(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112836050 A
(43)申请公布日 2021.05.25

(21)申请号 202110154722.2

(22)申请日 2021.02.04

(71)申请人 山东大学
地址 250101 山东省济南市高新区舜华路1500号

(72)发明人 刘士军 陈冠恒 郭子瑜 梅广旭 潘丽 杨承磊 孟祥旭

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 黄海丽

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/901(2019.01)
G06N 3/04(2006.01)

权利要求书3页 说明书8页 附图2页

(54)发明名称

针对关系不确定性的引文网络节点分类方法及系统

(57)摘要

本发明公开了针对关系不确定性的引文网络节点分类方法及系统，包括：获取待预测类别标签的论文，获取已知类别标签的引文网络；根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图；基于元路径邻居图，生成若干个泛化图；将所有泛化图，均输入到预训练后的图卷积神经网络中，输出待预测类别标签论文的类别标签。本发明通过对异质图的元路径邻居图进行重构，解决了异质图中关系的不确定性问题，同时经过泛化得到更多的图结构样本增加训练数据中的对抗性实例的数量，从而增强了模型的鲁棒性。

CN 112836050 A

1.针对关系不确定性的引文网络节点分类方法,其特征是,包括:

获取待预测类别标签的论文,获取已知类别标签的引文网络;

根据所述待预测类别标签的论文和所述已知类别标签的引文网络,构建元路径邻居图;

基于元路径邻居图,生成若干个泛化图;

将所有泛化图,均输入到预训练后的图卷积神经网络中,输出待预测类别标签论文的类别标签。

2.如权利要求1所述的针对关系不确定性的引文网络节点分类方法,其特征是,根据所述待预测类别标签的论文和所述已知类别标签的引文网络,构建元路径邻居图;具体步骤包括:

根据论文与论文之间的引用与被引用关系,和论文与论文之间作者是否相同,来构建元路径邻居图;

将每一篇论文视为一个节点;

如果论文与论文之间存在引用与被引用关系,则表示两个节点之间存在连接的边;或者,如果论文与论文之间存在作者相同,则表示两个节点之间存在连接的边;否则,表示两个节点之间不存在连接的边;得到元路径邻居图。

3.如权利要求1所述的针对关系不确定性的引文网络节点分类方法,其特征是,基于元路径邻居图,生成若干个泛化图;具体步骤包括:

基于分类混合隶属度随机块模型对元路径邻居图进行处理,得到若干个泛化图。

4.如权利要求1所述的针对关系不确定性的引文网络节点分类方法,其特征是,将所有泛化图,均输入到预训练后的图卷积神经网络中,输出待预测类别标签论文的类别标签;其中,预训练后的图卷积神经网络的训练步骤包括:

获取已知类别标签的引文网络;

根据所述已知类别标签的引文网络,构建元路径邻居图;

基于元路径邻居图,生成若干个泛化图;

将所有泛化图作为图卷积神经网络的输入值,将已知类别标签作为图卷积神经网络的输出值,结合MC-dropout方法对图卷积神经网络进行训练,得到训练后的图卷积神经网络。

5.如权利要求1所述的针对关系不确定性的引文网络节点分类方法,其特征是,基于元路径邻居图,生成若干个泛化图;具体步骤包括:

每个元路径邻居图G_{ϕ}均被视为一个随机图参数族的实现,使用a-MMSBM模型对G_{ϕ}建模,以此获取随机图参数并实现泛化图的生成;

对于包含M个节点的G_{ϕ},这些节点分为K类,而同时归属多种类别的任意节点a∈M,其归属类别通过一个维度为K的概率分布π_a表示,即$\pi_a=[\pi_{a1},\cdots,\pi_{ak}]^T$,其中$\pi_{ak}$表示节点a属于类别k的概率;

同时每种类别都有其类别强度$\beta_k\in(0,1)$,用于评估该类别的成员之间联系的紧密程度;

对于G_{ϕ}中任意节点对(a,b),存在指示变量$D_{a\to b}=k_1$表示节点a指向节点b时节点a所归属的类别为k_1,指示变量$D_{b\to a}=k_2$表示节点b指向节点a时节点b所归属的类别为k_2;

节点对(a,b)的连接概率$L_{ab}\in\{0,1\}$,取值为0时表示不存在连接的边,取值为1时表示

存在连接的边;总的来说,G_{ϕ}的节点之间是否连接取决于它们的类别成员的相似性和它们共享类别的强度;

基于元路径邻居图G_{ϕ}并利用a-MMSBM生成泛化图通过下面的描述定义:

(1)对于每个类别k,采样其类别强度β_k～Beta(η);

(2)对于任意节点a∈M,采样其归属的类别分布π_a～Dirichlet(α);

(3)对于任意节点对(a,b),分别采样其指示变量$D_{a\to b}$～π_a和$D_{b\to a}$～π_b;当$D_{a\to b}=D_{b\to a}=k$时,采样它们之间的边L_{ab}～Bernoulli(β_k);当$D_{a\to b}\neq D_{b\to a}$时$L_{ab}$～Bernoulli(δ),其中δ∈(0,1)是跨类别连接概率;

其中η和α均为超参数,上述的生成模型过程通过下面的联合后验公式描述:

$$p(\mathrm{G},D,\pi,\beta\,|\,\alpha,\eta)=\prod_{a=1}^{M}\prod_{b>a}^{M}p(L_{ab}\,|\,D_{a\to b},D_{b\to a},\beta)p(D_{a\to b}\,|\,\pi_a)p(D_{b\to a}\,|\,\pi_b)\prod_{a=1}^{M}p(\pi_a\,|\,\alpha)\prod_{k=1}^{K}p(\beta_k\,|\,\eta)$$

通过元路径邻居图G_{ϕ}获取a-MMSBM参数π和β的联合后验分布如下:

$$p(\pi,\beta\,|\,\mathrm{G}_{\phi})\propto p(\pi)p(\beta)p(\mathrm{G}_{\phi}\,|\,\pi,\beta)=\prod_{k=1}^{K}p(\beta_k)\prod_{a=1}^{M}p(\pi_a)\prod_{1\le a<b\le M}\sum_{D_{a\to b},D_{b\to a}}p(L_{ab},D_{a\to b},D_{b\to a}\,|\,\pi_a,\pi_b,\beta)$$ 。

6.如权利要求1所述的针对关系不确定性的引文网络节点分类方法,其特征是,所述方法还包括:

结合贝叶斯方法构建以节点标签为目标的推导公式,最终利用近似方法得到公式的近似表示,利用GCN中softmax函数得到每个权重样本的输出,并通过累加这些输出得到节点的标签以此得到节点的分类结果。

7.如权利要求1所述的针对关系不确定性的引文网络节点分类方法,其特征是,

节点标签的预测概率公式表达为:

$$p(\mathrm{Z}\,|\,\mathrm{X},\mathrm{Y_L},\mathrm{G_{obs}})\approx\frac{1}{\mathrm{H}}\sum_{\phi}^{\mathrm{H}}\frac{1}{\mathrm{N_G S}}\sum_{\mathrm{n=1}}^{\mathrm{N_G}}\sum_{\mathrm{s=1}}^{\mathrm{S}}p(Z\,|\,\mathrm{W}_{s,n,\phi},\mathrm{G}_{n,\phi},\mathrm{X})$$

其中,Z表示神经网络的输出向量,X表示节点的特征向量,Y_L表示节点的标签向量,G_{obs}表示输入的异质图,H表示采样元路径邻居图的样本数量,Φ表示一条元路径,其中一条元路径可以构建一个元路径邻居图,N_G表示采样泛化图的样本数量,n表示第n个泛化图样本,S表示采样权重样本的数量,s表示第s个权重样本,$G_{n,\phi}$表示基于元路径邻居图G_{ϕ}得到的第n个泛化图样本,$W_{s,n,\phi}$表示基于泛化图$G_{n,\phi}$得到的第s个权重样本。

8.针对关系不确定性的引文网络节点分类系统,其特征是,包括:

获取模块,其被配置为:获取待预测类别标签的论文,获取已知类别标签的引文网络;

构建模块,其被配置为:根据所述待预测类别标签的论文和所述已知类别标签的引文网络,构建元路径邻居图;

生成模块,其被配置为:基于元路径邻居图,生成若干个泛化图;

输出模块,其被配置为:将所有泛化图,均输入到预训练后的图卷积神经网络中,输出待预测类别标签论文的类别标签。

9.一种电子设备,其特征是,包括:一个或多个处理器、一个或多个存储器、以及一个或多个计算机程序;其中,处理器与存储器连接,上述一个或多个计算机程序被存储在存储器中,当电子设备运行时,该处理器执行该存储器存储的一个或多个计算机程序,以使电子设

备执行上述权利要求1-7任一项所述的方法。

10.一种计算机可读存储介质，其特征是，用于存储计算机指令，所述计算机指令被处理器执行时，完成权利要求1-7任一项所述的方法。

针对关系不确定性的引文网络节点分类方法及系统

技术领域

[0001] 本发明涉及图神经网络的人工智能分类技术领域，特别是涉及针对关系不确定性的引文网络节点分类方法及系统。

背景技术

[0002] 本部分的陈述仅仅是提到了与本发明相关的背景技术，并不必然构成现有技术。

[0003] 现实世界中存在的许多网络结构，比如引文网络、社交网络、交通网络等，吸引了研究人员的关注。由多种类型的节点和边构成的异质信息网络属于其中的一种，此类网络含有丰富的结构和语义信息且在现实世界中广泛存在，更加吸引了广泛的研究兴趣。近年来，针对异质信息网络，越来越多的异质图模型被构建出来解决异质图中的节点分类、节点聚类、链接预测等任务。这些模型虽然都表现出良好的性能，但是它们没有考虑到在异质图中存在关系的不确定性问题。现实应用中导致异质信息网络关系不确定性的主要原因有以下几种：

[0004] (1)异质图关系复杂，构建网络时导入信息不完备，缺失重要的关系。例如图2(a)中存在强关联性的两个论文节点P1、P2由于发表时间相同而没有相互引用，导致重要关系的缺失，网络结构中节点之间缺少连接的边。其中图2(b)为节点属性；

[0005] (2)异质图节点类型多样，关系之间的权重不一，过多的次要关系会影响重要关系。例如图2(a)中具有强关联性的论文节点P3、P2之间的重要关系存在连接的边，但是P3引用了更多的弱关联性的论文P4、P5，这样过多的次要关系会对重要关系造成影响。

[0006] (3)脏数据导致错误的关联关系。例如图2(a)中，由于脏数据的存在，发表于不同会议且无关联性的论文节点P5、P6之间存在错误的引用关系，导致网络结构中毫无联系的节点之间存在连接的边。其他从真实世界获取数据构建的异质信息网络所包含的关系中同样存在类似的不确定性。

[0007] 发明人发现，目前的异质图神经网络研究除了尚未系统地解决异质图存在关系的不确定性问题，还存在鲁棒性较弱的问题。鲁棒性较弱的模型无法保证稳定的训练效果，因为神经网络容易受到对抗性实例的影响，数据进行扰动后，模型训练效果容易下降。造成这种现象的原因是在数据集中对抗性实例样本数量太少，而神经网络的高度非线性特性导致模型学习不到这些特殊样本，因而在面对对抗性实例干扰时，模型表现出鲁棒性较弱的现象。这一缺陷严重限制了异质图神经网络在真实世界中的应用，因为真实世界中异质信息网络更容易受到各种因素干扰而发生改变。

发明内容

[0008] 为了解决现有技术的不足，本发明提供了针对关系不确定性的引文网络节点分类方法及系统；在解决异质图中关系不确定性问题的同时提高模型的鲁棒性。

[0009] 第一方面，本发明提供了针对关系不确定性的引文网络节点分类方法；

[0010] 针对关系不确定性的引文网络节点分类方法，包括：

[0011] 获取待预测类别标签的论文，获取已知类别标签的引文网络；

[0012] 根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图；

[0013] 基于元路径邻居图，生成若干个泛化图；

[0014] 将所有泛化图，均输入到预训练后的图卷积神经网络中，输出待预测类别标签论文的类别标签。

[0015] 第二方面，本发明提供了针对关系不确定性的引文网络节点分类系统；

[0016] 针对关系不确定性的引文网络节点分类系统，包括：

[0017] 获取模块，其被配置为：获取待预测类别标签的论文，获取已知类别标签的引文网络；

[0018] 构建模块，其被配置为：根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图；

[0019] 生成模块，其被配置为：基于元路径邻居图，生成若干个泛化图；

[0020] 输出模块，其被配置为：将所有泛化图，均输入到预训练后的图卷积神经网络中，输出待预测类别标签论文的类别标签。

[0021] 第三方面，本发明还提供了一种电子设备，包括：一个或多个处理器、一个或多个存储器、以及一个或多个计算机程序；其中，处理器与存储器连接，上述一个或多个计算机程序被存储在存储器中，当电子设备运行时，该处理器执行该存储器存储的一个或多个计算机程序，以使电子设备执行上述第一方面所述的方法。

[0022] 第四方面，本发明还提供了一种计算机可读存储介质，用于存储计算机指令，所述计算机指令被处理器执行时，完成第一方面所述的方法。

[0023] 与现有技术相比，本发明的有益效果是：

[0024] 本发明将异质图网络结构以及神经网络权重视为随机变量，结合贝叶斯方法构建异质图神经网络，从而实现更多可能性的预测，同时使用a-MMSBM对元路径邻居图建模，以此生成一定数量的重构结构的泛化图，从而解决了异质图中关系的不确定性问题，并增加了训练数据中对抗性实例的数量，从而提高模型的鲁棒性。

[0025] 本发明附加方面的优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0026] 构成本发明的一部分的说明书附图用来提供对本发明的进一步理解，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。

[0027] 图1为本发明的流程图；

[0028] 图2(a)为本发明的引文网络；

[0029] 图2(b)为本发明的节点类型。

具体实施方式

[0030] 应该指出，以下详细说明都是示例性的，旨在对本发明提供进一步的说明。除非另有指明，本文使用的所有技术和科学术语具有与本发明所属技术领域的普通技术人员通常

理解的相同含义。

[0031] 需要注意的是，这里所使用的术语仅是为了描述具体实施方式，而非意图限制根据本发明的示例性实施方式。如在这里所使用的，除非上下文另外明确指出，否则单数形式也意图包括复数形式，此外，还应当理解的是，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0032] 在不冲突的情况下，本发明中的实施例及实施例中的特征可以相互组合。

[0033] 本发明所使用的术语解释：

[0034] 引文网络是由文献间引用和被引用的关系构成的集合，这些文献资料包括科技期刊、专利文献、会议论文集、科技报告和学位论文等多种形式，其较好地描述了科学领域的发展、学科间的关系。

[0035] 引文网络被认为是社会网络的变型，该网络中的节点是文献，边代表了文献间的引用关系。引文网络的发展区别于一般的社会网络，它由文献的引用关系确定，不可随意添加或删除，其中的引用关系在时间上具有单向性，只能是后期的文献引用前期的文献。引文与被引文之间体现了文献内容的相关性以及知识的传递。实际上，引文网络中隐含了由文献作者组成的研究群体，该群体具有相似的研究内容，并代表着某个领域的研究现状及未来发展趋势，对促进科研的发展及加快学术成果的流动起着重要的作用。

[0036] 实施例一

[0037] 本实施例提供了针对关系不确定性的引文网络节点分类方法；

[0038] 如图1所示，针对关系不确定性的引文网络节点分类方法，包括：

[0039] S101：获取待预测类别标签的论文，获取已知类别标签的引文网络；

[0040] S102：根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图；

[0041] S103：基于元路径邻居图，生成若干个泛化图；

[0042] S104：将所有泛化图，均输入到预训练后的图卷积神经网络中，输出待预测类别标签论文的类别标签。

[0043] 作为一个或多个实施例，所述S102：根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图；具体步骤包括：

[0044] 根据论文与论文之间的引用与被引用关系，和论文与论文之间作者是否相同，来构建元路径邻居图；

[0045] 将每一篇论文视为一个节点；

[0046] 如果论文与论文之间存在引用与被引用关系，则表示两个节点之间存在连接的边；或者，如果论文与论文之间存在作者相同，则表示两个节点之间存在连接的边；否则，表示两个节点之间不存在连接的边；得到元路径邻居图。

[0047] 示例性的，所述S102：根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图；具体步骤包括：针对任意异质图$G_{obs}=(V,E)$，其中V表示节点集合，E表示边集合，设置节点类型集合为N，边类型集合为ε，由此可定义一条元路径Φ为$n_1 \xrightarrow{r_1} n_2 \xrightarrow{r_2} \cdots \xrightarrow{r_l} n_{l+1}$，其中$n_{l+1} \in N$，$r_1 \in \varepsilon$，本发明设置一条元路径表示单一关系，且

元路径两端连接相同类型的节点，由此可以构建一个元路径邻居图$G_{\phi}=(V',\phi)$，其中V'表示由元路径Φ连接的节点集。

[0048] 作为一个或多个实施例，所述S103：基于元路径邻居图，生成若干个泛化图；具体步骤包括：

[0049] 基于分类混合隶属度随机块模型(assortative mixed-membership stochastic block model，a-MMSBM)对元路径邻居图进行处理，得到若干个泛化图。

[0050] 示例性的，所述S103：基于元路径邻居图，生成若干个泛化图；具体步骤包括：

[0051] 每个元路径邻居图G_{ϕ}均被视为一个随机图参数族的实现，使用a-MMSBM模型对G_{ϕ}建模，以此获取随机图参数并实现泛化图的生成。

[0052] 对于包含M个节点的G_{ϕ}，这些节点分为K类，而同时归属多种类别的任意节点$a\in M$，其归属类别通过一个维度为K的概率分布π_a表示，即$\pi_a=[\pi_{a1},\cdots,\pi_{ak}]^T$，其中$\pi_{ak}$表示节点a属于类别k的概率。

[0053] 同时每种类别都有其类别强度$\beta_k\in(0,1)$，用于评估该类别的成员之间联系的紧密程度。

[0054] 对于G_{ϕ}中任意节点对(a,b)，存在指示变量$D_{a\to b}=k_1$表示节点a指向节点b时节点a所归属的类别为k_1，指示变量$D_{b\to a}=k_2$表示节点b指向节点a时节点b所归属的类别为k_2。

[0055] 节点对(a,b)的连接概率$L_{ab}\in\{0,1\}$，取值为0时表示不存在连接的边，取值为1时表示存在连接的边。总的来说，G_{ϕ}的节点之间是否连接取决于它们的类别成员的相似性和它们共享类别的强度。

[0056] 基于元路径邻居图G_{ϕ}并利用a-MMSBM生成泛化图通过下面的描述定义：

[0057] (1)对于每个类别k，采样其类别强度$\beta_k\sim$Beta(η)；

[0058] (2)对于任意节点$a\in M$，采样其归属的类别分布$\pi_a\sim$Dirichlet(α)；

[0059] (3)对于任意节点对(a,b)，分别采样其指示变量$D_{a\to b}\sim\pi_a$和$D_{b\to a}\sim\pi_b$。当$D_{a\to b}=D_{b\to a}=k$时，采样它们之间的边$L_{ab}\sim$Bernoulli(β_k)；当$D_{a\to b}\neq D_{b\to a}$时$L_{ab}\sim$Bernoulli(δ)，其中$\delta\in(0,1)$是跨类别连接概率。

[0060] 其中η和α均为超参数，上述的生成模型过程通过下面的联合后验公式描述：

$$p(\mathrm{G},D,\pi,\beta\,|\,\alpha,\eta)=\prod_{a=1}^{M}\prod_{b>a}^{M}p(L_{ab}\,|\,D_{a\to b},D_{b\to a},\beta)p(D_{a\to b}\,|\,\pi_a)p(D_{b\to a}\,|\,\pi_b)\prod_{a=1}^{M}p(\pi_a\,|\,\alpha)\prod_{k=1}^{K}p(\beta_k\,|\,\eta)$$

[0061] 通过元路径邻居图G_{ϕ}获取a-MMSBM参数π和β的联合后验分布如下：

$$p(\pi,\beta\,|\,\mathrm{G}_{\phi})\propto p(\pi)p(\beta)p(\mathrm{G}_{\phi}\,|\,\pi,\beta)=\prod_{k=1}^{K}p(\beta_k)\prod_{a=1}^{M}p(\pi_a)\prod_{1\le a<b\le M}\sum_{D_{a\to b},D_{b\to a}}p(L_{ab},D_{a\to b},D_{b\to a}\,|\,\pi_a,\pi_b,\beta)$$ 。

[0062] 作为一个或多个实施例，所述S104：将所有泛化图，均输入到预训练后的图卷积神经网络中，输出待预测类别标签论文的类别标签；其中，预训练后的图卷积神经网络的训练步骤包括：

[0063] S1041：获取已知类别标签的引文网络；

[0064] S1042：根据所述已知类别标签的引文网络，构建元路径邻居图；

[0065] S1043：基于元路径邻居图，生成若干个泛化图；

[0066] S1044：将所有泛化图作为图卷积神经网络的输入值，将已知类别标签作为图卷积神经网络的输出值，结合MC-dropout(Monte Carlo dropout)方法对图卷积神经网络

(Graph Convolutional Network,GCN)进行训练,得到训练后的图卷积神经网络。

[0067] 针对S1042～S1044结合贝叶斯方法构建以节点标签为目标的推导公式,最终利用近似方法得到公式的近似表示,利用GCN中softmax函数得到每个权重样本的输出,并通过累加这些输出得到节点的标签以此得到节点的分类结果。

[0068] 通过重构网络结构来解决异质图中存在的关系不确定性问题,并结合贝叶斯方法,将网络结构以及神经网络的权重视为随机变量,以节点的标签为推理目标,利用边缘化方法构建推导公式。

[0069] 在节点分类任务中,对于单个元路径邻居图的泛化图样本G,输入数据表示为节点特征X、节点标签Y_L与之相对应的输出表示为$Z=\{z_1,\cdots,z_n\}$,目标是通过神经网络训练得到一个能反映输入和输出之间关系的函数$z=f(x)$,本发明使用贝叶斯方法将神经网络权重参数W建模为随机变量,引入它们的先验分布,同时因为W是不确定的,所以神经网络的输出也是随机变量。

[0070] 对于新输入x的预测,根据相应的W的后验分布,通过积分函数得到:

[0071] $p(z|x,X,Y_L,G)=\int p(z|x,W,G)p(W|X,Y_L,G)dW$

[0072] 使用GCN建模$p(W|X,Y_L,G)$,并使用softmax函数实现$p(z|x,W,G)$以得到节点的分类标签。

[0073] 蒙特卡罗近似得到上式的近似公式:

[0074] $$p(z|x,X,Y_L,G)\approx\frac{1}{S}\sum_{s=1}^{S}p(z|x,W_s,G)$$

[0075] 其中S个权重样本W_s通过结合了MC-dropout方法的GCN得到。

[0076] 元路径邻居图G_ϕ、随机图参数$\lambda=\{\pi,\beta\}$、泛化图G的权重参数W以及节点标签Z都被视为随机变量,为本发明的最终目的是获得节点的标签,基于此,利用边缘化方法可以得到节点标签的后验概率计算公式:

[0077] $p(Z|X,Y_L,G_{obs})=\int p(Z|W,G,X)p(W|X,Y_L,G)p(G|\lambda)p(\lambda|G_\phi)p(G_\phi|G_{obs})dWdGd\lambda dG_\phi$

[0078] 其中$p(G_\phi|G_{obs})$表示从异质图G_{obs}中获取元路径邻居图G_ϕ的概率,$p(\lambda|G_\phi)$表示基于元路径邻居图G_ϕ获取随机图参数族$\lambda=\{\pi,\beta\}$的概率,$p(G|\lambda)$表示利用这些参数λ构建泛化图G的概率,$p(W|X,Y_L,G)$表示对于单个泛化图G采样神经网络权重参数W的概率,最终基于节点特征X、泛化图G以及权重样本W得到节点的标签分布$p(Z|W,G,X)$。

[0079] 对于$p(G_\phi|G_{obs})$的实现方式,针对不同数据集预定义多种元路径,基于这些元路径构建元路径邻居图样本集,再通过均匀采样的方式实现从异质图中采样元路径邻居图。对于$p(\lambda|G_\phi)$以及$p(G|\lambda)$,则通过随机图生成模型a-MMSBM对其建模来推理实现。对于权重后验$p(W|X,Y_L,G)$,本发明通过结合了MC-dropout方法的GCN实现权重样本W采样,最后通过GCN的softmax函数的结果对$p(Z|W,G,X)$建模。

[0080] 由此,得到上述节点标签的后验概率计算公式的蒙特卡洛近似:

[0081] $$p(Z|X,Y_L,G_{obs})\approx\frac{1}{H}\sum_{\phi}^{H}\frac{1}{I}\sum_{i}^{I}\frac{1}{N_G S}\sum_{n=1}^{N_G}\sum_{s=1}^{S}p(Z|W_{s,n,i,\phi},G_{n,i,\phi},X)$$

[0082] 其中,从$p(G_\phi|G_{obs})$采样H个元路径邻居图样本G_ϕ,对于每个G_ϕ通过随机图生成模型,从$p(\lambda|G_\phi)$获取I个参数族样本λ_i,从$p(G|\lambda_i)$采样出N_G个泛化图样本$G_{n,i,\phi}$,这些泛化图

样本的准确度取决于对元路径邻居图样本所构建的随机图生成模型，采用a-MMSBM作为随机图生成模型。

[0083] 对于权重样本的采样$p(W|X,Y_L,G_{n,i,\phi})$，通过结合了MC-dropout方法的GCN对每个$G_{n,i,\phi}$采样S个权重样本$W_{s,n,i,\phi}$，并对$W_{s,n,i,\phi}$使用softmax函数得到节点标签的概率分布，最后通过累加这些标签分布得到节点最终的标签分布$p(Z|X,Y_L,G_{obs})$。

[0084] 对于a-MMSBM的随机参数π、β可以采用随机优化方法学习，但是由于a-MMSBM的后验维度过高，对其参数采用随机初始化的方式影响了训练效果，所以本发明使用GCN预训练元路径邻居图G_ϕ，利用softmax函数的输出初始化参数π和β。同时为了避免因为参数π和β取值范围过大导致a-MMSBM生成的泛化图样本$G_{n,i,\phi}$与G_ϕ差异过大，本发明使用最大后验估计了替代π和β的积分，利用合适的π和β的先验得到近似公式：

[0085] $$\left\{\hat{\pi},\hat{\beta}\right\}=\arg\max_{\pi,\beta} p(\pi,\beta|G_\phi)$$

[0086] 由此$G_{n,i,\phi}$改写为$G_{n,\phi}$，$W_{s,n,i,\phi}$改写为$W_{s,n,\phi}$，上述节点标签的后验概率计算公式的蒙特卡洛近似可以进一步简写为：

[0087] $$p(Z|X,Y_L,G_{obs})\approx\frac{1}{H}\sum_{\phi}^{H}\frac{1}{N_G S}\sum_{n=1}^{N_G}\sum_{s=1}^{S}p(Z|W_{s,n,\phi},G_{n,\phi},X)$$

[0088] 其中，Z表示神经网络的输出向量，X表示节点的特征向量，Y_L表示节点的标签向量，G_{obs}表示输入的异质图，H表示采样元路径邻居图的样本数量，Φ表示一条元路径，其中一条元路径可以构建一个元路径邻居图，N_G表示采样泛化图的样本数量，n表示第n个泛化图样本，S表示采样权重样本的数量，s表示第s个权重样本，$G_{n,\phi}$表示基于元路径邻居图G_ϕ得到的第n个泛化图样本，$W_{s,n,\phi}$表示基于泛化图$G_{n,\phi}$得到的第s个权重样本。

[0089] 利用a-MMSBM从$p(G_{n,\phi}|\hat{\pi},\hat{\beta})$采样得到$G_{n,\phi}$，结合MC-dropout方法对$G_{n,\phi}$使用GCN实现从$p(W|X,Y_L,G_{n,\phi})$采样$W_{s,n,\phi}$，并使用softmax函数得到节点标签分布，最后采用累加的方式计算节点的标签分布。

[0090] 通过对异质图的元路径邻居图进行重构以及泛化得到新的图结构，原来异质图中本身具有强关系的节点之间的边会增强，而弱关系或脏数据带来的假边则会被忽略，解决了异质图中关系的不确定性问题，同时经过泛化得到更多的图结构样本能够增加训练数据中的对抗性实例的数量，从而增强了模型的鲁棒性。

[0091] 本发明基于异质图数据集的领域知识，预定义符合现实语义的元路径，以此构建元路径邻居图；所述元路径邻居图由相同类型节点构成，其中的边表示单一的关系；利用分类混合隶属度随机块模型a-MMSBM对被视为随机图的元路径邻居图建模，以此生成元路径邻居图的泛化图；所述泛化图的网络结构与元路径邻居图类似；通过结合MC-dropout方法的图卷积神经网络GCN训练泛化图得到神经网络权重参数样本；针对上述步骤结合贝叶斯方法构建以节点标签为目标的推导公式，最终利用近似方法得到公式的近似表示，利用GCN中softmax函数得到每个权重样本的输出，并通过累加这些输出得到节点的标签以此得到节点的分类结果。本发明通过对异质图的元路径邻居图进行重构，解决了异质图中关系的不确定性问题，同时经过泛化得到更多的图结构样本增加训练数据中的对抗性实例的数

量,从而增强了模型的鲁棒性。

[0092] 实施例二

[0093] 本实施例提供了针对关系不确定性的引文网络节点分类系统;

[0094] 针对关系不确定性的引文网络节点分类系统,包括:

[0095] 获取模块,其被配置为:获取待预测类别标签的论文,获取已知类别标签的引文网络;

[0096] 构建模块,其被配置为:根据所述待预测类别标签的论文和所述已知类别标签的引文网络,构建元路径邻居图;

[0097] 生成模块,其被配置为:基于元路径邻居图,生成若干个泛化图;

[0098] 输出模块,其被配置为:将所有泛化图,均输入到预训练后的图卷积神经网络中,输出待预测类别标签论文的类别标签。

[0099] 此处需要说明的是,上述获取模块、构建模块、生成模块和输出模块对应于实施例一中的步骤S101至S104,上述模块与对应的步骤所实现的示例和应用场景相同,但不限于上述实施例一所公开的内容。需要说明的是,上述模块作为系统的一部分可以在诸如一组计算机可执行指令的计算机系统中执行。

[0100] 上述实施例中对各个实施例的描述各有侧重,某个实施例中没有详述的部分可以参见其他实施例的相关描述。

[0101] 所提出的系统,可以通过其他的方式实现。例如以上所描述的系统实施例仅仅是示意性的,例如上述模块的划分,仅仅为一种逻辑功能划分,实际实现时,可以有另外的划分方式,例如多个模块可以结合或者可以集成到另外一个系统,或一些特征可以忽略,或不执行。

[0102] 实施例三

[0103] 本实施例还提供了一种电子设备,包括:一个或多个处理器、一个或多个存储器、以及一个或多个计算机程序;其中,处理器与存储器连接,上述一个或多个计算机程序被存储在存储器中,当电子设备运行时,该处理器执行该存储器存储的一个或多个计算机程序,以使电子设备执行上述实施例一所述的方法。

[0104] 应理解,本实施例中,处理器可以是中央处理单元CPU,处理器还可以是其他通用处理器、数字信号处理器DSP、专用集成电路ASIC,现成可编程门阵列FPGA或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0105] 存储器可以包括只读存储器和随机存取存储器,并向处理器提供指令和数据、存储器的一部分还可以包括非易失性随机存储器。例如,存储器还可以存储设备类型的信息。

[0106] 在实现过程中,上述方法的各步骤可以通过处理器中的硬件的集成逻辑电路或者软件形式的指令完成。

[0107] 实施例一中的方法可以直接体现为硬件处理器执行完成,或者用处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器、闪存、只读存储器、可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器,处理器读取存储器中的信息,结合其硬件完成上述方法的步骤。为避免重复,这里不再详细描述。

[0108] 本领域普通技术人员可以意识到，结合本实施例描述的各示例的单元及算法步骤，能够以电子硬件或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0109] 实施例四

[0110] 本实施例还提供了一种计算机可读存储介质，用于存储计算机指令，所述计算机指令被处理器执行时，完成实施例一所述的方法。

[0111] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。


获取待预测类别标签的论文，获取已知类别标签的引文网络
根据所述待预测类别标签的论文和所述已知类别标签的引文网络，构建元路径邻居图
基于元路径邻居图，生成若干个泛化图
将所有泛化图，均输入到预训练后的图卷积神经网络中，输出待预测类别标签论文的类别标签

图1

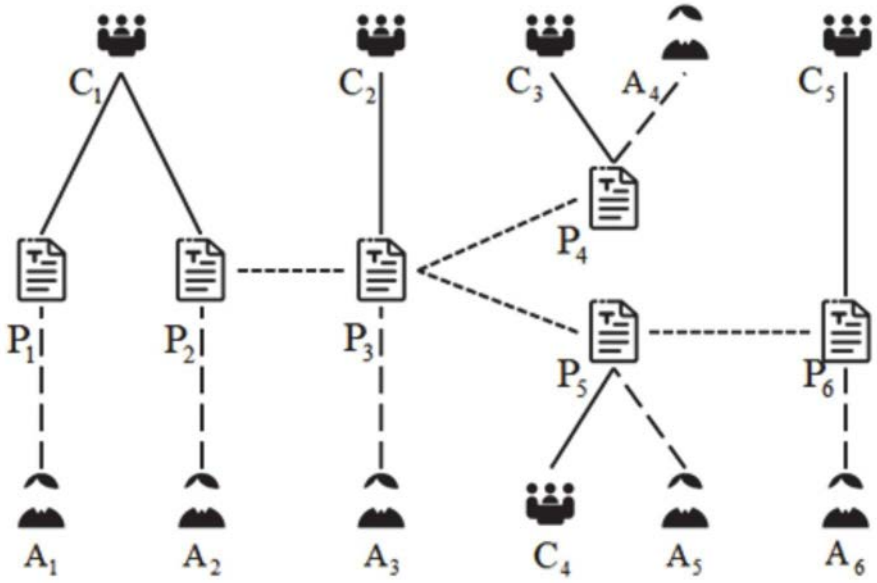
C1
C2
C3
A4
C5
P4
P1
P2
P3
P5
P6
A1
A2
A3
C4
A5
A6

图2(a)

作者(A)

论文(P)

会议(C)

图2(b)