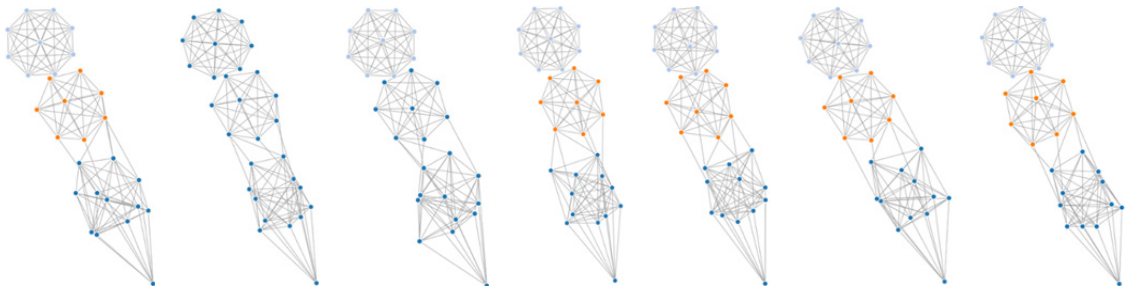

(a) Football 真实社区 (b) Radicchi (c) LPA (d) HANP (e) MB-DSGE (f) gCluSkeleton (g) CoDDA

Fig.8 Comparation of ground-truth and communities from different methods in Football

图 8 Football 数据集中,3 个真实社区的节点使用不同算法得到的社区比较

5.4.3 参数实验

本节分析跳数阈值 S、衰减因子σ和深度稀疏自动编码器的层数对实验结果的影响.比较 CoDDA 算法得到的社区与直接使用 k-均值算法对相似度矩阵聚类得到的社区,来展示 CoDDA 算法中的基于深度稀疏自动编码器的特征提取操作可以有效提高社区结果的准确性.

(1) 跳数阈值 S

针对 Strike 数据集,设置深度稀疏自动编码器各层的节点数是[24-16],衰减因子σ=0.5,分析不同跳数阈值 S 对 NMI 的影响.根据图 9 所示,在跳数阈值 S 的不同取值下,对比直接使用 k-均值算法对相似度矩阵进行聚类,CoDDA 算法得到的结果社区都更为准确.实验结果显示,CoDDA 算法基于深度稀疏自动编码器进行特征提取的操作可以很好地提高社区结果的准确性.

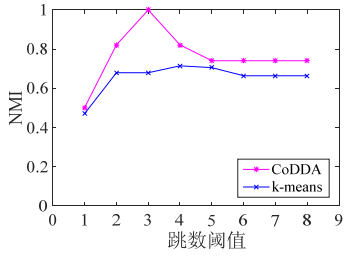

Fig.9 NMI of CoDDA and k-means with different values of S in Strike

图 9 Strike 数据集中不同跳数阈值 S 下,使用 CoDDA 算法与 k-均值算法的 NMI 值

同时,随着跳数阈值S的增加,NMI 呈现先增加后减小的趋势.这样的实验结果说明:考虑不直接相连但在一定跳数内可达的节点对的相似度,可以有效地反映每个节点局部的结构信息.但是如果跳数阈值过大,距离过远不在相同社区的节点之间也增加了一定的相似度值,这不利于社区边界的识别,使得社区准确性降低.对于小规模数据集 Strike 和 Football,分别选择小的跳数阈值 3-跳和 1-跳;对于大规模数据集 LiveJournal 和 Orkut,选择稍大的跳数阈值 8-跳,以达到最优的结果.

(2) 衰减因子σ

针对 Strike 数据集,设置深度稀疏自动编码器每层的节点数[24-16],跳数阈值 S=3,分析不同衰减因子σ的取值对实验结果的影响.根据图 10 所示,相比直接使用 k-均值算法对相似度矩阵的聚类操作,CoDDA 算法在不同衰减因子取值下得到的结果社区更精确.从实验的角度展示了 CoDDA 算法中基于深度稀疏自动编码器进行特征提取的操作在社区准确性方面的贡献.

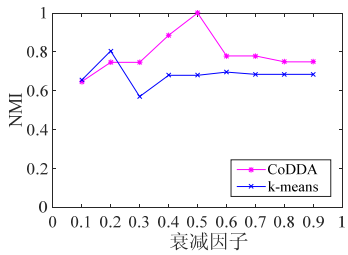

Fig.10 NMI of CoDDA and *k*-means with different values of σ in Strike

图 10 Strike 数据集中不同衰减因子σ下,使用 CoDDA 算法与 *k*-均值算法的 NMI 值

同时,随着衰减因子的增加,NMI 整体呈现先增加后减小的趋势.因为衰减因子控制相似度随着跳数的增加的衰减程度,所以对于小规模数据集 Strike 和 Football,选择稍大的衰减因子σ=0.5,以避免衰减因子过大时对社区边界产生的模糊作用;对于大规模数据集 LiveJournal 和 Orkut,选择小的衰减因子σ=0.1,增强节点的局部特征,以达到最优的结果.

(3) 深度稀疏自动编码器的层数

针对 LiveJournal 数据集,设置跳数阈值 S=8,衰减因子σ=0.1,分析不同层数的深度稀疏自动编码器对实验结果的影响,其中,1~6 层对应的节点数分别为 6 000,4 096,2 048,1 024,512 和 256.

根据图 11 所示,CoDDA 算法在 3 层深度稀疏自动编码器(每层节点是 6000-4096-2048)的设置下性能最好.随着深度稀疏自动编码器层数的增加,社区准确度先增加后减小.实验结果显示:深度稀疏自动编码器这一无监督深度学习方法可以提取出使社区结构更加明显的特征信息,提高结果社区的准确性.但是如果层数过高,会过滤掉一些特征信息,降低结果社区的准确性.因此,对于小规模数据集 Strike 和 Football,分别选择稍低的训练层数[24-16]和[180-128];对于大规模数据集 LiveJournal 和 Orkut,分别选择稍大的训练层数[6000-4096-2048]和[30692-16384-8192-4096-2048],以提供更准确的特征提取结果.

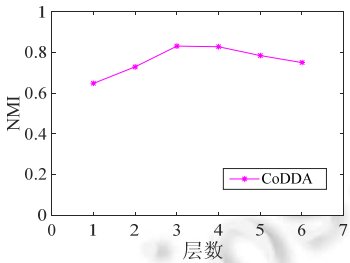

Fig.11 NMI of CoDDA with different numbers of deep spare autoencoder layers in Strike

图 11 Strike 数据集中在不同层数的深度稀疏自动编码器中使用 CoDDA 算法的 NMI 值

6 结束语

本文提出了基于深度稀疏自动编码器的社区发现算法 CoDDA,尝试使用无监督深度学习的方法,根据网络拓扑结构,提高使用例如 *k*-均值等经典的聚类方法进行社区发现的准确性.首先提出了基于 *s*-跳数的邻接矩阵预处理方法,得到能够更好地反映节点局部信息的相似度矩阵;然后构建深度稀疏自动编码器,对相似度矩阵进行深度特征提取;最后,使用 *k*-均值方法聚类得到社区.实验结果显示:与典型的社区发现算法相比,本文提出的 CoDDA 算法可以得到更准确的社区结构.并且,与直接使用高维相似度矩阵的 *k*-均值方法相比,CoDDA 算法得到的社区结果更为准确.

References:

[1] Fortunato S. Community detection in graphs. Physics Reports, 2010,486(3):75−174. [doi: 10.1016/j.physrep.2009.11.002]

[2] Huang FL, Zhang SC, Zhu XF. Discovering network community based on multi-objective optimization. Ruan Jian Xue Bao/Journal of Software, 2013,24(9):2062−2077 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4400.htm [doi: 10.3724/SP.J.1001.2013.04400]

[3] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4):046110. [doi: 10.1103/PhysRevE.78.046110]

[4] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/PhysRevE.69.026113]

[5] Newman MEJ. The structure and function of complex networks. SIAM Review, 2003,45(2):167−256. [doi: 10.1137/S00361445034 2480]

[6] Zhou XP, Liang X, Zhang HY. User community detection on micro-blog using R-C model. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2808−2823 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4720.htm [doi: 10.13328/j.cnki.jos. 004720]

[7] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. Proc. of the National Academy of Sciences of the United States of America, 2004,101(9):2658−2663. [doi: 10.1073/pnas.0400054101]

[8] Chen J, Saad Y. Dense subgraph extraction with application to community detection. IEEE Trans. on Knowledge and Data Engineering, 2012,24(7):1216−1230. [doi: 10.1109/TKDE.2010.271]

[9] Huang J, Sun H, Song Q, Deng H, Han J. Revealing density-based clustering structure from the core-connected tree of a network. IEEE Trans. on Knowledge and Data Engineering, 2013,25(8):1876−1889. [doi: 10.1109/TKDE.2012.100]

[10] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007,76(3):036106. [doi: 10.1103/PhysRevE.76.036106]

[11] Leung IXY, Hui P, Lio P, Crowcroft J. Towards real-time community detection in large networks. Physical Review E, 2009,79(6): 066107. [doi: 10.1103/PhysRevE.79.066107]

[12] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2014. 701−710. [doi: 10.1145/2623330.2623732]

[13] Hartigan JA, Wong MA. Algorithm AS 136: A *k*-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979,28(1):100−108. [doi: 10.2307/2346830]

[14] Tang L, Liu H. Community detection and mining in social media. Synthesis Lectures on Data Mining and Knowledge Discovery, 2010,2(1):1−137.

[15] Gan WY, He N, Li DY, Wang JM. Community discovery method in networks based on topological potential. Ruan Jian Xue Bao/ Journal of Software, 2009,20(8):2241−2254 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]

[16] Yang N, Gong DZ, Li X, Meng XF. Survey of communicaties identification. Journal of Computer Research and Development, 2005, 42(3):439−447 (in Chinese with English abstract).

[17] Wang M, Wang C, Yu JX, Zhang J. Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework. Proc. of the VLDB Endowment, 2015,8(10):998−1009. [doi: 10.14778/2794367.2794370]

[18] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: 10.1103/PhysRevE.69.066133]

[19] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6):066111. [doi: 10.1103/PhysRevE.70.066111]

[20] Tsourakakis C, Bonchi F, Gionis A, Gullo F, Tsiarli M. Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2013. 104−112. [doi: 10.1145/2487575.2487645]

[21] Xie J, Szymanski BK, Liu X. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: Proc. of the 2011 IEEE 11th Int'l Conf. on Data Mining Workshops. IEEE, 2011. 344−349. [doi: 10.1109/ICDMW.2011.154]

[22] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003,15(6): 1373−1396. [doi: 10.1162/089976603321780317]

[23] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In: Proc. of the 24th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2015. 891−900. [doi: 10.1145/2806416.2806512]

[24] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015,61:85−117. [doi: 10.1016/j.neunet.2014.09. 003]

[25] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(8):1798−1828. [doi: 10.1109/TPAMI.2013.50]

[26] Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Trans. on Signal and Information Processing, 2014,3:e2. [doi: 10.1017/atsip.2013.9]

[27] Zhang X, Gao Y. Face recognition across pose: A review. Pattern Recognition, 2009,42(11):2876−2896. [doi: 10.1016/j.patcog. 2009.04.017]

[28] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009,2(1):1−127. [doi: 10.1561/2200 000006]

[29] Deng L, Yu D. Deep learning: Methods and applications. Foundations and Trends in Signal Processing, 2014,7(3-4):197−387. [doi: 10.1561/2000000039]
[30] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015,521(7553):436−444. [doi: 10.1038/nature14539]
[31] Le QV, Ngiam J, Coates A, Lahiri A, Prochnow BY, Ng AY. On optimization methods for deep learning. In: Proc. of the 28th Int'l Conf. on Machine Learning (ICML-11). 2011. 265−272.
[32] Salakhutdinov R, Hinton G. Deep Boltzmann Machines. In: Proc. of the 12th Int'l Conf. on Artificial Intelligence and Statistics (AISTATS). 2009.
[33] Ranzato MA, Boureau YL, LeCun Y. Sparse feature learning for deep belief networks. Advances in neural information processing systems, 2008. 1185−1192.
[34] Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 2003,16(5):555−559. [doi: 10.1016/S0893-6080(03)00115-1]
[35] Mikolov T, Karafiát M, Burget L, Khudanpur S. Recurrent neural network based language model. Interspeech, 2010,2:3.
[36] Funahashi K, Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks. Neural Networks, 1993,6(6):801−806. [doi: 10.1016/S0893-6080(05)80125-X]
[37] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18(7):1527−1554. [doi: 10.1162/neco.2006.18.7.1527]
[38] Bengio Y, Lamblin P, Popovici D, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems, 2007,19:153.
[39] Lee H, Grosse R, Ranganath R, Ng AY. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. ACM, 2009. 609−616. [doi: 10.1145/1553374.1553453]
[40] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11):2278−2324. [doi: 10.1109/5.726791]
[41] Shin HC, Orton MR, Collins DJ, Doran S, Leach M. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(8):1930−1943. [doi: 10.1109/TPAMI.2012.277]
[42] Khorasgani RR, Chen J, Zaïane OR. Top leaders community detection approach in information networks. In: Proc. of the 4th SNA-KDD Workshop on Social Network Mining and Analysis. 2010.
[43] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems, 2015,42(1):181−213. [doi: 10.1007/s10115-013-0693-z]
[44] Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005,2005(9):P09008. [doi: 10.1088/1742-5468/2005/09/P09008]

附中文参考文献:

[2] 黄发良,张师超,朱晓峰.基于多目标优化的网络社区发现方法.软件学报,2013,24(9):2062−2077. http://www.jos.org.cn/1000-9825/4400.htm [doi: 10.3724/SP.J.1001.2013.04400]
[6] 周小平,梁循,张海燕.基于 R-C 模型的微博用户社区发现.软件学报,2014,25(12):2808−2823. http://www.jos.org.cn/1000-9825/4720.htm [doi: 10.13328/j.cnki.jos.004720]
[15] 淦文燕,赫南,李德毅,王建民.一种基于拓扑势的网络社区发现方法.软件学报,2009,20(8):2241−2254. http://www.jos.org.cn/1000-9825/3318.htm [doi: 10.3724/SP.J.1001.2009.03318]
[16] 杨楠,弓丹志,李忱,孟小峰.Web 社区发现技术综述.计算机研究与发展,2005,42(3):439−447.

尚敬文(1993－),女,山东济南人,硕士,主要研究领域为社交网络,社区发现.

辛欣(1990－),女,硕士,主要研究领域为机器学习,数据挖掘,社交网络,深度学习.

王朝坤(1976－),男,博士,副教授,博士生导师,CCF 会员,主要研究领域为图和社交数据管理,音乐计算,大数据系统.

应翔(1992－),男,硕士,主要研究领域为社区发现.