基于堆叠降噪自编码器的异质网络的层次构建与节点分类

蒋宗礼, 张津丽, 杜永萍, 王光亮
(北京工业大学信息学部, 北京 100124)

摘 要: 针对传统特征抽取方法不能很好解决含有丰富语义信息和复杂网络结构的异质网的数据稀疏和噪声问题,利用堆叠降噪自编码器进行特征抽取,有利于松弛策略建立其类别层次结构,完成节点的分类和排序. 在计算机科学文献库(digital bibliography & library project,DBLP)数据集上的实验结果表明:相比于其他分类算法,该方法分类性能更优,精确率可达 86.3%.

关键词: 异质网; 松弛策略; 堆叠降噪自编码器; 层次构建
中图分类号: TP 183 **文献标志码**: A **文章编号**: 0254 - 0037(2018)09 - 1217 - 10
doi: 10.11936/bjutxb2017040032

Hierarchy Construction and Classification of Heterogeneous Information Networks Based on Stacked Denoising Auto Encoder

JIANG Zongli, ZHANG Jinli, DU Yongping, WANG Guangliang
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: The problem of data with noise and sparsity of heterogeneous information networks can not be solved by the traditional feature extraction methods efficiently due to their semantics and complicated structure. Stacked denoising auto encoder was introduced to learn the features of sample. The relax strategy was employed to construct class hierarchy with high-quality, and then the nodes of the heterogeneous information network were classified and ranked. Experimental results on the dataset of DBLP (digital bibliography & library project) show that the method is effective, and the precision of classification is 86.3%.

Key words: heterogeneous information networks; relax strategy; stacked denoising auto encoder; hierarchy construction

随着互联网的普及,现实生活中由不同实体类型和关系组成的信息网络引起了广泛关注[1]. 例如科技文献信息网络[2]、群体知识(维基百科)网络[3]、Facebook 等. 目前大部分工作是基于同质网(即由同一种类型的边和节点组成的信息网络)的研究[4-5]. 相对于同质网,异质网中不同类型的节点和边包含了更加复杂的结构信息以及丰富的语义信息[6],异质网节点的分类等成为新的研究方向.

异质网中物体的分类具有很高的挑战性和复杂性. 由于不同的物体类型和物体之间复杂的关系,一种物体类型的标签信息不应该应用于另外

收稿日期: 2017-04-20
基金项目: 国家科技支撑计划子课题资助项目(2013BAH21B02-01);北京市自然科学基金资助项目(4153058);上海市智能信息处理重点实验室开放基金资助项目(IIPL-2014-004)
作者简介: 蒋宗礼(1956—),男,教授,主要从事网络信息搜索与处理、信息检索方面的研究,E-mail: jiangzl@bjut.edu.cn

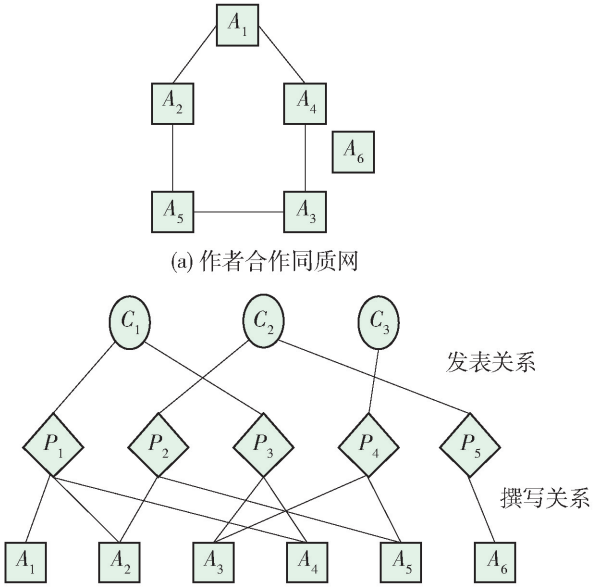
一种物体类型[7]. 例如,在科技文献网络中,用于分类的作者标签集合和文章标签集合是不同的. 除此以外,异质数据的一些特征,比如网络结构的复杂性、特征的缺乏、物体标签的匮乏,都增加了异质网络物体分类的困难. 传统的分类方法利用物体的局部特征和物体的属性进行监督学习. 在异质网络中,没有这种监督学习所需要的局部特征和属性. 如果把异质网络中的链接信息看作物体的属性,那么物体的维度会非常高,而且数据会非常的稀疏[8-9].

集合分类[10-11]是一种常用的网络分类方法,它通过与物体相邻的物体标签对物体进行分类. Zhou 等[12]提出学习局部与全局一致性(learning with local and global consistency, LLGC)方法对同质网进行分类. 该方法通过建立邻接矩阵来反映网络中任意 2 个物体之间的关联程度,在关联矩阵基础上推导出一个分类器. 如果 2 个物体相近,则被给予相同的类标签. LLGC 应用于同质信息网络中,但是可以通过摒弃异质网络中的物体类型,进而应用 LLGC 进行分类. Ji 等[13]提出了一种基于图的正则化框架对异质网进行分类. 该方法先通过每种类型的物体及其类标签生成一个预测函数,然后推导出目标函数,该目标函数使相关联的物体之间的预测值最小,预测标签与实际标签之间的值最小. 通过训练预测函数使得目标函数最优,从而得到基于预测函数的类别. Wan 等[14]提出异质网跳转元路径的概念,即 2 个物体类型相同,并且有相同的类标签,就可以实现跳转. 利用跳转原路径促进主动学习对包含训练标签的异质网中的物体进行分类. Ji 等[15]提出 RankClass 算法,该算法把异质网分类和排序结合起来,排名高的物体在异质网物体的分类中起到重要作用,同时,分类中重要的异质网中的类成员的信息影响排名的质量. 对于每个类标签,该算法都会分配物体一个概率值,通过迭代该概率值,得到每个物体的标签,进而实现对物体的排序.

1 基于堆叠降噪自编码器的异质网节点特征提取

1.1 科技文献信息网

科技文献信息网是一个典型的异质网络结构. DBLP 数据集中有 3 种实体对象:论文(paper)P,作者(author)A,会议(conference)C. 路径表达有 3 种关系:*A-P-A* 表示作者之间的合著关系;*A-P-C-P-A* 表示作者在同一会议发表文章;*A-P-A-P-A* 表示作者之间有相同的合著者. 对每一篇论文 $p_i \in P$,从原数据集中提取该论文所对应的网址,从网上爬下其所对应的摘要,摘要与论文之间存在着一一对应的关系. 利用堆叠降噪自编码器(stacked denoising auto encoder,SDAE)从摘要中提取特征,建立特征矩阵,并使用松弛策略建立层次结构,然后通过论文把与该论文相关的作者和会议关联起来,从而形成一个具有空间层次结构的异质网,如图 1 所示. 图 1(a)为用户关系同质网络. 它的节点是用户类型,边表示用户之间的评论、转发关系. 然而,现实中的大部分网络都是由不同类型的实体和关系组成,很难用同质网进行描述. 图 1(b)为社交异质网络,它包含用户、博文、标签 3 种实体,边用来描述用户和博文间发表、转发、评论关系,以及博文和标签之间的包含关系.



(a) 作者合作同质网

(b) 文献信息异质网

图 1 基于科技文献信息的同质网和异质网

Fig. 1 Scientific co-authorship network and bibliographic information network

1.2 用 SDAE 算法提取异质网特征

为了解决浅层神经网络提取特征困难、计算量大、目标函数[16]收敛速度慢、容易陷入局部极小点等问题,Hinton 等[17]提出深度学习的概念以及训练策略,继而产生了深度自编码器[18]. 为了使深度自编码器学到的特征能够对抗原始数据的污染、缺失,更具鲁棒性,Vincent 等[19]于 2008 年提出了降噪自编码器(denoising auto encoder, DAE),训练过程如图 2 所示.

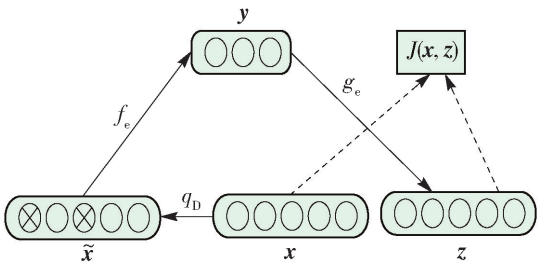


图 2　降噪自编码器训练过程

Fig. 2　Training process of DAE

DAE 由编码器、隐含层和解码器组成[20]. 这里用 $\boldsymbol{x}$ 表示输入向量，$\boldsymbol{y}$ 表示隐藏层向量，$\boldsymbol{z}$ 表示输出层向量. 训练过程如下.

1）通过一个随机的映射变换 q

$$\boldsymbol{x} \rightarrow q_D(\tilde{\boldsymbol{x}} | \boldsymbol{x}) \tag{1}$$

将随机噪声加到输入数据 $\boldsymbol{x}$，得到 $\tilde{\boldsymbol{x}}$. 式中 D 表示数据集.

2）编码器将 $\tilde{\boldsymbol{x}}$ 映射到隐含层，得到

$$\boldsymbol{y} = f_q(\boldsymbol{x}) = s(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b}) \tag{2}$$

该映射的参数集合 $\theta = \{\boldsymbol{W}, \boldsymbol{b}\}$，$\boldsymbol{W}$ 为输入层到隐藏层的连接权值矩阵，$\boldsymbol{b}$ 为隐藏层神经元的偏置向量. s 是一个非线性激活函数，本文中各神经元的激活函数均使用 sigmoid 函数.

3）解码器函数 $g_{\theta'}(\boldsymbol{y})$ 将 $\boldsymbol{y}$ 映射为 $\boldsymbol{z}$，即

$$\boldsymbol{z} = g_{q'}(\boldsymbol{y}) = s(\boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}') \tag{3}$$

该映射的参数集合 $\theta' = \{\boldsymbol{W}', \boldsymbol{b}'\}$，其中：$\boldsymbol{W}' = \boldsymbol{W}^{\mathrm{T}}$，$\boldsymbol{W}'$ 为隐藏层到输出层的连接权值矩阵；$\boldsymbol{b}'$ 为输出层神经元的偏置向量.

4）为使 $\boldsymbol{x}$ 与 $\boldsymbol{z}$ 尽可能接近，设最小化重构误差目标函数为 $J(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{z})$，本文实验采用交叉熵损失函数，即

$$\arg\min_{\theta, \theta'}[J(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{z})] = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{d}[x_{ik}\ln(z_{ik}) + (1 - x_{ik})\ln(1 - z_{ik})] \tag{4}$$

最后，通过梯度下降法不断调整模型的所有参数，从而获得最小重构误差，其中更新规则定义为

$$\boldsymbol{W} = \boldsymbol{W} - \eta\frac{\partial J(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{W}} \tag{5}$$

$$\boldsymbol{b} = \boldsymbol{b} - \eta\frac{\partial J(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{b}} \tag{6}$$

$$\boldsymbol{b}' = \boldsymbol{b}' - \eta\frac{\partial J(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{b}'} \tag{7}$$

式中 η 为学习率.

将多个 DAE 堆叠起来，前一个 DAE 的输出作为后一个 DAE 的输入，迭代更新连接权值矩阵和偏置向量进行预训练. 训练完成后，对整个网络进行微调，形成 SDAE，如图 3 所示.

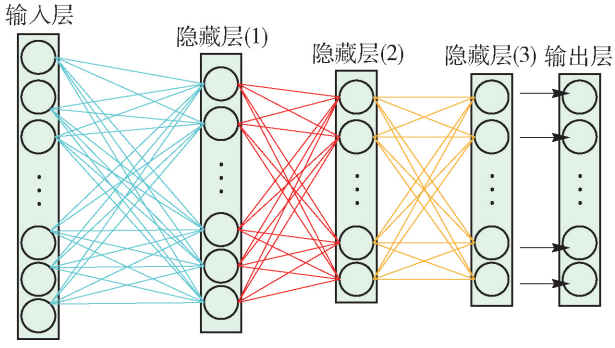


图 3　堆叠降噪自编码器

Fig. 3　Stack denoising auto encoder

构建隐含层层数为 k 的 SDAE 模型，每个隐含层所含神经元个数为：n_1，n_2，$\cdots$，n_k. 每个 DAE 中，$\boldsymbol{W}_y$ 为输入层到隐藏层的连接权值矩阵，$\boldsymbol{W}_z$ 为隐藏层到输出层的连接权值矩阵，$\boldsymbol{B}_y$ 为隐藏层神经元的偏置向量，$\boldsymbol{B}_z$ 为输出层神经元偏置向量. 为了叙述简洁，在算法中用 $\boldsymbol{W}$ 代表 $\boldsymbol{W}_y$ 和 $\boldsymbol{W}_z$. 各神经元的激活函数 s 使用 sigmoid 函数. 异质网节点的样本数据 $X = \{x_1, x_2, \cdots, x_m\}$，$x_i \in X$，为 d 维向量，类别 $L = \{l_1, l_2, \cdots, l_m\}$，$l \in L$ 表示类别，总类别数为 N. cost 为独立样本之间的似然函数. $y^{(1)}, y^{(2)}, \cdots, y^{(m)}$ 为最后一层 DAE 的输出.

异质网特征提取的 SDAE 算法如表 1 所示.

使用 SDAE 算法对异质网的节点进行特征提取，可以对原始节点特征进行多层次的建模训练，从而获得数据更深层次的结构. 同时，对于异质网节点中的噪声数据和稀疏性数据进行了有效处理，提高了模型的泛化能力.

2　基于松弛策略的异质网层次分类

异质网中多种不同类型对象之间蕴藏着大量丰富的信息，为了有效地对这些信息进行搜索与挖掘，需构建其层次分类结构. Wang 等[21]使用文本构建异质网，并对异质网进行递归分割，从而提出基于异质网节点内容的主题层次构建方法. 为了提高异质网节点层次分类的性能，本文采用松弛策略来构建异质网的层次结构. 为了构造出更加合理的网络层次结构，使用松弛策略推迟了异质网节点中不确定

表 1 异质网的 SDAE 算法

Table 1 SDAE algorithm of heterogeneous information networks

SDAE 算法
输入:异质网节点的样本数据个数$\|X\|=m$,类别个数$\|L\|=N$
输出:异质网节点的特征向量和 SDAE 的网络参数值$\theta_1,\theta_2,\cdots,\theta_k$
1 for $i=1$ to k do
1.1 用随机数据初始化各隐含层参数$\theta=\{\boldsymbol{W},\boldsymbol{B}_y,\boldsymbol{B}_z\}$;
2 for $i=1$ to k do
2.1 初始化初值收敛控制参数和学习率η;
2.2 while 不收敛 do
2.2.1 for $j=1$ to m do
2.2.1.1 使用式(1)得到x_j';
2.2.1.2 使用式(2)进行编码计算,得到隐藏层神经元y_j;
2.2.1.3 使用式(3)进行解码计算,得到输出层神经元z_j;
2.2.2 使用式(4)计算损失函数;
2.2.3 使用式(5)~(7),利用随机梯度下降法更新隐含层n_i的参数θ;
2.3 把y_i作为$i+1$层的输入:$x_{i+1}=y_i$;
3 反向微调:
3.1 初值微调学习率α;
3.2 for $i=k$ to 1 do
3.2.1 收敛控制参数;
3.2.2 while 不收敛 do
3.2.2.1 for $j=1$to m do
3.2.2.1.1 计算输入样本属于每个类别的归一化概率;
3.2.2.1.2 计算损失函数;
3.2.2.1.3 更新权值$\boldsymbol{W}$;

类别的判定,直到节点可以明确其所属类别,这样减少了异质网节点类别判定的“阻滞”问题对后续层次结构分类带来的性能影响. 如图 4 所示,将根节点 Node 的类别集合$C=\{A,B,C\}$分割为 2 个子集合C_{L}和C_{R},分别对应节点$\mathrm{Node_R}$和$\mathrm{Node_L}$.

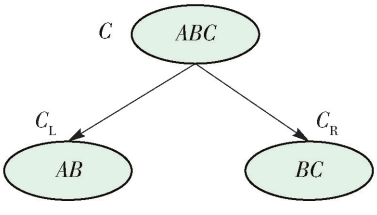


图 4 类别分割示意图

Fig. 4 Class collection partition

由 *K*-means 聚类结果还不能判定类别 *B* 的归属,依据松弛策略将其同时分配给节点$\mathrm{Node_R}$和$\mathrm{Node_L}$.

设某个节点类别集合为 C,c_i为 C 中的一个类别,该节点文本集合用 T 表示,t_i为属于类别 c_i的文本集合,d_{ik}为 t_i中的第 k 篇文本.

$$c_i\in C,i=1,2,\cdots,|C|$$

$$t_i\subseteq T,i=1,2,\cdots,|C|$$

$$d_{ik}\in t_i,i=1,2,\cdots,|C|,k=1,2,\cdots,|t_i|$$

本文对每个节点采用二值聚类,具体的层次结构划分过程如下:

1) 除叶子节点外,将每个类别集合 C 划分为 2 个子集C_{L}和C_{R}. 集合 T 中的文本采用 *K*-means 聚类算法进行二值聚类,为 t_i中的每篇文本 d_{ik}赋值 p_{ik},即

$$p_{ik}(d_{ik})=\begin{cases}1\\-1\end{cases}\tag{8}$$

2) 根据 t_i中的文本的聚类集合决定每个类别 c_i的值 q_i,用来表示 c_i应该属于哪个集合,公式为

$$q_i=\frac{1}{|t_i|}\sum_{k=1,2,\cdots,|t_i|}p_{ik}\tag{9}$$

3) 引入松弛因子 δ,结合 q_i判断 c_i应该属于哪个子集合,公式为

$$\begin{cases} c_i \in C_{\mathrm{L}}, q_i < \delta - 1 \\ c_i \in C_{\mathrm{R}}, q_i > 1 - \delta \\ c_i \in C_{\mathrm{L}} \text{ 且 } c_i \in C_{\mathrm{R}}, \delta - 1 < q_i < 1 - \delta \end{cases} \tag{10}$$

如图 5 所示，δ 越小，被同时分到左右子节点的类别越多.

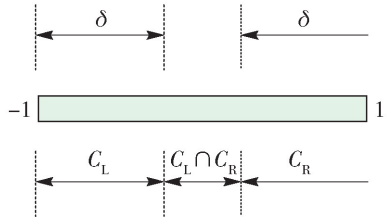


图 5　δ 对类别分割的影响

Fig. 5　Class collection partition influenced by δ

4）递归地将每个节点的类别集合划分为 2 个相交或者不相交的类别子集合，终止条件为：节点的类别集合只包含一个节点或者该类别集合不可再分（即左子节点或右子节点的类别集合和父节点的类别集合完全相同）.

5）类别层次建立后，通过类别集合中的每个节点文本把与该文本相关的作者和会议关联起来，形成有层次结构的异质网，如图 6 所示.

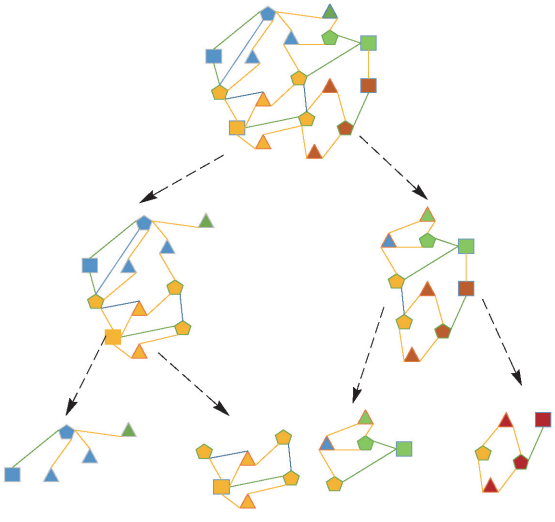


图 6　异质网的层次构建

Fig. 6　Hierarchy constructed by heterogeneous information network

在异质网的类别层次结构中，用每个非叶子节点构造一个二值分类器，类别为 C_{L}-C_{R} 的作为正例文本，类别为 C_{R}-C_{L} 的作为负例文本，类别为 $C_{\mathrm{L}} \cap C_{\mathrm{R}}$ 的不作为训练文本. 再使用分类器、决策树（decision tree，DT）、k-近邻（k-nearest neighbor，KNN）等为每个文本进行分类，直到文本被分到叶子节点为止. 若叶子节点所对应的子异质网中只包含一个类别，那么该类别就是该文本最终预测类别；若叶子节点所对应的异质网中有多个类别，那么构造一个分类数目大于等于 2 的分类器来进行类别预测.

3　基于迭代的异质网节点的排序

本文是在基于松弛策略建立的异质网的层次结构基础上，对节点内部的数据元素进行迭代排序. 其中，异质网分成不同主题的子网.

本文以 DBLP 数据集作为实验对象，它是由德国特里尔大学的 Michael Ley 建立的计算机领域内的英文文献的集成数据库系统[13]. 首先建立基于 DBLP 数据集的数据模型以及关联矩阵. 排序结构数据模型如图 7 所示，G 由 2 个子网构成：作者之间合作关系网 G_{AA} 和作者与文章的关系网 G_{AP}. G 中两关系网分别对应作者-作者关系矩阵 $\boldsymbol{M}_{AA}$ 和作者-文章关系矩阵 $\boldsymbol{M}_{AP}$，$\boldsymbol{M}_{AA}(i,j)$ = 作者 i 与 j 合作文章的篇数，即

$$\boldsymbol{M}_{AP}(i,j) = \begin{cases} 1, & i \text{ 是 } j \text{ 的作者} \\ 0, & \text{其他} \end{cases} \tag{11}$$

式中 $\boldsymbol{M}_{PA}$ 为 $\boldsymbol{M}_{AP}$ 的转置矩阵.

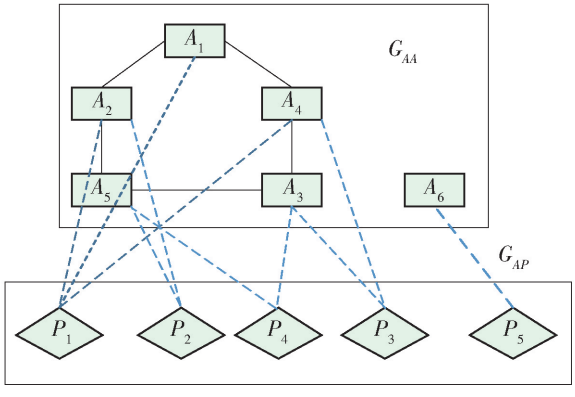


图 7　排序结构数据模型

Fig. 7　Ranking of data model

根据作者影响力，基于如下 3 条规则进行排序.

规则 1　作者的排序越高，其合作的作者的排序越高.

$$\mathrm{Rank}_A(k) = \sum_{r=1}^{n} \boldsymbol{M}_{AA}(k,r)\mathrm{Rank}_A(r)a + \mathrm{Rank}_A(k)(1-a) \tag{12}$$

规则 2　作者的排序越高，其所著文章的排序越高.

$$\mathrm{Rank}_P(j) = \sum_{r=1}^{n} \boldsymbol{M}_{PA}(j,k)\mathrm{Rank}_A(k) \tag{13}$$

规则 3　文章的排序越高，其作者的排序越高.

$$\mathrm{Rank}_A(r) = \sum_{r=1}^{n} \boldsymbol{M}_{AP}(r,j)\,\mathrm{Rank}_P(j)\,b + \mathrm{Rank}_A(r)(1-b) \quad (14)$$

式中:$\mathrm{Rank}_A(k)$为第 k 个作者的排序值;$\mathrm{Rank}_P(k)$为第 k 篇文章的排序值. 每个作者的 Rank 值包含 2 部分:一部分是通过规则计算出的 Rank 值,另一部分是通过相关联的规则计算出的作者本身的 Rank 值. 参数 α 和 β 表示通过上一条规则计算出的作者自身的 rank 值在这一条规则的 Rank 值计算中所占的权重,取值为 0 ~ 1.

根据上述规则,作者的排序是根据文章和其合作作者确定的,而文章又是反向根据作者排序的. 作者 A 与文章 P 之间相互影响,每一条排序规则都把异质网节点之间的相互影响关系迭代到了权值的计算中.

在排序的过程中,先根据异质网中对象之间的关联关系初始化矩阵 $\boldsymbol{M}_{AA}$ 和 $\boldsymbol{M}_{AP}$. 矩阵 $\boldsymbol{M}_{AA}$ 中的元素 m_{ij} 表示作者 i 和作者 j 之间的合作次数,矩阵 $\boldsymbol{M}_{AP}$ 中的元素 m_{ip} 表示作者 i 发表文章 p. 初始值对最后的排序结果影响并不大,只是对排序迭代收敛的速度有一定影响. 使用规则 1 ~ 3 进行迭代排序,在迭代过程中,设置一个阈值 δ,当 2 次迭代结果之间差值小于 δ 时,迭代结束. 计算 2 次迭代结果之间差值的公式为

$$D(t,t+1) = \frac{\sum_{i=1}^{|V|} R(i,t+1) - R(i,t)}{|V|} \quad (15)$$

式中:$|V|$ 为异质网中作者的数目;$R(i,t+1)$ 和 $R(i,t)$ 分别为作者 i 在第 $t+1$ 次和第 t 次迭代时的排序值.

4 实验与结果分析

4.1 实验数据集

DBLP 按年代列出了作者的科研成果,包括国际期刊和会议等公开发表的论文. 实验选取其中数据库(database,DB)、数据挖掘(data ming,DM)、信息检索(information retrieval,IR)和人工智能(artificial intelligence,AI)4 个领域中 2012—2016 年的 14 个会议、5 134 个作者、5 516 篇文章,组成 3 种类型的异质网. 从文献的相关元数据中提取出文章链接,爬取文章摘要作为内容分析对象,见表 2.

表 2 DBLP 中 4 个不同领域的数据集分布

Table 2 Four different datasets of DBLP

领域	类别数目	类别集合	文本数目
DB	3	VLDB	94
		ICDE	300
		DASFAA	189
DM	4	ICDM	474
		SDM	192
		PKDD	240
		PAKDD	210
IR	2	ECIR	199
		WWW	64
AI	5	AAAI	461
		CVPR	1042
		ICML	691
		IJCAI	461
		NIPS	899

4.2 实验结果及分析

4.2.1 特征数量和分类方法对性能的影响

不同的特征数目以及不同的特征选择方法会影响分类准确率. 一般情况下,更多的特征数目会提高算法分类准确精度,但是过多特征数目会导致信息冗余,分类效率降低. 同样,不同的特征选择方法也会对算法的分类性能产生影响. 本文采用 2 折交叉验证把数据集分成测试集(2 755 篇)和训练集(2 761 篇),然后分别采用卡方 χ^2、信息增益(information gain,IG)、词频-逆文件频率(term frequency-inverse document frequency,TF-IDF)以及词频-逆文件频率与信息增益相结合(term frequency-inverse document frequency & information gain,TF-IDF & IG)4 种特征选择方法在不同的特征数目下进行实验,分类方法采用 softmax 方法. 实验结果如图 8 所示.

从实验结果可以看出,特征数量为 2 500 时,4 种特征选择方法具有最高的分类准确率. 特征数量从 2 000 增加到 2 500 后,算法分类准确率提升,而特征数量从 2 500 增加到 2 800 后,算法分类准确率略有下降,可以近似认为特征数量取 2 500 时,算法具有最高的分类准确率. 对于数据集 DBLP 而言,特征选择方法 IG 和 TF-IDF 均高于 χ^2,当选用 TF-IDF&IG时,分类准确率最高,达到75.3%. 在之后

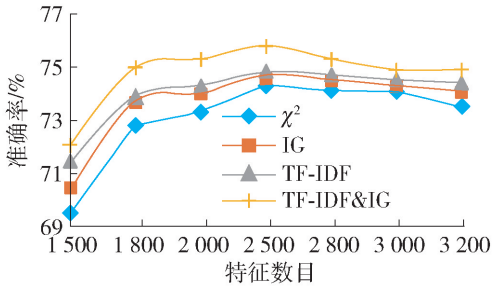


图 8　特征数目和不同特征抽取方法下的分类性能

Fig. 8　Performance influenced by different feature selection method and feature number

的实验中，选取 2 500 作为数据集 DBLP 的特征数量，特征计算方法选择 TF-IDF&IG.

4.2.2　层次结构构建结果

图 9 是本实验得到的层次结构. 由于类别之间比较相近，$\partial = 0.45$，步长为 0.05. 当类别之间的相似度很高时，算法不会进行分割，直到 2 个类别之间相似度较小时，算法才会进行分割，因此，使用松弛策略构造出的层次结构是一个有向无环图. 如果叶子节点中包含多个类别，则采用多值分类器进行划分.

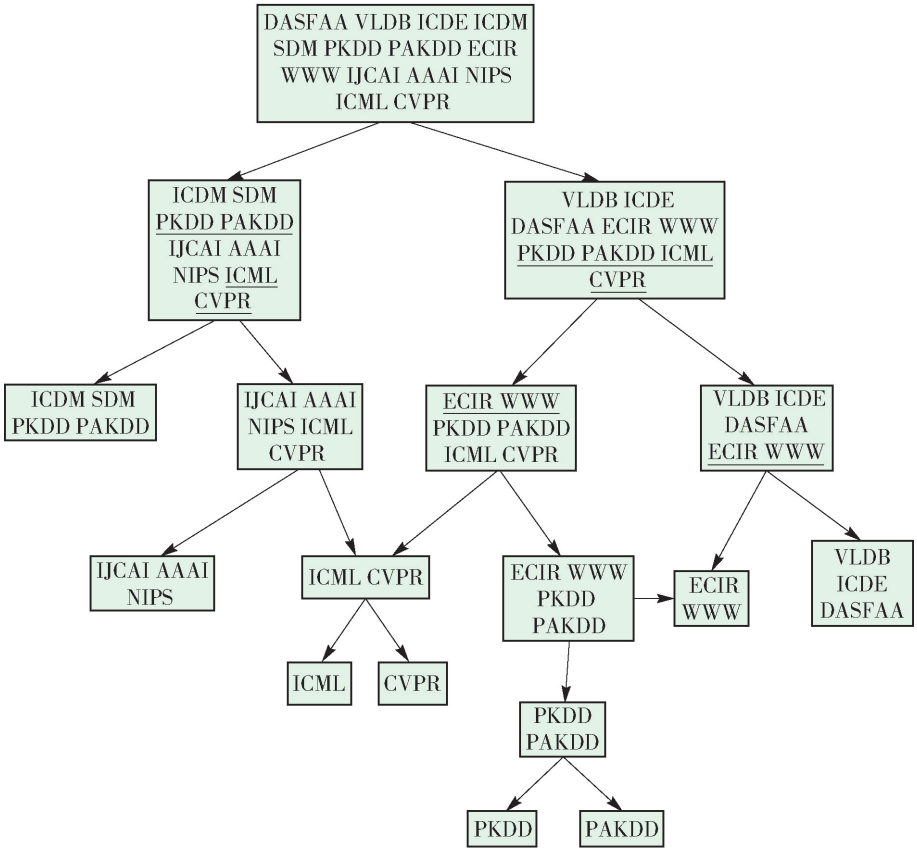


图 9　DBLP 数据集中文章摘要构造出的层次结构

Fig. 9　Hierarchy constructed by abstract of DBLP

4.2.3　基于堆叠降噪自编码器与松弛策略的异质网络分类性能评测

构建 SDAE 模型，各层神经元节点个数为：2 000、1 500、1 200. 在图 10 给出的松弛策略构建的类别层次基础上，分别使用 DT、KNN、支持向量机（support vector machine，SVM）、朴素贝叶斯（native Bayes，NB）以及随机森林（random forest，RF）5 种算法与 SDAE 算法做对比. 本文采用精确率（precision）、召回率（recall）和 F 值作为分类算法性能评价指标. 这 5 种算法都使用 TF-IDF&IG 选取特征，在不同特征数目下的结果如图 10（a）~（f）所示. 实验结果表明，使用松弛策略对异质网建立分类层次结构后，在特征数量分别为 1 500、1 800、2 000、2 500、2 800、3 000、3 200 时，SDAE 的性能要优于其他的分类器.

从图 11（a）~（c）的结果中可以看出，本文选取特征数目分别为 2 000、2 200、2 500、2 800、3 000，使用 SDAE 提取出的特征，相比于使用 TF-IDF&IG 提取的特征的准确率、召回率、F 值都提高显著，准确率最高达到 86.3%. 同时做了 t-检验来

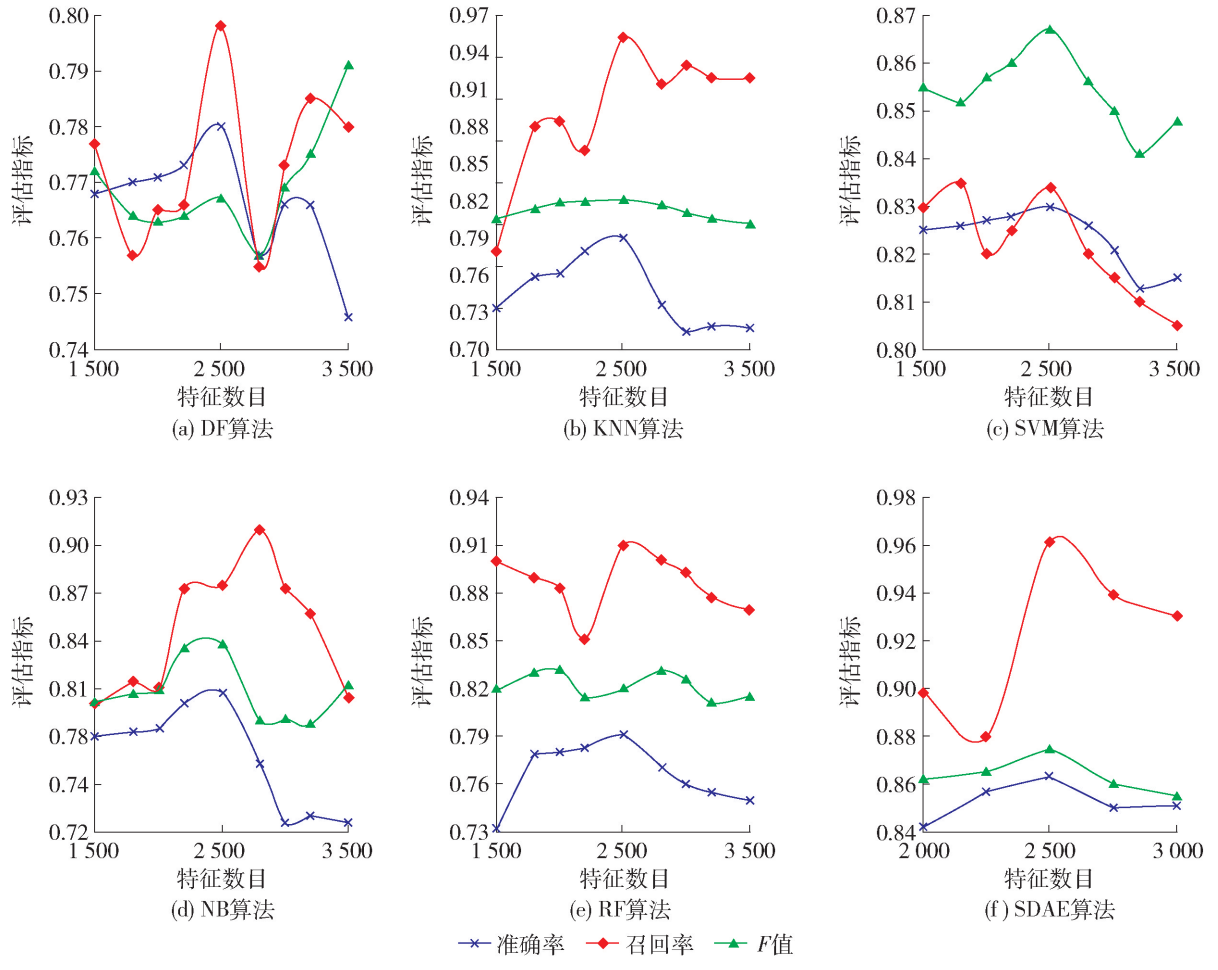


图 10 不同类型的分类器在不同的特征下对分类性能的影响

Fig. 10 Performance influenced by different classifier on different features

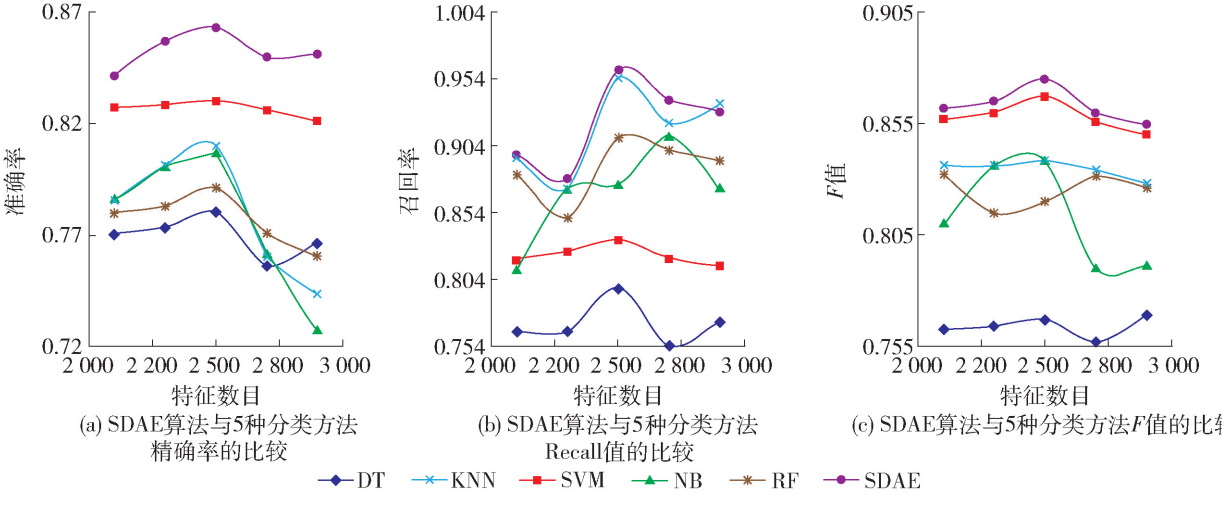


图 11 SDAE 与 5 种分类算法的性能比较

Fig. 11 Performance comparison between SDAE and five different methods

测试该方法性能提高的显著性,如表 3 所示. 5 种分类方法下的 p 值均小于 0.05,SDAE 算法所带来的性能的提高是显著的.

4.2.4 异质网中作者影响力的排序

异质网节点分类完成后,根据式(5)~(7)分别在 DB、DM、IR、AI 四个领域内对异质网内的作

者影响力进行排序，排序结果如表 4 所示. 从表中可以看出，在各个领域内对异质网节点进行排序比不分类直接对异质网节点进行排序更有比较意义.

表 3　测试 SDAE 性能提高显著性的 *t*-检验结果

Table 3　*t*-test of SDAE

成对比较	SDAE&&DT	SDAE&&KNN	SDAE&&SVM	SDAE&&NB	SDAE&&RF
t-检验	1.860	2.015	2.015	2.132	1.895
p 值	1.18×10^{-7}	0.001 2	0.005 1	0.003 7	3.48×10^{-6}

表 4　4 个研究领域中的前 5 名作者排序

Table 4　Top-5 authors in four research areas

领域排名	DB	DM	IR	AI
作者	Kai Zheng	Abdul Quamar	Lei Chen	Feiping Nie
	Lei Chen	Jeffrey Xu Yu	Alex Q Chen	Yuhong Guo
	Quoc Viet Hung Nguyen	StevenEuijong Whang	Jianzhong Qi	Ping Li
	Mohamed F Mokbel	Rene Mueller	Weixiong Rao	Chris H Q Ding
	Martin Faust	QuocTrung Tran	Hyuk-Yoon Kwon	Stefano Ermon

5　结论

1）利用深度学习模型 SDAE 对异质网节点内容进行特征提取，实验结果表明，在相同的分类方法下取得了较高的分类准确率，同时有效缓解了异质网节点内容的数据稀疏问题.

2）在使用 SDAE 提取出的特征的基础上，利用松弛策略思想来构建异质网的类别层次结构，并在该层次结构中对异质网的节点进行分类，获得了较高的分类精确率. 异质网节点分类完成后，本文在类内基于迭代算法对异质网节点进行排序，给出作者影响力的排序结果.

3）在公开数据集 DBLP 上的实验结果表明，基于 SDAE 与松弛策略的异质网络的分类层次结构与其他的分类方法相比有更优秀的表现，在以后的工作中，将基于深度学习的方法对异质网关联分析进行更深入的研究.

参考文献：

[1] SUN Y Z, HAN J W, YAN X, et al. PathSim: meta path-based top-*K* similarity search in heterogeneous information networks[J]. Proceedings of the Vldb Endowment, 2011, 4(11): 992-1003.

[2] SUN Y Z, YU Y, HAN J W. Ranking-based clustering of heterogeneous information networks with star network schema[C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28- July 01, 2009. New York: ACM, 2009: 797-806.

[3] ELLISON N B. Social network sites: definition, history, and scholarship [J]. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230.

[4] VÖLKEL M, KRÖTZSCH M, VRANDECIC D, et al. Semantic Wikipedia [C] // Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, May 23-26, 2006. New York: ACM, 2006: 585-594.

[5] MUKUL G, PRADEEP K, BHARAT B. A new relevance measure for heterogeneous networks [C] // Big Data Analytics and Knowledge Discovery, Valencia, Spain, September1-4, 2015. Berlin: Springer-Verlag, 2015: 165-177.

[6] CAO B, LIU N N, YANG Q. Transfer learning for collective link prediction in multiple heterogeneous domains [C] // Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, June 21-24, 2010. New York: ACM, 2010: 159-166.

[7] ANGELOVA R, KASNECI G, WEIKUM G. Graffiti: graph-based classification in heterogeneous networks [J]. World Wide Web, 15(2): 139-170.

[8] CHEN L, GUAN R, WANG Z, et al. HetPathMine: a novel transductive classification algorithm on heterogeneous information networks [C] // ECIR 2014: Advances in Information Retrieval. Cham: Springer, 2014: 210-221.

[9] ROSSI R, LOPES A, REZENDE S. Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts[J]. Information Processing & Management, 2012, 52(2): 217-257.

[10] PRITHVIRAJ S, GALILEO N, MUSTAFA B, et al. Collective classification in network data articles [J]. Artificial Intelligence Magazine, 2008, 29(3): 93-106.

[11] 施培蓓, 刘贵全, 汪中. 一种基于类别不平衡数据的层次分类模型[J]. 中国科学技术大学学报, 2015, 45(1): 61-68.
SHI P B, LIU G Q, WANG Z. A hierarchical classification model for class-imbalanced data [J]. Journal of University of Science and Technology of China, 2015, 45(1): 61-68. (in Chinese)

[12] ZHOU D Y, OLIVIER B, THOMAS N L, et al. Learning with local and global consistency[C]// Proceedings of the 16th International Conference on Neural Information Processing Systems, Whistler, British Columbia, Canada, December 09-11, 2003. New York: MIT Press Cambridge, 2003: 321-328.

[13] JI M, SUN Y Z, MARINA D, et al. Graph regularized transductive classification on heterogeneous information networks [C] // Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases, Barcelona, Spain, September 20-24, 2010. Berlin: Springer-Verlag, 2010: 570-586.

[14] WAN C, LI X, KAO B, et al. Classification with active learning and meta-paths in heterogeneous information networks[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Melbourne, Australia, October 18-23, 2015. New York: ACM, 2015: 443-452.

[15] JI M, HAN J W, MARINA D. Ranking-based classification of heterogeneous information networks[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, August 21-24, 2011. New York: ACM, 2011: 1298-1306.

[16] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[17] HINTON G E, OSINDER S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[18] 唐朝辉, 朱清新, 洪朝群, 等. 基于自编码器及超图学习的多标签特征提取[J]. 自动化学报, 2016, 42(7): 1014-1021.
TANG C H, ZHU Q X, HONG C Q, et al. Multi-label feature selection with autoencoders and hypergraph learning[J]. Acta Automatica Sinica, 2016, 42(7): 1014-1021. (in Chinese)

[19] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th International Conference on Machine Learning. New York: ACM, 2008: 1096-1103.

[20] 尹宝才, 王文通, 王立春. 深度学习研究综述[J]. 北京工业大学学报, 2015, 41(1): 48-59.
YIN B C, WANG W T, WANG L C. Review of deep learning[J]. Journal of Beijing University of Technology, 2015, 41(1): 48-59. (in Chinese)

[21] WANG C, DANILEVSKY M, LIU J, et al. Constructing topical hierarchies in heterogeneous information networks [J]. Knowledge and Information System, 2015, 44(3): 529-558.

(责任编辑 梁 洁)