基于 Node2vec 的改进算法的研究

杜阳阳,李华康,李 涛
(南京邮电大学 计算机学院,江苏 南京 210046)

摘 要:针对图节点的多标签分类任务,在 Node2vec 算法的基础上进行了改进,在原来随机游走的基础上加上部分标签信息的指导,然后对节点进行向量表示。算法首先根据每一个图节点及其邻居节点的标签信息和事先设定好的游走参数的值,计算当前节点的邻居节点被游走的概率;然后由概率值和其他设定好的游走的参数开始游走,得到若干条路径;之后再调用 Word2vec 方法对若干条游走路径进行训练,将每个图节点表示成向量。最后,通过使用逻辑分类模型对节点的特征表示进行多标签分类来验证算法的有效性。实验结果证明,通过使用标签信息的指导,多标签分类的准确率有明显提升。

关键词:数据挖掘;随机游走;节点表示;多标签分类
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2018)07-0006-05
doi:10.3969/j.issn.1673-629X.2018.07.002

Research on Improved Algorithm Based on Node2vec

DU Yang-yang,LI Hua-kang,LI Tao
(School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210046,China)

Abstract:In view of the multi-label classification task of graph nodes,we carry on the improvement on the basis of theNode2vec algorithm,adding some label information to the original random walk,and then representing the nodes with vectors. The algorithm firstly calculates the probability that the neighbor node of the current node is traveling according to the label information of each graph node and its neighbor node and the value of the walking parameter set in advance. Then,the probability and other parameters begin to walk,and get a number of paths. After that,the Word2vec method is called to train a number of walking path,each node will be expressed as a vector. Finally,the validity of the algorithm is verified by using the logical classification model to label the feature representation of the nodes. The experiment shows that the accuracy of multi-label classification is remarkably improved by using the guidance of label information.

Key words:data mining;random walk;node representation;multi-label classification

0 引 言

信息网络在现实世界中无处不在,规模从几百个到几百亿个不等[1]。邻接矩阵、邻接表等是常用的数据表示方法,但在计算效率以及数据稀疏的问题上往往不能达到很好的效果。随着深度学习的发展和深入研究,图数据的表示学习为图数据的挖掘问题提供了新的思路。节点表示成向量可以应用于多个领域,如可视化[2]、节点分类[3]、链路预测[4]以及推荐[5]等。

在图数据挖掘中,评价节点表示算法效果的主要方式是对节点进行多标签分类[6]。和每个节点只有一个标签的问题相比,显然每个节点有多个标签的问题要更为复杂。目前,基于游走的图节点的表示学习算法都使用在多标签分类的结果来评价算法的有效性。

对于图节点的多标签分类问题,文中在随机游走算法的基础上考虑了标签信息的指导作用。该算法设置了一个游走参数,使游走时倾向于走和当前节点有共同标签的邻居节点,而不是随机游走,并通过实验进行验证。

1 国内外研究现状

对数据进行分析需要将数据表示成计算机能够识别的形式[7]。表示学习算法分为监督表示学习和无监督表示学习[8],其中无监督表示学习算法使用无标注的数据集,通过将输入数据变换到不同维度的向量空

收稿日期:2017-07-23 **修回日期**:2017-11-16 **网络出版时间**:2018-03-07
基金项目:国家自然科学基金(61502247,11501302,61502243,91646116);国家博士后科学基金(2016M600434);江苏省自然科学基金(BK20140895,BK20150862);江苏省博士后科研资助计划(1601128B)
作者简介:杜阳阳(1993-),女,硕士,研究方向为大数据挖掘;李华康,博士,讲师,研究方向为大数据挖掘、用户行为分析、网络空间安全等;李涛,博士,通讯作者,美国佛罗里达国际大学教授,研究方向为数据挖掘、机器学习和信息检索及生物信息学等。

间来计算输入数据的特征表示。常见算法包括局部线性嵌入[9]、独立成分分析[10]、无监督字典学习[11]以及受限玻尔兹曼机等。对于表示学习效果的评价,除了考察机器学习算法的效果外,Bengio 等[12]对各种表示学习算法进行了综述,并讨论了表示学习的目标及评价标准。刘知远等[13]系统地介绍了知识表示学习的进展和主要的表示学习算法。

图数据的表示学习为使用机器学习算法提供了可能。Chen 等[14]提出了一种根据有向图的连接结构,将有向图中的节点表示为一维向量的方法。目前,图数据表示学习算法[15]包括基于谱方法、基于最优化、基于概率生成模型以及基于深度学习的方法。其中,基于谱方法的图数据表示学习算法只考虑了网络的结构信息,难以引入网络节点的属性信息以扩展应用。基于最优化的图数据表示学习算法通常与特定的网络数据处理需求相关,通用性较差。基于概率生成模型的图数据表示学习算法要求网络节点需具有文本属性,同样存在通用性差的问题。

1.1　Deepwalk 算法

Deepwalk 算法是 Bryan Perozzi 等[16]提出的将 Word2vec 的思想用于图节点表示学习的算法。Yang 等证明 Deepwalk 算法相当于将矩阵 $\boldsymbol{M}$ 分解为两个矩阵的乘积,最终得到的节点特征向量可以由这两个矩阵进行拼接得到。Yu 等[17]使用正则化的低秩矩阵分解来得到 $\boldsymbol{M}$ 的分解矩阵。同时,根据分解矩阵的原理,Yang 等[18]将节点的属性信息纳入考虑之中,提出改进 Deepwalk 的 TADW 模型。

Deepwalk 算法在游走过程中,完全随机选择下一步游走的节点,没有一个明确的目标来指导游走,难以针对特定的学习目标来选择游走路径。

1.2　Line 算法

Line 算法[19]通过定义两种节点之间的相似性,并设计相应的目标函数,来优化节点特征的学习。该算法的提出者分别使用两种相似性指标下得到的节点向量表示,以及两种向量表示的拼接,与 Deepwalk 得到的节点向量表示作了对比实验,得到了较优的节点标签预测结果。另外,Tang 等提出的 PTE 算法[20]通过将文档、词语、标签组织成一个异构网络,以进行文档标签的预测任务,利用了 Line 算法的思想训练得到各种网络节点的向量表示,提高了预测的准确度。

1.3　Node2vec 算法

Node2vec 算法[21]是另一种对 Deepwalk 算法中的随机游走过程进行改进的算法,由 Aditya 等提出。Aditya 同样给出了两种图结构中节点相似度的评价标准,分别叫做同质性(homophily)[22]和结构对等性(structural equivalence)[23]。Node2vec 算法通过指定 p,q 两个参数,给图中的边分配权值,通过权值的大小指导游走过程,实现了指定游走是更趋向于图数据的深度优先遍历还是更趋向于图数据的广度优先遍历。通过多次尝试,在多标签分类任务上,Node2vec 取得了比 Deepwalk 和 Line 算法较好的预测效果。

2　基于标签信息的指导游走算法

2.1　算法实现

假设用 $G(V,E)$ 表示一个图,其中 V 代表节点,E 代表边[24],算法设置一个参数 p $(0<p<1)$ 来调节游走过程,$p=0$ 表示完全随机游走,$p=1$ 表示游走和当前节点有共同标签的邻居节点。下面介绍算法的具体实现过程。

设图 $G(V,E)$ 中有 N 个节点,当前游走的节点为 c,下面需要选择下一个 c 的邻居节点作为游走的节点。假设 c 节点有 T 个邻居节点,表示为:

$$\text{neighbors}(c)=\{n_1,n_2,\cdots,n_T\},0\leqslant T<N \quad (1)$$

假设 T 个邻居节点中有 K 个邻居节点和 c 节点有共同的标签,表示为:

$$\text{common}(c)=\{m_1,m_2,\cdots,m_k\},0\leqslant K\leqslant T \quad (2)$$

从上面可以看出,common(c)是 neighbors(c)的子集。若 c 节点下一个游走的节点为 d,则显然 $d\in$ neighbors(c)。设定 p 参数,使得 $p=P(d\in$ common(c)),表示 d 属于 common(c)的概率,从而得到一组新的变量 $f(n)$ 为 c 的每一个邻居节点分配游走到的概率:

$$f(n)=\begin{cases}\dfrac{1-p}{N-T}, n\notin \text{common}(c)\\ \dfrac{p}{T}, n\in \text{common}(c)\end{cases}, n\in \text{neighbors}(c) \quad (3)$$

最后将这组概率值传递给 Alias Method。

首先加载图数据,记录每一个图节点对应的邻居节点以及每一个图节点对应的标签信息,然后根据每一个图节点及其邻居节点的标签信息和事先设定好的参数 p 的值,计算当前节点的邻居节点被游走的概率,然后用 Alias Method 随机选择邻居节点,每个邻居节点被选中的概率等于计算的被游走的概率值。由概率值和其他设定好的游走的参数(如游走的路径长度等)开始游走,会得到若干条路径。之后调用 Word2vec 方法对这若干条游走路径进行训练,将每个图节点表示成向量。最后对图节点进行多标签分类,检验算法的分类效果。分类效果越好,图节点向量表示方法的有效性越好。

在第二步中,根据当前节点和邻居节点的标签信息,以及 p 参数的值,得到了每个节点被游走的概率

值。然后将这组概率值用于 Alias Method 的 alias_setup 方法,建立 alias_nodes 变量。alias_nodes 变量是一种 key-value 的数据结构,key 代表图中的所有节点,value 与该节点的邻居列表等长,是对游走概率序列进行调整之后的两个概率序列。在 Alias Method 算法的 alias_draw 方法中通过使用随机数与这两个概率序列进行比较,将返回一个下标索引,然后选取对应该下标索引的邻居节点作为下一个游走的节点。当重复多次地调用 alias_draw 方法时,返回的下标索引的概率分布将符合指定的被游走的概率值序列。为每一个邻居节点计算被游走到的概率的伪代码如算法 2 所示。游走部分的伪代码如算法 1 所示。

算法 1:probability_walk(G ,start_node,path_length, alias_nodes)

Input: Graph: $G(V,E)$, the node which path starts from: start_node,

The specified length of path: path_length

The structure of alias method: alias_nodes

Output:path_list

path_list append start_node

current_node=start_node

while the length of path_list < path_length

neighbors_list= G[current_node]

index= alias_draw(alias_nodes[current_node][0], alias_nodes[current_node][1])

next_node=neighbors_list[index]

path_list append next_node

current_node=next_node

算法 2:generate_probability(G , T , p)

input: Graph: $G(V,E)$, Tags info: T (V , tag_list), the specified percentage of tags info: p

Output:alias_nodes

for each node in$V$

neighbors_list= G[node]

Initial probability_list with zeros,its length equals to neighbors_list

for each neighbor in neighbors_list

if T[node] and T[neighbor] has common element

assign the element in probability_list accordingly to one

count+=1

for each element in probability_list

if element==1

change the element to$p$/count

else

change the element to (1-p)/(length_of_neighbors-count)

alias_nodes[node] = alias_setup(probability_list)

2.2　实验验证

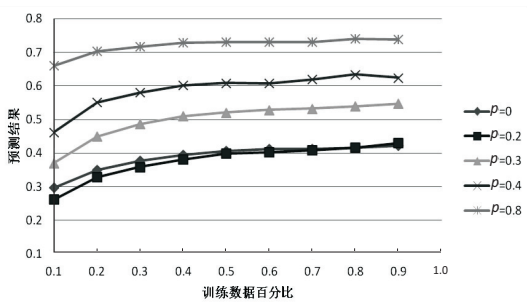
实验中使用了 blogcatalog 数据集,作为在多标签分类问题方面的数据集。在 blogcatalog 数据集中,节点代表博客作者,共包含 10 312 个节点。图中的边代表两个博客作者的社交关系。图中节点的标签是博客作者发表博客的内容类别,平均每一个博客作者包含 1.6 个标签。

在标签分类结果的评价上,使用 F_1 函数进行比较。F_1 函数是对分类结果的召回率和正确率的加权平均,计算公式表示为:

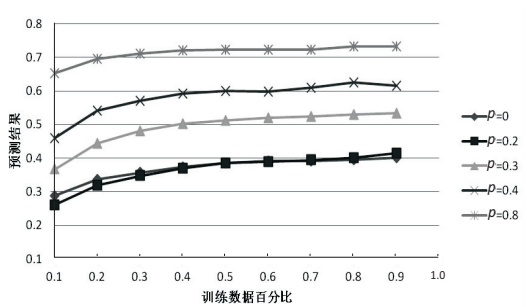
$$F_1 = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}} \tag{4}$$

而对于多标签分类问题,考虑到每一个类别标签在数量上的不平衡性,需要对每一个类别上的 F_1 函数再进行一个加权平均。对 F_1 函数的加权平均方式又可以分为“micro”、“macro”、“samples”和“weighted”四种,分别定义为“米克”、“麦克”、“赛普”以及“威特”。其中,“米克”方式从整体上统计每一个类别正确和错误预测的次数;“麦克”方式对每一个类别标签进行简单的求平均,没有考虑标签数量上的不平衡性;“赛普”和“威特”方式分别从每一次预测的角度和每一个标签的角度对预测的 F_1 结果进行了加权平均。本次实验结果的分析将用这四种方式分别进行考量。

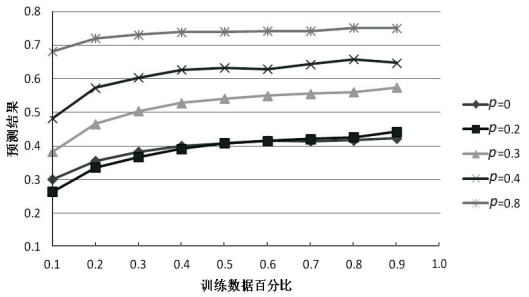
在 blogcatalog 数据集上的实验结果如图 1 所示。其中四张图分别展示了多标签分类结果使用 F_1 函数在四种评价方法-“米克”、“麦克”、“赛普”和“威特”下的预测效果。每一张图上,横轴表示在多标签分类时训练数据所占的比例,图中的五条折线分别显示了该算法中 p 参数取值 0,0.2,0.3,0.4 以及 0.8 时对应的预测结果的提升情况。当 p 取 0 时,表示在游走过程中,基于标签信息的指导作用为零,即完全的随机游走。



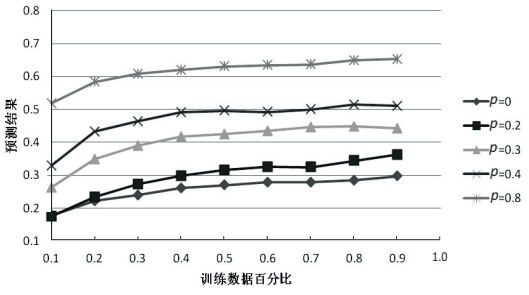
(a)米克的预测效果



(b)麦克的预测效果



(c)赛普的预测效果



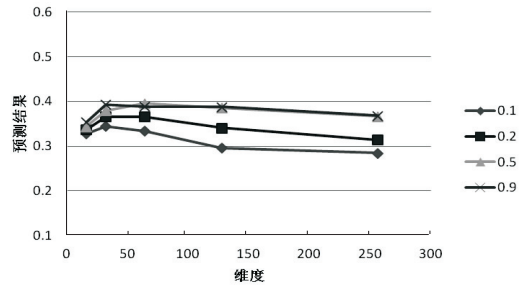
(d)威特的预测效果

图 1　在 blogcatalog 数据集上的预测结果

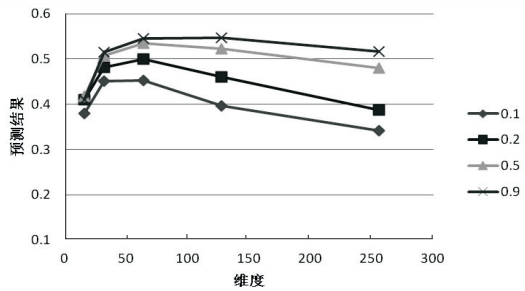
从图 1 可以看到,随着参数 p 的逐渐变大,预测效果有了明显提升。当 p 取值 0.2 时,在“麦克”和“赛普”评价方式下,其预测效果与随机游走效果基本相当。当 p =0.2 且训练数据低于 50% 时,在“米克”评价方式下,预测效果相比于随机游走的预测效果低了 0.01,表明从总体而言,指定参数 p 的取值覆盖到的标签信息的比例低于随机游走时覆盖到的标签信息的比例。但在“威特”评价方式下,预测效果相比于随机游走的预测效果已经有了较大的提高,表明从每一个标签的角度,通过较少的标签指导信息,游走过程中在兼顾遍历整个图结构的同时,算法已经开始倾向于游走与本次多标签分类任务关系更密切,特征更显著的图节点,从而在预测结果上体现出了正确率的提升。

图 2 是游走的参数(包括向量表示的维度、游走的次数)以及训练数据的比例在使用标签信息指导游走前后对标签预测结果影响的对比图。算法中的标签指导信息参数 p 取值 0.3,预测结果使用 F_1 评价标准。

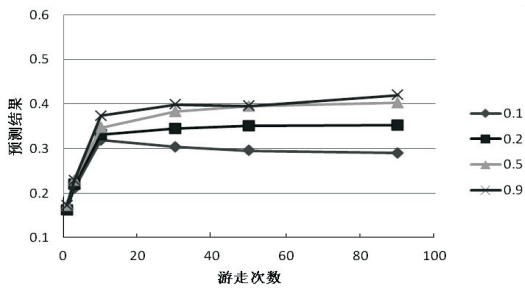
其中,图 2(a)、图 2(b)分别显示了使用指导游走前后随着向量表示维度(16, 32, 64, 128, 256)的增加,不同训练数据占比(0.1, 0.2, 0.5, 0.9)的标签预测结果。图 2(c)、图 2(d)分别显示了使用指导游走前后随着游走次数(1, 3, 10, 30, 50, 90)的增加,不同训练数据占比(0.1, 0.2, 0.5, 0.9)的标签预测结果。可以看到,随着游走次数从 1 次增加到 10 次,预测的准确率迅速提升。在使用标签指导之后,不同的训练数据比例下,预测效果均超过随机游走的预测效果。随着游走次数超过 10 次,由过多游走带来的噪声使得预测效果有所下降,如图 2(c)中训练比例 0.1 时的结果。但可以看到,当训练数据的比例增大时,这种下降得到了补偿,如图 2(c)中训练比例大于 0.2 时的结果基本持平或缓慢提升。在使用标签指导后,过多游走带来的噪声对标签特征的影响更加严重,可以看到图 2(d)中,训练比例 0.2 时预测结果仍然有所下降。因此,合适的游走次数对于标签指导游走算法的效果同样重要。



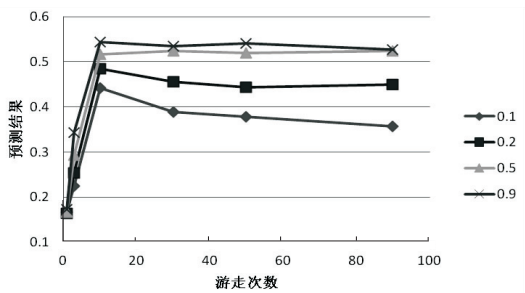
(a)指导游走前随维度变化趋势



(b)指导游走后随维度变化趋势



(c)指导游走前随游走次数变化趋势



(d)指导游走后随游走次数变化趋势

图 2　不同游走参数及训练数据比例的预测效果对比

3　结束语

针对图节点的多标签分类任务,设计了一种基于标签信息指导游走的图节点表示学习算法。该算法通

过设置一个游走参数 p，可以做到在游走过程中对有共同标签的邻居进行倾向性可调的选择，达到了在学习图节点的整个连接关系和学习节点之间的标签相似性特征的平衡。最后，实验对比了在使用标签指导游走前后以及不同的参数 p 和训练数据比例下，预测效果的变化情况。使用标签指导游走之后，预测效果提升较显著。同时，实验也对比了在使用标签指导游走前后，其他游走参数（包括游走次数、节点向量的维度和训练数据占比）对标签分类的影响情况。

在 Node2vec 的基础上考虑了标签信息，但随着机器学习算法的发展，对表示学习的特征提取有了更高的要求，节点表示的方法有待进一步研究。

参考文献：

[1] 赫 南，李德毅，淦文燕，等. 复杂网络中重要性节点发掘综述[J]. 计算机科学，2007，34(12)：1-5.

[2] VAN DER MAATEN L，HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research，2008，9(11)：2579-2605.

[3] BHAGAT S，CORMODE G，MUTHUKRISHNAN S. Node classification in social networks[M]//Social network data analytics. US：Springer，2011：115-148.

[4] LIBEN-NOWELL D，KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology，2007，58(7)：1019-1031.

[5] YU Xiao，REN Xiang，SUN Yizhou，et al. Personalized entity recommendation：a heterogeneous information network approach[C]//Proceedings of the 7th ACM international conference on web search and data mining. New York，NY，USA：ACM，2014：283-292.

[6] 郑 伟，王朝坤，刘 璋，等. 一种基于随机游走模型的多标签分类算法[J]. 计算机学报，2010，33(8)：1418-1426.

[7] 马 强. 浅谈自动识别技术的发展[J]. 合作经济与科技，2012(16)：128.

[8] 李 凯，陈新勇. 基于核策略的半监督学习方法[J]. 计算机工程，2009，35(15)：170-172.

[9] ROWEIS S T，SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science，2000，290(5500)：2323-2326.

[10] HYVÄRINEN A，OJA E. Independent component analysis：algorithms and applications[J]. Neural Networks，2000，13(4-5)：411-430.

[11] LEE H，BATTLE A，RAIN R，et al. Efficient sparse coding algorithms[C]//Proceedings of the 19th international conference on neural information processing systems. Vancouver，Canada：IEEE，2006：801-808.

[12] BENGIO Y，COURVILLE A，VINCENT P. Representation learning：a review and new perspectives[J]. IEEE Transactions on Software Engineering，2013，35(8)：1798-1828.

[13] 刘知远，孙茂松，林衍凯，等. 知识表示学习研究进展[J]. 计算机研究与发展，2016，53(2)：247-261.

[14] CHEN Mo，YANG Qiong，TANG Xiaoou. Directed graph embedding[C]//Proceedings of the 20th international joint conference on artificial intelligence. Hyderabad，India：[s. n.]，2007：2707-2712.

[15] 陈维政，张 岩，李晓明. 网络表示学习[J]. 大数据，2015，1(3)：8-22.

[16] PEROZZI B，AL-RFOU R，SKIENA S. DeepWalk：online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]：ACM，2014：701-710.

[17] YU H F，JAIN P，KAR P，et al. Large-scale multi-label learning with missing labels[C]//Proceedings of the 31st international conference on machine learning. [s. l.]：[s. n.]，2014：593-601.

[18] YANG Cheng，LIU Zhiyuan，ZHAO Deli，et al. Network representation learning with rich text information[C]//Proceedings of the 24th international conference on artificial intelligence. Buenos Aires，Argentina：AAAI Press，2015：2111-2117.

[19] TANG Jian，QU Meng，WANG Mingzhe，et al. LINE：large-scale information network embedding[C]//Proceedings of the 24th international conference on world wide web. Florence，Italy：[s. n.]，2015：1067-1077.

[20] TANG Jian，QU Meng，MEI Qiaozhu. PTE：predictive text embedding through large-scale heterogeneous text networks[C]//ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]：ACM，2015：1165-1174.

[21] GROVER A，LESKOVEC J. Node2vec：scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco，California，USA：ACM，2016：855-864.

[22] YANG J，LESKOVEC J. Overlapping communities explain core-periphery organization of networks[J]. Proceedings of the IEEE，2014，102(12)：1892-1902.

[23] HENDERSON K，GALLAGHER B，ELIASSI-RAD T，et al. RolX：structural role extraction & mining in large graphs[C]//Proceedings of the 18th ACM international conference on knowledge discovery and data mining. Beijing，China：ACM，2012：1231-1239.

[24] 叶 嘉，黄桂敏. 一种非结构化 P2P 的随机有向图拓扑模型[J]. 计算机应用与软件，2007，24(4)：64-66.