



(12)发明专利申请

(10)申请公布号 CN 111046664 A

(43)申请公布日 2020.04.21

(21)申请号 201911188601.9

G06N 3/08(2006.01)

(22)申请日 2019.11.26

(71)申请人 哈尔滨工业大学(深圳)

地址 518055 广东省深圳市南山区桃源街  
道深圳大学城哈尔滨工业大学校区

(72)发明人 廖清 胡国勇 吴君宇 丁焯  
王轩 李京竹

(74)专利代理机构 广州三环专利商标代理有限  
公司 44202

代理人 郭浩辉 麦小婵

(51)Int.Cl.

G06F 40/289(2020.01)

G06F 40/205(2020.01)

G06K 9/62(2006.01)

G06N 3/04(2006.01)

权利要求书3页 说明书9页 附图2页

(54)发明名称

基于多粒度的图卷积神经网络的假新闻检测方法及系统

(57)摘要

本发明公开了基于多粒度的图卷积神经网络的假新闻检测方法及系统。在一个实施例中：所述方法，通过构建不同深度距离的概率转移矩阵，形成多粒度的信息输入，借助图卷积神经网络对含有不同节点特征的关系图分别进行卷积，得到不同维度的特征在不同节点间的相关性分析结果；采用注意力机制对所述相关性分析结果进行加权求和，得到不同的注意力权重系数，进而得到新闻节点的表示，获得新闻真假性判断的结果。本发明能够对不同维度的特征在不同的节点间进行相关性分析，通过找到假新闻在不同信息特征之间的相似性，提高重要特征的权重，忽略无关特征信息的。



1. 一种基于多粒度的图卷积神经网络的假新闻检测方法,其特征在于,包括:

构建不同深度距离的概率转移矩阵,形成多维度的图卷积神经网络,对含有不同节点特征的图卷积神经网络分别进行卷积,得到不同维度的特征在不同节点间的相关性分析结果;

采用注意力机制对所述相关性分析结果进行加权求和,得到不同的注意力权重系数,进而得到新闻真假性判断的结果。

2. 根据权利要求1所述的基于多粒度的图卷积神经网络的假新闻检测方法,其特征在于,所述每个图卷积神经网络的输入为一个邻接矩阵和一个特征矩阵,所述邻接矩阵根据辅助信息中的每一种关系图构成,所述特征矩阵由文本内容的词向量和信用历史组成。

3. 根据权利要求2所述的基于多粒度的图卷积神经网络的假新闻检测方法,其特征在于,所述邻接矩阵由以下公式确定:

$$A_{ij} = \begin{cases} 1, & \text{如果新闻 } i, j \text{ 发布者有相同的职业;} \\ 0, & \text{其他} \end{cases};$$

图卷积神经网络的分层传播规则为:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)});$$

其中,  $\tilde{A} = A + I$  是具有自连接的邻接矩阵,  $\tilde{D} = \sum_j \tilde{A}_{ij}$  是自连接邻接矩阵的度矩阵,  $H^{(1)}$  是图卷积中第1层的激活矩阵,  $H^{(0)} = X$  是初始的输入特征矩阵  $X$ ,  $\sigma(\cdot)$  是激活函数,  $\tilde{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$  是单步随机游走的转移概率矩阵,通过对该概率矩阵进行  $k$  阶近邻处理,  $\hat{A}_{ij}$  是节点  $i$  和节点  $j$  进行  $k$  阶近邻之后的矩阵,  $\hat{A}^k = \Pi_{l=1}^k \hat{A}_l$ , 通过对单步随机游走的转移概率矩阵进行  $k$  阶近邻,计算不同距离接近矩阵来描述节点之间的相关性,保留多粒度的信息提高各节点表示的多样性;

图卷积神经网络就所述近邻矩阵进行迭代更新的结果为:

$$Z_k = \hat{A}^k \text{ReLU}(\hat{A}^k X W_k^{(0)}) W_k^{(1)}, k = 1, 2, 3, \dots;$$

其中,  $W_k^{(0)}$  是初试输入层的权矩阵,  $W_k^{(1)}$  是第二层权矩阵,根据上述公式得到不同粒度的输出。

4. 根据权利要求3所述的基于多粒度的图卷积神经网络的假新闻检测方法,其特征在于,将每一个输出输入到注意力权重计算中,利用一个非线性函数来计算它们对应的注意力权重  $u_i$  ( $1 \leq i \leq m$ ),

$$u_i = \tanh(w_i z_i + b_i);$$

采用softmax函数对  $u_i$  ( $1 \leq i \leq m$ ) 进行归一化得到:

$$\alpha_i = \frac{\exp(u_i)}{\sum_{l=1}^m \exp(u_l)};$$

其中,  $\alpha_i$  取值范围在0到1之间;

利用求得的注意力权重  $u_i$  对迭代更新的结果  $z_k$  进行加权求和,得到一个新闻的判断表示  $p_j$ ,

$$p_j = \sum_{i=1}^m \alpha_i z_i ;$$

利用交叉熵来评估训练得到的判断和标记样本实际结果之间的偏差,损失函数为:

$$L = - \sum_{l \in Y_L} \sum_{f=1}^c Y_{lf} \ln p_{lf} ;$$

其中, $Y_L$ 是标记样本分类标签的集合, $c$ 是最终表示的维数,目标是要最小化交叉熵损失函数 $L$ ,通过反向传播不断修改权重使损失变小。

5. 一种基于多粒度的图卷积神经网络的假新闻检测系统,其特征在于,包括:

相关性分析单元,用于构建不同深度距离的概率转移矩阵,形成多维度的图卷积神经网络,对含有不同节点特征的图卷积神经网络分别进行卷积,得到不同维度的特征在不同节点间的相关性分析结果;

注意力机制融合单元,用于采用注意力机制对所述相关性分析结果进行加权求和,得到不同的注意力权重系数,进而得到新闻真假性判断的结果。

6. 根据权利要求5所述的基于多粒度的图卷积神经网络的假新闻检测系统,其特征在于,所述每个图卷积神经网络的输入为一个邻接矩阵和一个特征矩阵,所述邻接矩阵根据辅助信息中的每一种关系图构成,所述特征矩阵由文本内容的词向量和信用历史组成。

7. 根据权利要求6所述的基于多粒度的图卷积神经网络的假新闻检测系统,其特征在于,所述邻接矩阵由以下公式确定:

$$A_{ij} = \begin{cases} 1, & \text{如果新闻}i,j\text{发布者有相同的职业;} \\ 0, & \text{其他} \end{cases} ;$$

图卷积神经网络的分层传播规则为:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)});$$

其中, $\tilde{A} = A + I$ 是具有自连接的邻接矩阵, $\tilde{D} = \sum_j \tilde{A}_{ij}$ 是自连接邻接矩阵的度矩阵, $H^{(1)}$ 是图卷积中第1层的激活矩阵, $H^{(0)} = X$ 是初始的输入特征矩阵 $X$ , $\sigma(\cdot)$ 是激活函数, $\tilde{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 是单步随机游走的转移概率矩阵,通过对该概率矩阵进行 $k$ 阶近邻处理, $\hat{A}_{ij}$ 是节点 $i$ 和节点 $j$ 进行 $k$ 阶近邻之后的矩阵, $\hat{A}^k = \prod_{i=1}^k \hat{A}_i$ ,通过对单步随机游走的转移概率矩阵进行 $k$ 阶近邻,计算不同距离接近矩阵来描述节点之间的相关性,保留多粒度的信息提高各节点表示的多样性;

图卷积神经网络就所述近邻矩阵进行迭代更新的结果为:

$$Z_k = \hat{A}^k \text{ReLU}(\hat{A}^k X W_k^{(0)}) W_k^{(1)}, k = 1, 2, 3, \dots ;$$

其中, $W_k^{(0)}$ 是初试输入层的权矩阵, $W_k^{(1)}$ 是第二层权矩阵,根据上述公式得到不同粒度的输出。

8. 根据权利要求7所述的基于多粒度的图卷积神经网络的假新闻检测系统,其特征在于,将每一个输出输入到注意力权重计算中,利用一个非线性函数来计算它们对应的注意力权重 $u_i$  ( $1 \leq i \leq m$ ),

$$u_i = \tanh(w_i z_i + b_i);$$

采用softmax函数对 $u_i$  ( $1 \leq i \leq m$ )进行归一化得到:

$$\alpha_i = \frac{\exp(u_i)}{\sum_{i=1}^m \exp(u_i)};$$

其中,  $\alpha_i$ 取值范围在0到1之间;

利用求得的注意力权重 $u_i$ 对迭代更新的结果 $z_k$ 进行加权求和,得到一个新闻的判断表示 $p_j$ ,

$$p_j = \sum_{i=1}^m \alpha_i z_i;$$

利用交叉熵来评估训练得到的判断和标记样本实际结果之间的偏差,损失函数为:

$$L = - \sum_{l \in Y_L} \sum_{f=1}^c Y_{lf} \ln p_{lf};$$

其中,  $Y_L$ 是标记样本分类标签的集合,  $c$ 是最终表示的维数,目标是要最小化交叉熵损失函数 $L$ ,通过反向传播不断修改权重使损失变小。

## 基于多粒度的图卷积神经网络的假新闻检测方法及系统

### 技术领域

[0001] 本发明涉及新闻检测技术领域,尤其涉及基于多粒度的图卷积神经网络的假新闻检测方法及系统。

### 背景技术

[0002] 传统假新闻检测的缺点是需要人工设置特征,让新闻内容分词后根据这些特征来进行检测判断,由于假新闻的写作风格、主题以及来源各异,使得人工设计特征这个过程费时费力,十分繁复。

[0003] 卷积神经网络(CNN)的方法使用了新闻的所有特征,将新闻的所有的背景信息以及文本信息都作为特征平等地作为判断新闻真假性的依据,每一种特征在判断中都起到相同的作用,但是由于新闻中存在一些无用的信息,这些无用的信息与重要的信息在判断中起到相同的作用,有用的信息就受到无用信息(噪声)的影响,使得判断准确率下降,同时,由于新闻是短文本,上下文信息很少,输入到CNN中的特征不足以准确地判断新闻的真假,所以准确率较低。

[0004] 基于注意力的长短期记忆网络使用了记忆神经网络结合注意力机制处理上下文的相关信息,将上下文的相关信息作为特征来判断新闻的真假性,同时利用注意力机制进行加权,调节不同辅助信息特征的权重,以最大化提升分类表现。提取重要的信息作为判断特征的方法准确率比卷积神经网络高,但是这个方法忽略了新闻之间的联系,只关注了新闻本身的特征,没有考虑假新闻在各个维度间的关系,准确率还存在提高空间。

[0005] 上述两类深度学习方法从短新闻中能够提取的特征有限,通过对输入的特征向量进行分析计算后得到最终的表示,通过计算误差,根据对比误差反向传播修改特征权重,可以筛选出在分类中起重要作用的信息,提高分类表现,但没有考虑新闻与新闻之间的联系,无法很好的判断新闻的真假,判断准确率较低。

[0006] 使用图卷积神经网络的方式,图卷积网络将新闻上下文信息和背景信息都考虑到特征分类中,通过网络中不同新闻节点之间的信息传播,也把新闻之间的相关性进行学习。但由于新闻特征众多,不同的节点特征对应的网络迭代层数不同,即传播到的相邻节点的层数不同(迭代一次,就把节点周围连接节点的信息结合,以此类推),如果迭代次数过多,就会出现过度平滑的问题,所有节点的特征迭代到最后都相似,就无法完成区分新闻真假性的功能,因此每个节点只能包含一种背景特征,没有办法将所有的节点背景特征都进行迭代计算用于分类,无法用于多种关系的多维场景下。

### 发明内容

[0007] 本发明所要解决的技术问题在于,提供一种基于多粒度的图卷积神经网络的假新闻检测方法及系统,能够对不同维度的特征在不同的节点间进行相关性分析,通过找到假新闻特征之间的相似性,提高重要特征的权重,忽略无关特征信息的权重,减少无关信息在判断中产生的噪声影响,以更准确地对新闻的真假性进行判断。

[0008] 为解决上述问题,本发明实施例提供一种基于多粒度的图卷积神经网络的假新闻检测方法,其特征在于,包括:

[0009] 构建不同深度距离的概率转移矩阵,形成多维度的图卷积神经网络,对含有不同节点特征的图卷积神经网络分别进行卷积,得到不同维度的特征在不同节点间的相关性分析结果;

[0010] 采用注意力机制对所述相关性分析结果进行加权求和,得到不同的注意力权重系数,进而得到新闻真假性判断的结果。

[0011] 具体的,所述每个图卷积神经网络的输入为一个邻接矩阵和一个特征矩阵,所述邻接矩阵根据辅助信息中的每一种关系图构成,所述特征矩阵由文本内容的词向量和信用历史组成。

[0012] 具体的,所述邻接矩阵由以下公式确定:

$$[0013] \quad A_{ij} = \begin{cases} 1, & \text{如果新闻} i, j \text{ 发布者有相同的职业;} \\ 0, & \text{其他} \end{cases};$$

[0014] 图卷积神经网络的分层传播规则为:

$$[0015] \quad H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)});$$

[0016] 其中,  $\tilde{A} = A + I$  是具有自连接的邻接矩阵,  $\tilde{D} = \sum_j \tilde{A}_{ij}$  是自连接邻接矩阵的度矩阵,  $H^{(1)}$  是图卷积中第1层的激活矩阵,  $H^{(0)} = X$  是初始的输入特征矩阵,  $\sigma(\cdot)$  是激活函数,  $\tilde{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$  是单步随机游走的转移概率矩阵,通过对该概率矩阵进行k阶近邻处理,  $\hat{A}_{ij}$  是节点i和节点j进行k阶近邻之后的矩阵,  $\hat{A}^k = \Pi_{i=1}^k \hat{A}_i$ ,通过对单步随机游走的转移概率矩阵进行k阶近邻,计算不同距离接近矩阵来描述节点之间的相关性,保留多粒度的信息提高各节点表示的多样性;

[0017] 图卷积神经网络就所述近邻矩阵进行迭代更新的结果为:

$$[0018] \quad Z_k = \hat{A}^k \text{ReLU}(\hat{A}^k X W_k^{(0)}) W_k^{(1)}, k = 1, 2, 3, \dots;$$

[0019] 其中,  $W_k^{(0)}$  是初试输入层的权矩阵,  $W_k^{(1)}$  是第二层权矩阵,根据上述公式得到不同粒度的输出。

[0020] 具体的,将每一个输出输入到注意力权重计算中,利用一个非线性函数来计算它们对应的注意力权重  $u_i$  ( $1 \leq i \leq m$ ),

$$[0021] \quad u_i = \tanh(w_i z_i + b_i);$$

[0022] 采用softmax函数对  $u_i$  ( $1 \leq i \leq m$ ) 进行归一化得到:

$$[0023] \quad \alpha_i = \frac{\exp(u_i)}{\sum_{i=1}^m \exp(u_i)};$$

[0024] 其中,  $\alpha_i$  取值范围在0到1之间;

[0025] 利用求得的注意力权重  $u_i$  对迭代更新的结果  $z_k$  进行加权求和,得到一个新闻的判断表示  $p_j$ ,

$$[0026] \quad p_j = \sum_{i=1}^m \alpha_i z_i;$$

[0027] 利用交叉熵来评估训练得到的判断和标记样本实际结果之间的偏差,损失函数为:

$$[0028] \quad L = - \sum_{l \in Y_L} \sum_{f=1}^c Y_{lf} \ln p_{lf};$$

[0029] 其中, $Y_L$ 是标记样本分类标签的集合, $c$ 是最终表示的维数,目标是要最小化交叉熵损失函数 $L$ ,通过反向传播不断修改权重使损失变小。

[0030] 本发明实施例还提供一种基于多粒度的图卷积神经网络的假新闻检测系统,包括:

[0031] 相关性分析单元,用于构建不同深度距离的概率转移矩阵,形成多维度的图卷积神经网络,对含有不同节点特征的图卷积神经网络分别进行卷积,得到不同维度的特征在不同节点间的相关性分析结果;

[0032] 注意力机制融合单元,用于采用注意力机制对所述相关性分析结果进行加权求和,得到不同的注意力权重系数,进而得到新闻真假性判断的结果。

[0033] 实施本发明实施例具有如下有益效果:

[0034] 假新闻之间一定存在一些相似性,可以利用这些相似性来对新闻的真假性进行判断,多粒度图卷积神经网络模型采用了图卷积神经网络学习不同距离节点间的信息,通过迭代变换传播捕捉新闻间的相关性,相比现有技术仅仅采用新闻本身的特征作为分类依据有较大提高。

[0035] 同时多粒度图卷积网络采用构建多个不同深度距离的概率转移矩阵,得到不同特征下各个节点的相关性信息,可以将新闻的不同特征之间的相关性都挖掘出来,与单一图卷积神经网络相比,改善了单一图卷积神经网络没有完全利用到所有的新闻背景信息的缺陷。

[0036] 相比于卷积神经网络以及记忆网络都对新闻词向量的特征进行同等的处理,没有区分不同特征的重要性,有较大的噪声影响判断结果,多粒度的图卷积神经网络采用注意力机制,其核心是在预测新闻类别时所用的上下文的特征是不同的,可以直观地解释各个句子和词对分类类别的重要性。通过使用注意力机制将不同的多个维度的图卷积神经网络结果进行融合,充分考虑到了新闻之间的联系以及新闻本身的各种特征。

## 附图说明

[0037] 图1是本发明一个实施例提供的一种基于多粒度的图卷积神经网络的假新闻检测方法的流程示意图;

[0038] 图2是本发明一个实施例提供的卷积神经网络示意图示意图;

[0039] 图3是本发明一个实施例提供的LSTM-attention模型示意图;

[0040] 图4是本发明一个实施例提供的多维图构造示意图;

[0041] 图5是本发明一个实施例提供的多粒度图卷积网络模型示意图。

## 具体实施方式

[0042] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于

本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0043] 应当理解,当在本说明书和所附权利要求书中使用时,术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在,但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0044] 还应当理解,在此本发明说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本发明。如在本发明说明书和所附权利要求书中所使用的那样,除非上下文清楚地指明其它情况,否则单数形式的“一”、“一个”及“该”意在包括复数形式。

[0045] 还应当进一步理解,在本发明说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合,并且包括这些组合。

[0046] 传统的假新闻检测,通过对文本进行预处理(去除停用词、分词),构造其统计特征,如词频、句法特征等,训练分类模型,其中特征表示更影响表现的词在文本分类中的作用更大。通过将新闻文本输入,根据人工选出的特征进行分析检测的方法,需要人工设置特征,费时费力。

[0047] 假新闻自动检测能够缓解该问题,自动检测一般是基于两类特征对假新闻进行检测。一类特征是基于新闻的文本内容,另一类特征是结合新闻的背景辅助信息展开,提升分类效果。

[0048] 请参阅图2,常用于端到端的文本分类模型——卷积神经网络(TextCNN),是将新闻的文本内容等信息,都作为特征输入到卷积神经网络中。文本首先被转化为词向量(即用一个向量来表示文本中的一个词),词向量可以反映词与词之间的关系,随后将这些词向量输入到卷积神经网络中进行训练,在训练中不断修改词向量使它能够更好地表现文本的特征。卷积神经网络通过卷积层池化层对输入的文本背景特征进行提取捕捉文本特征的局部相关性,再经过池化等操作,得到根据输入特征学习得到的表示。该模型通过共享参数来减少参数量、捕捉局部不变形以提高训练性能,但是忽略了新闻之间的关系,也没有办法区分不同特征对判断结果的影响,将各种特征平等地作为判断新闻真假的依据。

[0049] 此外,请参阅图3,使用长短期记忆网络与注意力机制相结合(LSTM-attention)来判断短新闻真假的方法,利用长短期记忆网络对短文本信息进行上下文分析。通过寻找上下文和目标词语的关联来找到新闻文本的特征,并用这些特征来进行判断,随后通过注意力机制对得到的这些特征进行加权处理,对不同的特征赋予不同的注意力权重,突出对判断结果影响大的特征,减小噪声特征对判断结果的影响。注意力机制的核心是在预测新闻类别时所用的上下文的特征是不同的,可以直观地解释各个句子和词对分类类别的重要性。但是该方法没有对新闻之间的相关性进行分析,忽略的新闻之间的关系在判断中的作用。

[0050] 上述监督学习模型,需要人工对训练的样本进行分析,而训练样本越多训练的效果就越好,这就导致了需要大量的人力成本。同时,由于短新闻的文本信息含量少、表达习惯的差异,很难学习到潜在的模式进行分类;此外,只关注新闻本身的一些特征信息,没有将新闻、新闻背景信息(辅助信息)间的相关性结合考虑,因此它的检测准确率不高。

[0051] 研究个体间相关性的模型——图卷积神经网络,通过对由边权和节点(每个节点都有各自的向量表示)构成的图进行卷积操作,可以将每个节点的信息进行传播,使各个节



点的特征融合邻居节点信息,通过多次迭代,节点的信息也可以传播到距离较远的节点,经过该种信息整合方式就可以找到这些节点的相关性,同时也可以对节点自身的特征进行利用分析。最终的目标是学习得到节点(新闻)的表示,并进行分类。

[0052] 请参阅图1。

[0053] 一种基于多粒度的图卷积神经网络的假新闻检测方法,其特征在在于,包括:

[0054] S100、构建不同深度距离的概率转移矩阵,形成多维度的图卷积神经网络,对含有不同节点特征的图卷积神经网络分别进行卷积,得到不同维度的特征在不同节点间的相关性分析结果。

[0055] 请参阅图4,在具体的实施例当中,利用图卷积网络获取多粒度信息,结合节点特征,文本表示和信用历史,对虚假新闻进行分类。一个大型的基准数据集LIAR,根据虚假程度将新闻分成六类:很假,假,半真,较真,大部分真,真。其本质是六分类的文本问题。首先,针对原来的卷积神经网络只能处理单一维度的信息,将原来的图卷积神经网络拓展为多维的图卷积神经网络,为每一种背景特征构建图卷积网络形成多维的图卷积神经网络。

[0056] 具体的,所述每个图卷积神经网络的输入为一个邻接矩阵和一个特征矩阵,所述邻接矩阵根据辅助信息中的每一种关系图构成,所述特征矩阵由文本内容的词向量和信用历史组成。

[0057] 以职业关系为例,如果新闻*i*,*j*的发布者属于同一种职业,就在两个节点间建立边,那么所述邻接矩阵由以下公式确定:

$$[0058] \quad A_{ij} = \begin{cases} 1, & \text{如果新闻 } i, j \text{ 发布者有相同的职业;} \\ 0, & \text{其他} \end{cases};$$

[0059] 图卷积神经网络的分层传播规则为:

$$[0060] \quad H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)});$$

[0061] 上述公式可实现节点间信息的传播,找到节点之间的相关性。其中, $\tilde{A} = A + I$ 是具有自连接的邻接矩阵, $\tilde{D} = \sum_j \tilde{A}_{ij}$ 是自连接邻接矩阵的度矩阵, $H^{(1)}$ 是图卷积中第1层的激活矩阵, $H^{(0)} = X$ 是初始的输入特征矩阵 $X$ , $\sigma(\cdot)$ 是激活函数, $\tilde{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 是单步随机游走的转移概率矩阵,通过对该概率矩阵进行*k*阶近邻处理, $\hat{A}_{ij}$ 是节点*i*和节点*j*进行*k*阶近邻之后的矩阵, $\hat{A}^k = \prod_{i=1}^k \hat{A}_i$ ,通过对单步随机游走的转移概率矩阵进行*k*阶近邻,计算不同距离接近矩阵来描述节点之间的相关性,保留多粒度的信息提高各节点表示的多样性;

[0062] 图卷积神经网络就所述近邻矩阵进行迭代更新的结果为:

$$[0063] \quad Z_k = \hat{A}^k \text{ReLU}(\hat{A}^k X W_k^{(0)}) W_k^{(1)}, k = 1, 2, 3 \dots;$$

[0064] 其中, $W_k^{(0)}$ 是初试输入层的权矩阵, $W_k^{(1)}$ 是第二层权矩阵,根据上述公式得到不同粒度的输出。

[0065] S200、采用注意力机制对所述相关性分析结果进行加权求和,得到不同的注意力权重系数,进而得到新闻真假性判断的结果。

[0066] 在具体的实施例当中,由于不是每一个输出在分辨假新闻中地位都相同,利用注意力机制来将多深度的信息聚合起来,注意力机制从本质上讲和人类的选择性视觉注意力

机制类似,核心目标也是从众多的信息中选出对当前任务目标更加关键的信息。通过注意力模型不但可以把不同的特征整合在一起,还可以通过对不同的特征分配不同的注意力权重来重点突出一些重要的特征。

[0067] 具体的,将每一个输出输入到注意力权重计算中,利用一个非线性函数来计算它们对应的注意力权重 $u_i$  ( $1 \leq i \leq m$ ),

$$[0068] \quad u_i = \tanh(w_i z_i + b_i);$$

[0069] 采用softmax函数对 $u_i$  ( $1 \leq i \leq m$ ) 进行归一化得到:

$$[0070] \quad \alpha_i = \frac{\exp(u_i)}{\sum_{i=1}^m \exp(u_i)};$$

[0071] 其中, $\alpha_i$ 取值范围在0到1之间;

[0072] 利用求得的注意力权重 $u_i$ 对迭代更新的结果 $z_k$ 进行加权求和,得到一个新闻的判断表示 $p_j$ ,

$$[0073] \quad p_j = \sum_{i=1}^m \alpha_i z_i;$$

[0074] 利用交叉熵来评估训练得到的判断和标记样本实际结果之间的偏差,损失函数为:

$$[0075] \quad L = - \sum_{l \in Y_L} \sum_{f=1}^c Y_{lf} \ln p_{lf};$$

[0076] 其中, $Y_L$ 是标记样本分类标签的集合, $c$ 是最终表示的维数,目标是要最小化交叉熵损失函数 $L$ ,通过反向传播不断修改权重使损失变小。

[0077] 请参阅图5,整个处理流程为先构建多粒度图卷积网络,对网络进行卷积,利用注意力机制将各个深度的网络结合起来,得到最后的分类结果。

[0078] 本发明实施例通过研究新闻在各个维度的相似性,提出多粒度图卷积神经网络模型以提高检测效果。具体地,该模型通过利用辅助信息构造新闻间在各个维度的关系图,利用图卷积学习节点信息在不同节点之间传播迭代的特点,找到新闻与新闻之间的相关性,将新闻之间的关联加入到特征之中,并利用多深度的图卷积块捕获邻居的多粒度信息,随后通过注意机制进行信息整合。在一个公开的基准假新闻数据集上的实验结果表明,所提出的模型比最新的五种方法取得了更好的表现。

[0079] 与卷积神经网络、基于注意力机制的长短期记忆网络模型相比,显然解决了这几个方法没有考虑将新闻相关性考虑到判断中的缺点。对于卷积神经网络平等利用各种背景信息导致产生噪声影响结果的缺点,多粒度图卷积网络模型改变了利用辅助信息的方式,并加入注意力机制将所有输出结果进行加权结合,让新闻文本内容、新闻背景内容以及新闻之间的关系有效融合,再利用最终的样本表示进行分类,有效利用了影响分类表现的重要特征,减小了无用特征的噪声影响。

[0080] 而对于图卷积网络只能处理单一维度的特征信息,并且存在过度平滑的问题,多粒度图卷积神经网络针对不同维度的关系进行多维拓展,处理不同维度的特征信息,并且构建了多粒度的图卷积神经网络,在学习新闻间关系的同时显式地把各个距离的邻居信息结合,随后利用注意力机制将特征信息有效结合,保证了短新闻特征的多样性和有效性,从而更好地对假新闻进行区分,提升了分类表现。

[0081] 本发明实施例还提供一种基于多粒度的图卷积神经网络的假新闻检测系统,包括:

[0082] 相关性分析单元,用于构建不同深度距离的概率转移矩阵,形成多维度的图卷积神经网络,对含有不同节点特征的图卷积神经网络分别进行卷积,得到不同维度的特征在不同节点间的相关性分析结果;

[0083] 请参阅图4,在具体的实施例当中,利用图卷积网络获取多粒度信息,结合节点特征,文本表示和信用历史,对虚假新闻进行分类。一个大型的基准数据集LIAR,根据虚假程度将新闻分成六类:很假,假,半真,较真,大部分真,真。其本质是六分类的文本问题。首先,针对原来的卷积神经网络只能处理单一维度的信息,将原来的图卷积神经网络拓展为多维的图卷积神经网络,为每一种背景特征构建图卷积网络形成多维的图卷积神经网络。

[0084] 具体的,所述每个图卷积神经网络的输入为一个邻接矩阵和一个特征矩阵,所述邻接矩阵根据辅助信息中的每一种关系图构成,所述特征矩阵由文本内容的词向量和信用历史组成。

[0085] 以职业关系为例,如果新闻*i*,*j*的发布者属于同一种职业,就在两个节点间建立边,那么所述邻接矩阵由以下公式确定:

$$[0086] \quad A_{ij} = \begin{cases} 1, & \text{如果新闻 } i, j \text{ 发布者有相同的职业;} \\ 0, & \text{其他} \end{cases};$$

[0087] 图卷积神经网络的分层传播规则为:

$$[0088] \quad H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)});$$

[0089] 上述公式可实现节点间信息的传播,找到节点之间的相关性。其中, $\tilde{A} = A + I$ 是具有自连接的邻接矩阵, $\tilde{D} = \sum_j \tilde{A}_{ij}$ 是自连接邻接矩阵的度矩阵, $H^{(1)}$ 是图卷积中第1层的激活矩阵, $H^{(0)} = X$ 是初始的输入特征矩阵 $X$ , $\sigma(\cdot)$ 是激活函数, $\tilde{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 是单步随机游走的转移概率矩阵,通过对该概率矩阵进行*k*阶近邻处理, $\hat{A}_{ij}$ 是节点*i*和节点*j*进行*k*阶近邻之后的矩阵, $\hat{A}^k = \prod_{i=1}^k \hat{A}_i$ ,通过对单步随机游走的转移概率矩阵进行*k*阶近邻,计算不同距离接近矩阵来描述节点之间的相关性,保留多粒度的信息提高各节点表示的多样性;

[0090] 图卷积神经网络就所述近邻矩阵进行迭代更新的结果为:

$$[0091] \quad Z_k = \hat{A}^k \text{ReLU}(\hat{A}^k X W_k^{(0)}) W_k^{(1)}, k = 1, 2, 3, \dots;$$

[0092] 其中, $W_k^{(0)}$ 是初试输入层的权矩阵, $W_k^{(1)}$ 是第二层权矩阵,根据上述公式得到不同粒度的输出。

[0093] 注意力机制融合单元,用于采用注意力机制对所述相关性分析结果进行加权求和,得到不同的注意力权重系数,进而得到新闻真假性判断的结果。

[0094] 在具体的实施例当中,由于不是每一个输出在分辨假新闻中地位都相同,利用注意力机制来将多深度的信息聚合起来,注意力机制从本质上讲和人类的选择性视觉注意力机制类似,核心目标也是从众多的信息中选出对当前任务目标更加关键的信息。通过注意力模型不但可以把不同的特征整合在一起,还可以通过对不同的特征分配不同的注意力权重来重点突出一些重要的特征。

[0095] 具体的,将每一个输出输入到注意力权重计算中,利用一个非线性函数来计算它们对应的注意力权重 $u_i$  ( $1 \leq i \leq m$ ),

[0096]  $u_i = \tanh(w_i z_i + b_i)$ ;

[0097] 采用softmax函数对 $u_i$  ( $1 \leq i \leq m$ ) 进行归一化得到:

[0098] 
$$\alpha_i = \frac{\exp(u_i)}{\sum_{i=1}^m \exp(u_i)}$$
;

[0099] 其中, $\alpha_i$ 取值范围在0到1之间;

[0100] 利用求得的注意力权重 $u_i$ 对迭代更新的结果 $z_k$ 进行加权求和,得到一个新闻的判断表示 $p_j$ ,

[0101] 
$$p_j = \sum_{i=1}^m \alpha_i z_i$$
;

[0102] 利用交叉熵来评估训练得到的判断和标记样本实际结果之间的偏差,损失函数为:

[0103] 
$$L = - \sum_{l \in Y_L} \sum_{f=1}^c Y_{lf} \ln p_{lf}$$
;

[0104] 其中, $Y_L$ 是标记样本分类标签的集合, $c$ 是最终表示的维数,目标是要最小化交叉熵损失函数 $L$ ,通过反向传播不断修改权重使损失变小。

[0105] 请参阅图5,整个处理流程为先构建多粒度图卷积网络,对网络进行卷积,利用注意力机制将各个深度的网络结合起来,得到最后的分类结果。

[0106] 在假新闻分类过程中,要尽可能地将所有对分类有效的特征都利用到分类中,使分类效果尽可能好。现有的技术如传统的文本分类方法、基于神经网络的方法:LSTM-attention模型,卷积网络模型等,都没有利用到新闻与新闻之间的联系。直观地看,假新闻之间一定存在一些相似性,可以利用这些相似性来对新闻的真假性进行判断,多粒度图卷积神经网络模型采用了图卷积神经网络学习不同距离节点间的信息,通过迭代变换传播捕捉新闻间的相关性,相比现有技术仅仅采用新闻本身的特征作为分类依据有较大提高。

[0107] 同时多粒度图卷积网络采用构建多个不同深度距离的概率转移矩阵,得到不同特征下各个节点的相关性信息,可以将新闻的不同特征之间的相关性都挖掘出来,与单一图卷积神经网络相比,改善了单一图卷积神经网络没有完全利用到所有的新闻背景信息的缺陷。

[0108] 相比于卷积神经网络以及记忆网络都对新闻词向量的特征进行同等的处理,没有区分不同特征的重要性,有较大的噪声影响判断结果,多粒度的图卷积神经网络采用注意力机制,其核心是在预测新闻类别时所用的上下文的特征是不同的,可以直观地解释各个句子和词对分类类别的重要性。通过使用注意力机制将不同的多个维度的图卷积神经网络结果进行融合,充分考虑到了新闻之间的联系以及新闻本身的各种特征。

[0109] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access

Memory, RAM) 等。

[0110] 以上所述是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也视为本发明的保护范围。

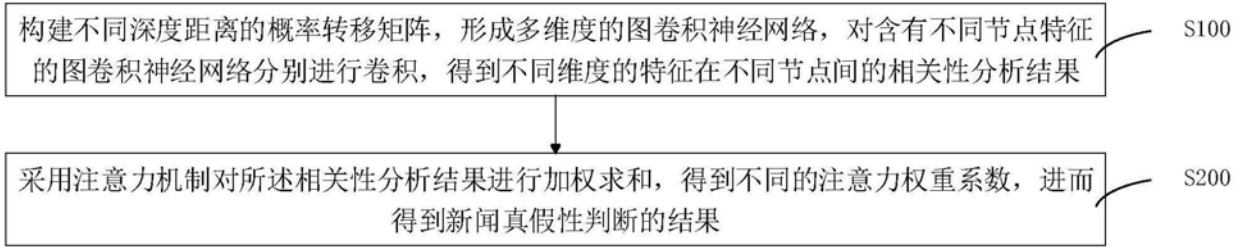


图1

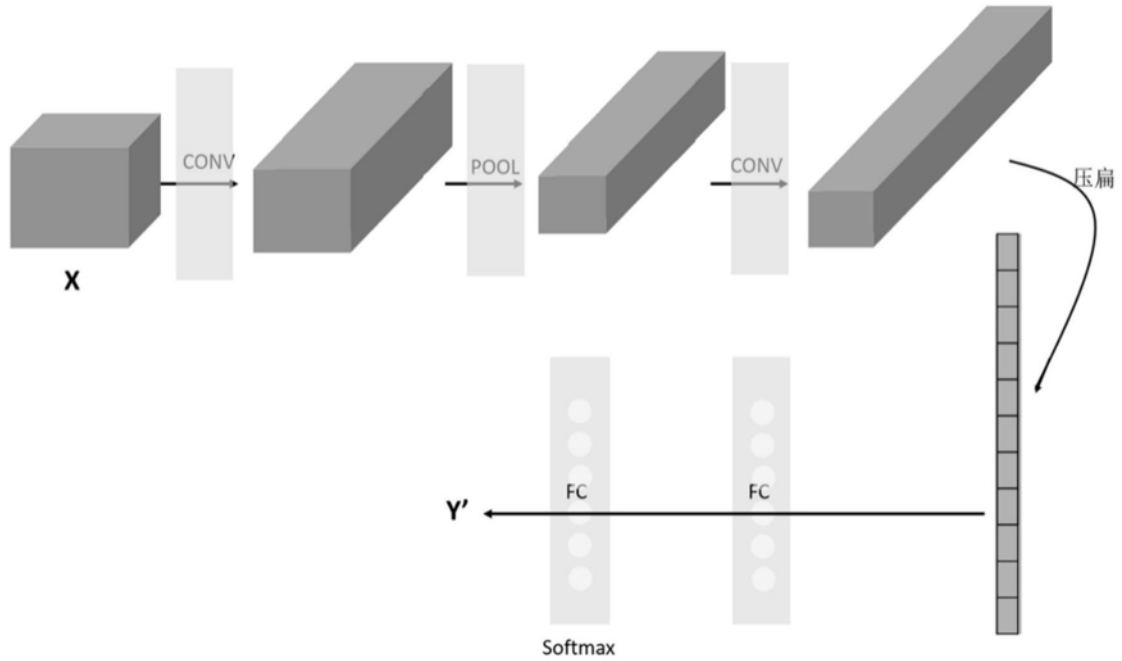


图2

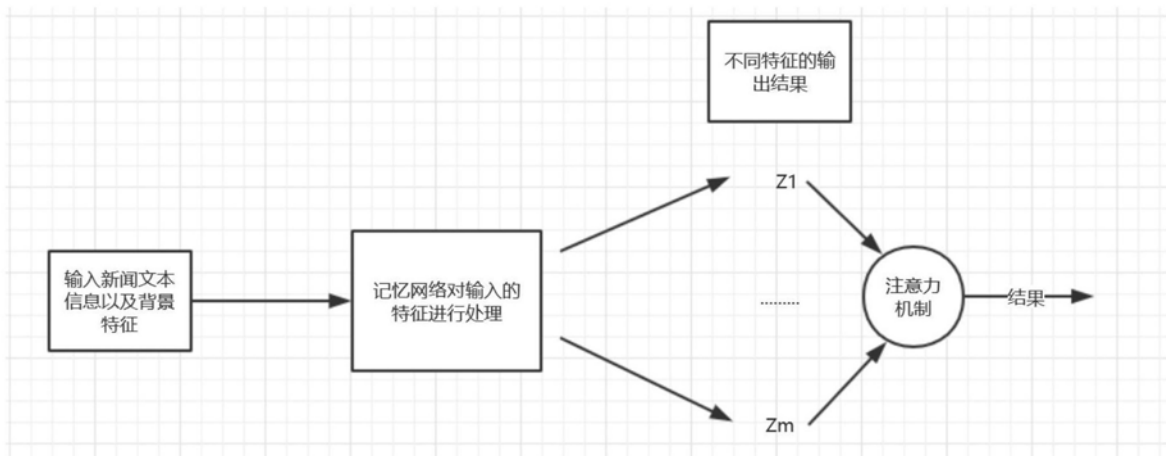


图3

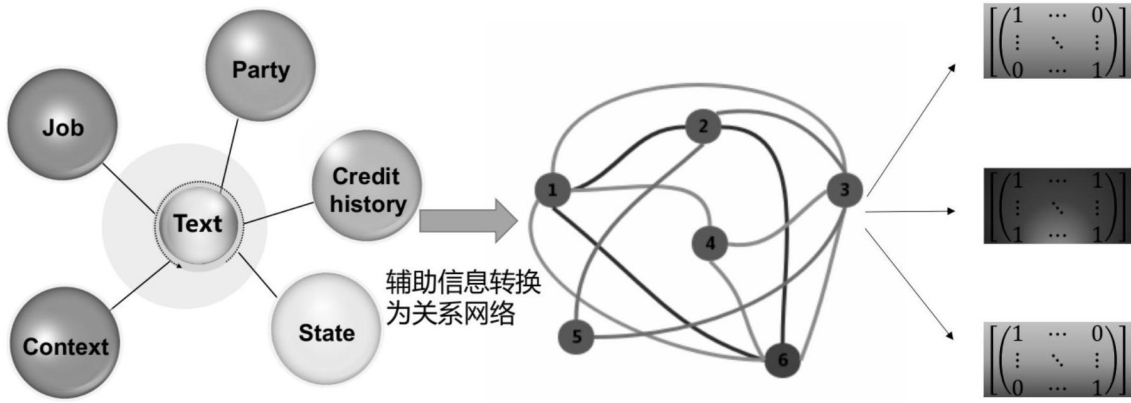


图4

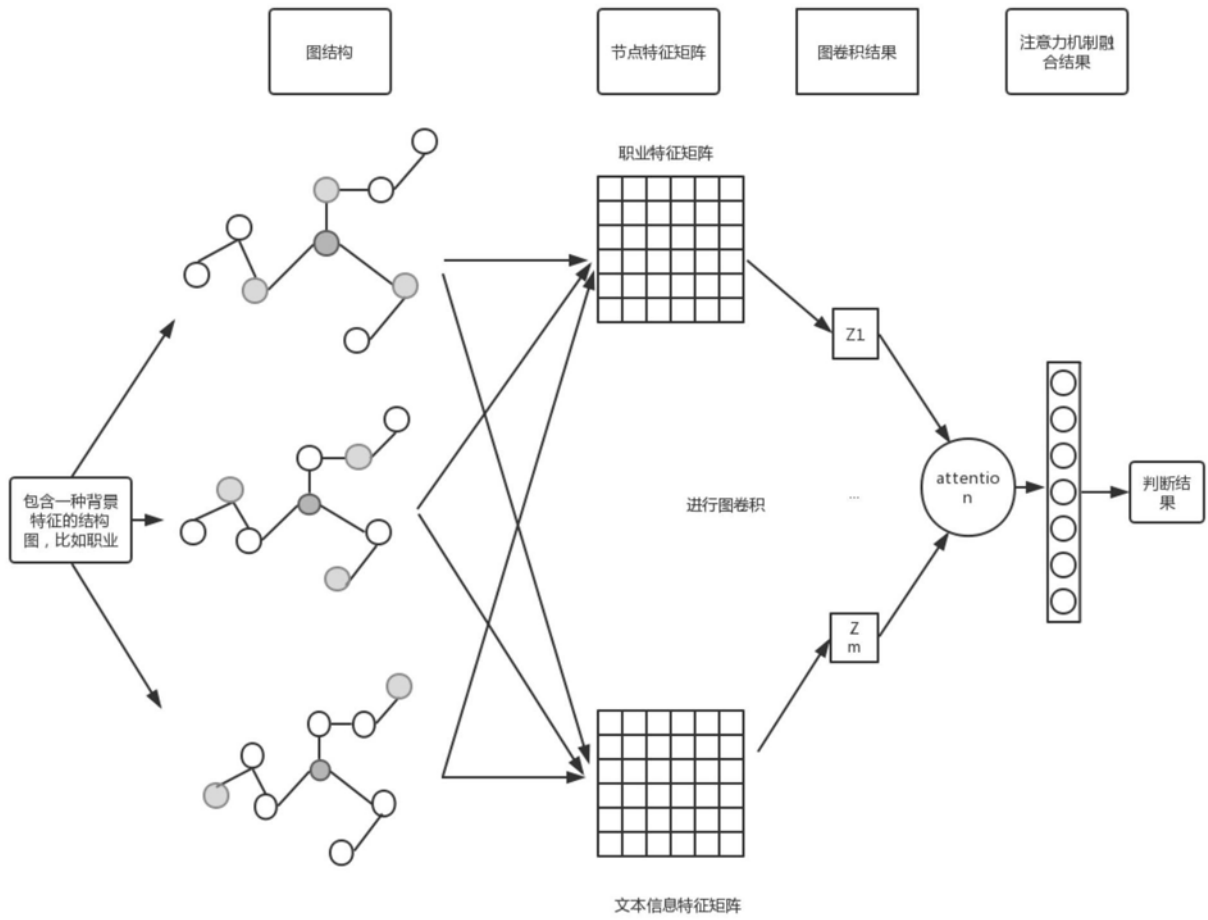


图5