信度的无偏估计

在应用改进的图模型得到模型的参数后，本节对 c_u 的置信度进行量化分析。本文提出的节点条件概率方法与联合概率方法相比可以更好地对边缘值进行刻画，然而实验表明这两种方法都不能很好地对节点标注标签。对于观测到的节点的直接邻居节点，这两种方法得到的 $\mu_u(c_u)$ 的预测值要大于真实值，并且不能在保证图中标签有效扩张的同时对上述差异进行补偿。为此，本文提出一种解耦模型 C，该模型依据图模型的输出 c_u 和 $\mu_u(c_u)$，对每个 $u\in R$ 输出一个关于 c_u 的置信度 p_u 来描述其正确性。

假设 R 的一个子集 $D\subset R$ 中节点的标签已知，并且 $D\cap O=\emptyset$，那么 C 为一个概率二元分类器。本文基于 D 应用逻辑回归模型对 C 进行训练。对于 $\forall j\in D$，如果预测值与真实值相等即 $\hat{c}_j=c_j$，那么建立一个实例并且含有标签 $y_j=1$，否则 $y_j=0$；在 j 的邻居节点中依据观测到的和预测的标签和边缘值推导特征集合 z_j。该特征包括带有标签 c_j 的邻居节点的比例、边缘值 $\mu_u(c_u)$、除了 c_j 以外含有其它标签的节点比例等等，特征的详细信息见文献[14]。文献[14]提出的方法的缺点是：为了对模型 C 进行训练，除了需要集合 O 外，还需要标签集合 D。本文的目标是最大化正确预测节点的节点个数，因而需要 O 尽可能大。如果从 O 中分割出部分集合 D，那么上述目标函数值将打折扣，从而影响模型 C 的准确性。

文本提出一种不包含额外标签数据的无偏算法用于模型 C 的训练。当在完整的图 G 上对集合 R 中的元素预测 $\hat{c}_u$ 时，O 中节点的标签被附着在 c 上，因而无法得到待预测的标签。本文采用的思想是：对每一个 $\forall i\in O$ 和 $u\in\{i\}\cup N(i)$，应用第 i 个节点条件概率训练得到的子图 G_i 与除 i 以外的所有节点的观测标签 c_{-i} 计算边缘值 $\bar{\mu}_u^i(l)=Pr(x_u=l\mid c_{-i},w)$，预测值 $\bar{c}_u^i=\arg\max_l\bar{\mu}_u^i(l)$。由于 G_i 远远小于 G，因此该方法具有非常高的效率。对 $\forall j\in O$，如果 $\hat{c}_j=c_j$，建立一个标签实例 $y_j=1$，否则 $y_j=0$；在 j 的邻居节点中依据观测到的和预测的标签及边缘值推导特征集合 z_j。细节如算法 2 所示。

算法 2 无偏标签预测算法

输入：G,O,c;

1. 初始化参数 w；根据 2.1 节所示方法对 w 进行学习；

2. 对 $\forall u \in R$ 调用推理函数 infer (G,w,c)得到 c_u 和 p_u；

3. For $i \in O$ do

4. 依据 $Pr(c_i | c_{-i})$对 G 进行剪枝得到 G_i；

5. 调用 infer (G_i,w,c_{-i})得到$\bar{c}_u^i$ 和$\bar{\mu}_u^i$($u=\{i\} \cup N(i)$)；

6. 依据$\bar{c}_i^i$,$\bar{\mu}_i^i$,$\bar{c}_{N(i)}^i$,$\bar{\mu}_{N(i)}^i$和 $c_{N(i)}$得到特征向量 z_i；

7. 如果$\bar{c}_i^i = c_i$,$y_i=1$,否则 $y_i=0$；

8. End for

9. 应用训练数据$\{(z_i,y_i): i \in O\}$得到逻辑回归模型 C；

10. For $u \in R$ do

11. 依据$\bar{c}_u$,$\bar{\mu}_u$,$\bar{c}_{N(u)}$,$\bar{\mu}_{N(u)}$和 $c_{N(u)}$得到特征向量 z_u；

12. $p_u = C(z_u)$；

13. End for

14. Return $\{(u,p_u): u \in R\}$；

3 实验结果与分析

3.1 数据集

实验采用两个公开的真实数据集 Twitter[15] 和 Pokec[16]，数据集的基本信息如表 1 所列。Twitter 数据集用于位置预测，包含 2012 年 6 月和 7 月的部分数据，大约 8200 个包含地理信息的消息。实验从中取出 350 万个用户以及这些用户经常访问的地点，并移除了那些超过 1000 个关注者的名人用户。Pokec 数据集用于属性的预测，其中每个用户包含了性别、年龄、爱好、婚姻状况、子女、职业以及教育背景等属性信息。本实验将用户的年龄分为 10 组，每组跨度为 5 岁，其目的是依据含有年龄的数据预测未知用户的年龄。

表 1 数据集的基本信息

数据集	节点个数	边个数	单向边百分比	标签个数
Twitter[15]	1071254	3863698	42.87	2113
Pokec[16]	1136049	10773722	61.24	10

3.2 实验结果

实验将本文提出的算法与迭代式分类[8]、随机游走[9]、张量分解[10]、标签扩张[11]和共同引用[13]算法进行了性能对比，主要对比了算法的预测准确性和算法的执行效率。

首先，实验对比了不同算法的位置或属性预测准确性。在对比算法的预测准确性时，采用位置或属性正确预测的百分比作为评价指标，实验结果如图 1 和图 2 所示。从这两幅图中可以看出：所有算法的预测准确性都随着训练数据的增加而提高，迭代式分类算法在进行预测时准确性最低，本文提出的算法的预测准确性最高，其余 4 种算法在进行预测时准确性相近，都处于中间位置。该实验表明了本文提出的算法在标签预测时具有更高的准确性。

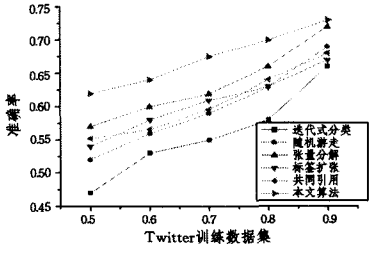

图 1 算法的预测准确性对比(Twitter)

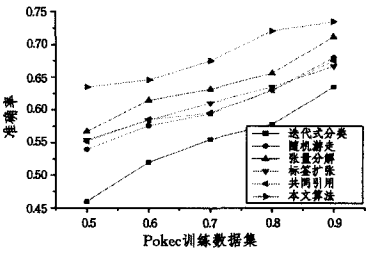

图 2 算法的预测准确性对比(Pokec)

其次，实验对比了不同算法的运行效率。在对比算法的执行效率时，实验取 50%的数据作为训练数据，分别对比了算法的训练和预测阶段所需要的运行时间。从图 3 和图 4 中可以看出：所有算法的学习时间都明显多于预测时间，张量分解算法在训练和预测两个阶段所用的时间都是最多的，而本文提出的算法的训练和预测时间在两个数据集上都是最少的。该实验表明本文提出的算法在标签预测时具有更高的执行效率。

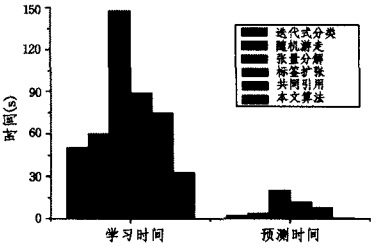

图 3 算法的运行时间对比(Twitter)

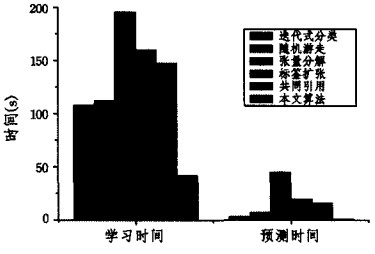

图 4 算法的运行时间对比(Pokec)

最后，实验评价了本文提出的算法中子图的大小对算法的性能影响。由于算法在 Twitter 和 Pokec 数据集上表现出相同的特征，图 5 示出了算法在 Twitter 数据集上的实验结果。图中横轴为子图包含的节点个数，大小从 5 增加到 5000，左边的纵轴为本文提出的算法的误差，右边的纵轴为算法的运行时间(包含训练和预测阶段)。从该图可以看出，算法的运行时间随着子图节点数呈近似线性地增加，而算法的误差减小率却低于线性。该实验表明，算法采用子图进行计算虽然产生了更多的误差，但却节约了更多的计算时间，从而使得算法可以在有限的时间内获得更大的准确性提升。

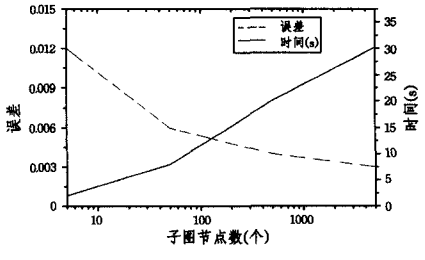

图 5 本文算法中子图大小对性能的影响(Twitter)

(下转第 259 页)

(1):105-129

[7] Wang S W, Duan Y M, Shu W N, et al. Differential evolution with elite mutation strategy[J]. Journal of computational information systems, 2013, 9(3): 855-862

[8] Sun J, Zhang Q, Tsang E. DE/EDA: A new evolutionary algorithm for global optimization[J]. Information Sciences, 2005, 169(3): 249-262

[9] Rahnamayan S, Tizhoosh H, Salama M. Opposition-based differential evolution[J]. IEEE Transactions on Evolutionary Computation, 2008, 12(1): 64-79

[10] Liu H, Cai Z, Wang Y. Hybridizing particle swarm optimization with differential evolution for constrained numerical and engineering optimization[J]. Applied Soft Computing, 2010, 10(2): 629-640

[11] Wang Y, Cai Z X, Zhang Q F. Differential evolution with composite trial vector generation strategies and control parameters [J]. IEEE Transactions on Evolutionary Computation, 2011, 15(1): 55-66

[12] Mallipeddi R, Suganthan P, Pan Q, et al. Differential evolution algorithm with ensemble of parameters and mutation strategies [J]. Applied Soft Computing, 2011, 11(2): 1679-1696

[13] Qin A, Huang V, Suganthan P. Differential evolution algorithm with strategy adaptation for global numerical optimization[J]. IEEE Transactions on Evolutionary Computation, 2009, 13(2): 398-417

[14] Yao X, Liu Y, Liu G. Evolutionary programming made faster [J]. IEEE Transactions on Evolutionary Computation, 1999, 3(2): 82-102

[15] Suganthan P, Hansen N, Liang J, et al. Problem definitions and evaluation criteria for the cec 2005 special session on real-parameter optimization[R]. Nanyang Technological University, 2005

(上接第 250 页)

结束语 标签预测在社会网络中具有广泛的应用前景。本文提出了一种改进的无偏节点标签预测算法。基于所有观察数据的训练目标的联合概率最大化与以这些数据为条件的单变量边缘预测值的不匹配现象，提出了一种改进的图模型训练方法。然后，通过对置信度的无偏估计，基于子图方法提出一种不包含额外标签数据的无偏算法用于模型的训练。Twitter 和 Pokec 数据集实验表明，与相关的标签预测算法相比，本文提出的算法在准确性和运行效率方面都得到了明显的提升。

参 考 文 献

[1] Sadilek A, Kautz H, Bigham J P. Finding your friends and following them to where you are[C]// Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012: 723-732

[2] 张子柯. 社会化标签系统的结构，演化和功能[J]. 上海理工大学学报，2011，33(5)：445-451
Zhang Zi-ke. Social tagging systems: structure, dynamics and function[J]. Journal of University of Shanghai For Science and Technology, 2011, 33(5): 445-451

[3] Weinsberg U, Bhagat S, Ioannidis S, et al. BlurMe: Inferring and obfuscating user gender based on ratings[C]// Proceedings of the Sixth ACM Conference on Recommender Systems. ACM, 2012: 195-202

[4] Hong L, Ahmed A, Gurumurthy S, et al. Discovering geographical topics in the twitter stream[C]// Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 769-778

[5] Mahmud J, Nichols J, Drews C. Where Is This Tweet From? Inferring Home Locations of Twitter Users[C]// ICWSM. 2012: 511-514

[6] 魏建良，朱庆华. 基于社会化标注的个性化推荐研究进展 [J]. 情报学报，2010，29(4)：625-633
Wei Jian-liang, Zhu Qing-hua. Advances in Personalized Information Recommendation Based on Social Tagging[J]. Journal of The China Society for Scientific and Technical Information, 2010, 29(4): 625-633

[7] 吴超，周波. 基于复杂网络的社会化标签分析[J]. 浙江大学学报(工学版)，2010，44(11)：2194-2197
Wu Chao, Zhou Bo. Complex network analysis of tag as a social network[J]. Journal of Zhejiang University (Engineering Science), 2010, 44(11): 2194-2197

[8] Sen P, Namata G, Bilgic M, et al. Collective classification in network data[J]. AI magazine, 2008, 29(3): 93-106

[9] Azran A. The rendezvous algorithm: Multiclass semi-supervised learning with markov random walks[C]// Proceedings of the 24th International Conference on Machine Learning. ACM, 2007: 49-56

[10] 廖志芳，李玲，刘丽敏，等. 三部图张量分解标签推荐算法[J]. 计算机学报，2012，35(12)：2625-2632
Liao Zhi-fang, Li Ling, Liu Li-min, et al. A Tripartite Decomposition of Tensor for Social Tagging[J]. Chinese Journal of Computers, 2012, 35(12): 2625-2632

[11] Talukdar P P, Reisinger J, Paşca M, et al. Weakly-supervised acquisition of labeled class instances using graph random walks [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 582-590

[12] 袁柳，张龙波. 基于概率主题模型的标签预测[J]. 计算机科学，2011，38(7)：175-180
Yuan Liu, Zhang Long-bo. Social Tag Predication Based on Probabilistic Topic Model[J]. Computer Science, 2011, 38(7): 175-180

[13] Bhagat S, Cormode G, Muthukrishnan S. Node classification in social networks[M]// Social network data analytics. Springer US, 2011: 115-148

[14] Chaudhari G. High confidence predictions in social networks [D]. IIT Bombay, 2013

[15] Java A, Song X, Finin T, et al. Why we twitter: understanding microblogging usage and communities[C]// Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, 2007: 56-65

[16] Bae S H, Halperin D, West J, et al. Scalable Flow-Based Community Detection for Large-Scale Network Analysis[C]// 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW). IEEE, 2013: 303-310