工程应用

doi: 10.3969/j.issn.1673-5692.2020.04.014

基于集成聚类的社区发现算法

江逸楠[1]，刘家琛[1,2]，王亚珅[1]，张欣海[1]，张　博[1]

（1. 中国电子科学研究院，北京　100041；
2. 西安电子科技大学，陕西 西安　710071）

摘　要： 社区发现对于理解复杂系统的整体组织及其功能特性具有重要意义，在个性化服务、广告营销、舆情传播甚至犯罪团伙发现等领域具有广泛的应用场景。近年来，在机器学习聚类问题方面获得较好效果的集成聚类被引入社区发现问题的研究中，以提升社区发现的精确度和稳定性。本文通过将集成方法、集成聚类选择与社区发现相结合，提出了一个社区发现的集成聚类方法框架。首先通过集成选择获得较高质量的集成成员子集，之后通过模块度赋予成员不同的重要性权值，从而集成成员中发现的信息来改进社区发现效果。在几个真实网络上与其他经典社区发现算法进行对比实验的结果显示出基于该框架能有效提高社区发现的精确度。

关键词： 社区发现；集成聚类；集成选择

中图分类号： TP301.6　　**文献标志码：** A　　**文章编号：** 1673-5692(2020)04-382-06

An Ensemble Clustering Framework for Community Detection

JIANG Yi-nan[1], LIU Jia-chen[1,2], WANG Ya-shen[1], ZHANG Xin-hai[1], ZHANG Bo[1]

(1. China Academy of Electronics and Information Technology, Beijing 100041, China;
2. Xidian University, Xi'an 710071, China)

Abstract: As an integral part of network analysis, the issue of community detection is of great significance to understand the overall organization and functional characteristics of complex systems, and has extensive application scenarios in various fields such as personalized service, precise advertising, propagation prediction and even identification of criminal groups. Recently, Ensemble clustering approaches, which have been successfully applied in areas such as machine learning for clustering, are introduced in the issue of community detection to improve the accuracy and stability. An ensemble clustering framework for community detection has been proposed by aggregating ensemble selection and ensemble clustering to community detection. Firstly, the clustering ensemble selection approach is devised to the base clustering set that generated by multiple runs of community detection algorithms, to compute a subset of the base clustering set that maximize both the quality and the diversity. Then by weighting the importance of each clustering according to the modularity, the subset is integrated based on hierarchical clustering to improve community detection. Experiments on benchmark networks show the validity of our approach and the comparisons with other typical methods indicate the enhance of accuracy.

Key words: Community Detection, Ensemble Clustering, Ensemble Selection

收稿日期： 2020-01-06　　**修订日期：** 2020-03-10

基金项目： 国家重点研发计划(2017YFC0820503)；北京市科技新星计划项目(Z181100006218041)；北京市科技计划项目(Z181100009018008)

0 引 言

近年来随着社会媒体的发展,社会网络分析以及相关的应用受到了广泛关注[1-5]。尤其是随着网络通信、数字媒体、计算机技术的快速进步,社会网络数据的获取、生成、处理和分析技术也得到迅速发展,这为社会网络分析带来了更多的研究手段和更广泛的应用场景。社区发现作为网络分析的一个重要手段,对于理解复杂系统的整体组织及其功能特性具有重要意义,在多媒体大数据时代更是具有广泛的应用,其典型的应用领域包括个性化服务[6]、精准广告投放[7]、传播预测[8]甚至犯罪团伙发现[9]等。

社区结构(Community Structure)是复杂网络中最普遍也是最重要的拓扑特性之一,表现为社区内部各节点之间连接紧密,而不同社区之间的节点则连接稀疏[10,11]。自从复杂网络社区发现的问题提出以来,各种社区定义和发现算法陆续被提出,包括最早的图论的图分割方法[12-14]、基于聚类思想的方法[15-17]、基于模块度的方法[18-19]、标签传播算法[20-21]、动态算法以及其他针对重叠社区[22-24]、动态社区[25]、网络局部信息和先验信息[26,27]等的方法。这些算法及其改进算法往往需要在算法复杂度、计算速度、社区划分精度和稳定性之间进行权衡。

集成方法(Ensemble Methods),作为一种可以将不同来源的数据有效融合从而提高统计估算结果的方法[28],近年来被广泛应用于基于机器学习的分类和聚类问题中。而考虑到社区发现问题可以看作是对网络相似节点的一种聚类,集成聚类(Ensemble Clustering)方法被引入到社区发现问题的研究中以提升社区发现的精确度和稳定性[29]。其基本思路是将一些计算速度快但稳定性不佳的社区发现聚类算法得到的结果作为基聚类集,通过对其进行集成获得最终的更为准确和鲁棒的一致性聚类结果。更进一步地,研究表明集成聚类的结果质量与基聚类集的质量和多样性密切相关[30]。因此,有研究者对集成聚类选择(Cluster Ensemble Selection)方法进行研究,通过对基聚类集进行选择后再集成来获得更高质量的社区发现结果[31-32]。

本文将集成方法、集成聚类选择与社区发现相结合,提出了一个社区发现的集成聚类方法框架。在该框架中,首先通过集成选择获得较高质量的集成成员子集,之后通过模块度赋予成员不同的重要性权值,从而集成成员中发现的信息来获得较好的社区发现效果。本文在第 2 节中介绍了基于集成聚类和集成选择的社区发现算法框架,在第 3 节中给出了该算法框架在几个真实网络中的应用结果,在第 4 节中对结果进行了讨论。

1 算法框架

本文的社区发现算法框架是在集成聚类思想[33]的基础上提出的。图 1 展示了一个常见的集成聚类框架,它由三个部分组成: 集成成员生成、一致性函数和评估。可见,集成聚类框架的输入是一个给定的待聚类数据集,输出是该数据集的最终聚类结果。

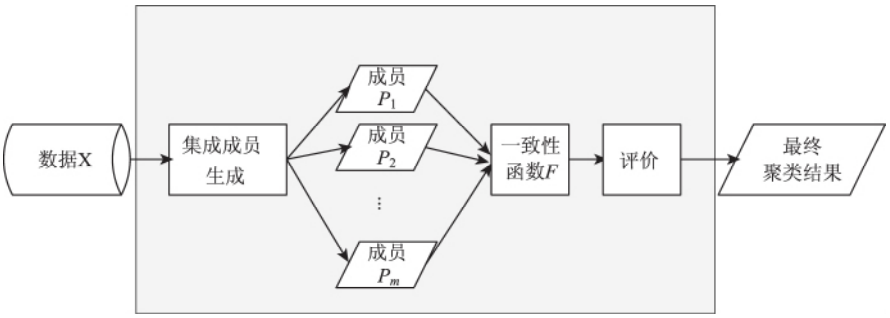

图 1　集成聚类框架

集成聚类框架的第一个阶段,主要目的是生成 m 个聚类作为集成成员。生成的成员应尽可能彼此不同,从而可以捕获数据中更多有用的信息。集成聚类的任务是通过一致性函数 F 集成成员$\{P_1,P_2,P_3,\cdots,P_m\}$来找到数据集 X 的一个划分 P^*,使得 P^* 在评价指标方面比集成中的单个成员更好。针对该方法的思想,本文将其应用在社区发现的问题上,并通过集成选择的方法来更好的生成集成成员。

2 集成选择与集成聚类

2.1 集成选择方法

不同的研究表明,集成聚类的输出质量与聚类集中每个基聚类的质量以及这些基聚类的多样性密切相关[30]。针对社区发现问题,社区发现结果的质量大多采用模块度[34]进行评价,而标准互信息(NMI)[35]可用来衡量两个社区发现结果的相似度。

结合以上分析,为了选择出一个既具有多样性又具有较高质量的优选基聚类子集,本文设计了一种集成选择方法。该算法的基本思想是:首先针对获得的多个社区发现候选方案,构建一个相似度矩阵,其中矩阵的值为每两个候选方案的 NMI 值。基于此矩阵,构建出一个候选方案间的相似图。之后通过一种社区发现算法对候选方案进行划分,被划分到同一社区的方案则有着较高的相似度。之后选出每个社区中一定比例的高模块度的社区方案,将这些选出的方案作为接下来用于集成的候选方案集。该算法具体步骤可描述如下:

1) 令 $Y=\{Y_1,Y_2\cdots,Y_n\}$ 为多种社区发现算法生成的社区发现候选结果集合。分别求 Y 中每两个结果之间的相似值,获得规模为 $n^*\ n$ 的相似度矩阵 S^y。

2) 基于相似矩阵构建结果间的相关邻近图(Relative neighborhood graph, RNG),应用社区发现算法将其分为不同的组,组内为相似度较高的社区发现结果。

3) 计算每个社区发现结果的模块度 Q,用来评价该结果的质量。

4) 在每个组里选取模块度前 $X\%$ 的候选结果,得到用于集成的结果集。

步骤 1 中本文用标准互信息(NMI)来衡量两种社区发现结果的相似度:其计算方法如下:

$$\mathrm{NMI}(A,B)=\frac{-2\sum_{i=1}^{F_A}\sum_{j=1}^{F_B}F_{ij}\log\left(\frac{n^*\ F_{ij}}{F_i^*\ F_j}\right)}{\sum_{i=1}^{F_A}F_i\log\left(\frac{F_i}{n}\right)+\sum_{j=1}^{F_B}F_j\log\left(\frac{F_j}{n}\right)} \tag{1}$$

其中,矩阵 F 中的 F_{ij} 用来表示同时在结果 A 中属于社区 i 且在结果 B 中属于社区 j 的节点数量。矩阵 F 中第 i 和 j 行元素之和分别用 F_i 和 F_j 表示。F_A(F_B)分别为社区发现结果 A(B)中的社区数量。通常 NMI 的取值范围在 [0,1] 之间,当 NMI 的取值为 1 时,则意味着 A 和 B 的划分情况完全一致。

步骤 2 中相关邻近图(Relative neighborhood graph, RNG)最初由 Toussaint 等人[36]提出,RNG 图是由以下构造规则定义的:两点 x_i 和 x_j 如果满足下列性质,则有边连接:

$$d(x_i,x_j)\leqslant\max_l\{d(x_i,x_l),d(x_j,x_l)\},\forall l\neq i,j \tag{2}$$

其中 $d(x_i,x_j)$ 为距离函数。

步骤 3 中采用模块度来评价每个社区发现结果的质量,其计算方法由 Newman 等提出:

$$Q=\sum_{c=1}^{k}\left[\frac{l_c}{L}-\left(\frac{d_c}{2L}\right)^2\right] \tag{3}$$

其中,k 用来表示我们所得到的社区个数。L 表示整个网络中所有连边的数量,l_c 代表某一社区 c 中内部的连边数目,而 d_c 则由该社区 c 中内部节点的度求和得到。可以看出,当一个社区内节点间连接更紧密时,该社区的模块度 Q 就会更大。

2.2 基于集成聚类的社区发现算法

通过 2.1 小节的集成选择算法,可以得到多样性和质量较优的社区发现结果集合作为构建集成的成员,成员之间的多样性意味着它们捕获了关于数据的不同信息,有助于提高集成的性能。

如图 2 所示,本文提出的基于集成聚类的社区发现算法主要包括以下五个阶段:

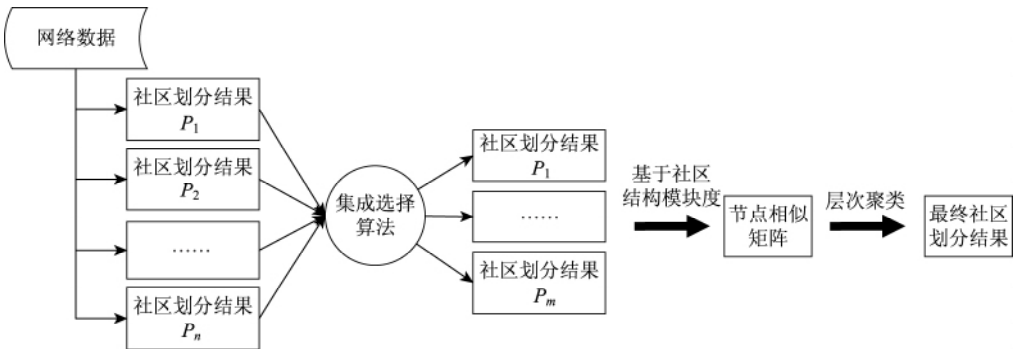

图 2 算法流程

(1) 通过 LPA、Louvain 等社区发现算法多次运行获得 N 个社区发现算法结果 $P=\{P_1, P_2, \cdots, P_n, \cdots, P_N\}$,其中 P_n 为第 n 个社区发现算法结果。

(2) 运用 2.1 章节中提出的集成选择算法,选出 m 个结果作为较优候选结果子集 $P'=\{P_{i1}, P_{i2}, \cdots, P_{im}\}$,其中 $i_m \in [\![1, n]\!]$。

(3) 计算结果子集中每个社区发现结果的模块度 Q_t,根据公式(3)计算得到每个结果的权值,其中 ω_t 为第 t 组结果的重要性权值。

$$\omega_t = \frac{Q_t}{\sum_{t=1}^{m} Q_t} \tag{4}$$

(4) 根据公式(4)构建一个节点间加权相似度矩阵 S,其中 S_{ij} 为节点 i 和节点 j 之间的相似度,计算方式如下公式(4)所示:

$$S(i,j) = \sum_{t=1}^{m} \omega_t f(X_{it}, X_{jt}) \tag{5}$$

其中,ω_t 为公式(3)中计算所得第 t 组集成成员的权值,m 为集成成员的数量。而 X_{it},X_{jt} 分别为第 t 组社区发现结果中网络节点 i 和 j 的标签,当 $X_{it}=X_{jt}$ 时,$f(X_{it}, X_{jt})=1$,否则 $f(X_{it}, X_{jt})=0$。通过模块度给不同集成结果分配不同的重要性权值,然后综合每个集成结果的社区划分情况来反映节点间的相似性。

(5) 在获得的相似性矩阵上应用凝聚层次聚类算法来生成最终的社区发现结果,层次聚类算法中的计算两个组合数据点距离的方式采用 Average Linkage。

凝聚层次聚类[37]方法的思想是: 开始把每一个对象视为一个聚类,之后计算两两间的距离,合并距离较近的两个聚类。其中 Average Linkage 计算距离的方式是取两个组合中的所有点两两之间距离的平均值。

3　实验结果

3.1　数据集

针对本文所提出的算法,本文选取四个常见的真实网络数据集用于验证社区发现效果。四个数据集的内容介绍如下:

(1) Karate club 网络[38]: 该数据是学者 Zachary 通过分析大学空手道俱乐部中 34 个成员间关系得到的经典网络。在真实网络中所有成员根据成员间关系被分为两个群体。

(2) Dolphins 网络[39]: 该网络是通过观察 62 只海豚在 7 年的时间互动情况得到的。海豚是网络中的节点,边表示海豚之间存在频繁互动。根据海豚日常活动发现,海滩内部可分为两个群体。

(3) American football 网络[15]: 该网络为美国大学生橄榄球球队比赛情况组成的网络。115 个节点分别为 115 个球队,连边表示球队间发生过比赛。这些球队分成 12 个联盟,同联盟内球队之间比赛较多。

(4) Political books 网络[18]: 该网络是政治书籍联合购买网络,节点为书籍,边表示两本书被同时购买过。

表 1　数据集规模

网络	节点数	边数	社区数
Karate club	34	78	2
Dolphins	62	159	2
American football	115	613	12
Political books	105	441	3

3.2　实验结果分析

目前已有一些经典的社区发现算法如 LPA、Louvain 算法等,我们应用本文算法及其他典型社区发现算法在四个真实网络数据集上进行实验。由于已知四个网络数据集的真实社区划分结果,因此计算算法结果与网络真实划分的标准化互信息(NMI)可以更好的评价算法的有效性。

在其他社区发现算法对比实验上,我们采用运行 50 次求均值的方法得到其结果与真实划分的 NMI 值,实验结果如表 2 所示。在表 2 中,本文提出算法的输入为 100 个由其他算法多次运行获得的社区发现结果。

表 2　不同算法在多个数据集的 NMI 值比较

数据集	LPA	Louvain	Girvan-Newman	本文算法
Karate Club	0.654	0.657	0.617	0.921
Football	0.874	0.803	0.851	0.926
Dolphins	0.579	0.571	0.599	0.889
Polbooks	0.529	0.583	0.561	0.572

从表 2 中可以看出,本文算法在四个数据集上的 NMI 指标均高于别的算法,即本文算法社区发现结果更接近于真实社区。在 Karate Club 和 Football 数据集上,本文算法 NMI 值相比于其他三种算法有着 30% 的提高。

4 结 语

本文中，我们在集成聚类的思想上，提出一种基于集成的社区发现算法框架，首先考虑多样性和质量两个方面，从多种社区发现获得的结果中选出较优的结果子集，用来提高后续集成输出的质量。之后，通过模块度评价指标计算每个结果的重要性权值，集成所有结果中划分情况计算出节点的相似性矩阵，最后通过层次聚类方法得到网络社区结构。通过在真实数据集上与其他算法对比，本文算法在 NMI 评价指标上均有所提高，可以有效地发现更真实的社区结构。具有精确性和稳定性的社区发现算法，在多媒体大数据时代具有广泛的应用前景，如在商业领域提供个性化服务、发现潜在客户，在新媒体领域预测信息传播行为，在社会安全领域挖掘犯罪团伙等。

参考文献:

[1] Shi J, Salmon CT. Identifying Opinion Leaders to Promote Organ Donation on Social Media: Network Study [J]. Journal of Medical Internet Research, 2018, 20(1): 1-7.

[2] Al-Garadi M A, Varathan K D, Ravana S D, et al. Analysis of Online Social Network Connections for Identification of Influential Users [J]. Acm Computing Surveys, 2018, 51(1): 1-37.

[3] Ghosh R, Lerman K. Rethinking centrality: The role of dynamical processes in social network analysis [J]. Discrete and Continuous Dynamical Systems - Series B (DCDS-B), 2014, 19(5): 1355-1372.

[4] Fan L, Wu W, Lu Z, et al. Influence Diffusion, Community Detection, and Link Prediction in Social Network Analysis [M]. Berlin: Springer Publishing, 2013.

[5] Wolfe A W. Social Network Analysis: Methods and Applications by Stanley Wasserman; Katherine Faust [J]. 1995, 91(435): 219-220.

[6] Feng H, Tian J, Wang H J, et al. Personalized recommendations based on time-weighted overlapping community detection [J]. Information & Management, 2015, 52(7): 789-800.

[7] Gurini D F, Gasparetti F, Micarelli A, et al. iSCUR: Interest and Sentiment-Based Community Detection for User Recommendation on Twitter [M]. Berlin: Springer Publishing, 2014.

[8] Zhu H, Ma J. How the contact differences and individuals' similarity affect the rumor propagation process in complex heterogeneous networks [J]. International Journal of Modern Physics C, 2018, 29(8): 1-5.

[9] Gangopadhyay A, Chen S. Health Care Fraud Detection with Community Detection Algorithms [C] // IEEE International Conference on Smart Computing. St. Louis: IEEE Press, 2016: 1-5.

[10] Newman, J. M E. The Structure and Function of Complex Networks [J]. Siam Review, 2003, 45(2): 167-256.

[11] 姜雅文. 网络空间用户行为的复杂网络特性研究[J]. 中国电子科学研究院学报, 2017, 12(5): 452-457,480.

[12] Kernighan B W, Lin S. An Efficient Heuristic Procedure for Partitioning Graphs [J]. Bell System Technical Journal, 1970, 49(2): 291-307.

[13] Pothen A, Simon H D, Liou K-P. Partitioning Sparse Matrices with Eigenvectors of Graphs [J]. Siam Jmatrix Analappl, 1990, 11(3): 430-452.

[14] Flake G W, Lawrence S, Giles C L, et al. Self-organization and identification of Web communities [J]. Computer, 2002, 35(3): 66-70.

[15] Girvan M, Newman M E J. community structure in social and biological networks [J]. PNAS, 2002, 99(12): 7821-7826.

[16] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2003, 69(6 Pt 2): 066133.

[17] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks [J]. Physical review E, Statistical, nonlinear, and soft matter physics, 2004, 70: 264-277.

[18] Newman M E J. Modularity and community structure in networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.

[19] Girvan M, Newman M E J, Girvan M, et al. Community structure in social and biological networks [J]. Proceedings of the National Academy of ences, 2002, 99(12): 7821-7826.

[20] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 76(3): 036106.

[21] Subelj L, Bajec M. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2011, 83

(3): 036103.

[22] Zhang y, Yeung D-Y. Overlapping community detection via bounded nonnegative matrix tri-factorization [J]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, 8: 606-614.

[23] Gregory S. Finding Overlapping Communities Using Disjoint Community Detection Algorithms [M]. Berlin: Springer, 2009.

[24] Sun P G, Gao L, Han S S. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks [J]. information sciences, 2011, 181(6): 1060-1071.

[25] Greene D, Doyle D, Cunningham P. Tracking the Evolution of Communities in Dynamic Social Networks [C] // 2010 International Conference on Advances in Social Networks Analysis and Mining. Odense: IEEE Press, 2010: 176-183.

[26] Ma X, Gao L, Yong X, et al. Semi-supervised clustering algorithm for community structure detection in complex networks [J]. Physica A Statistical Mechanics & Its Applications, 2010, 389(1): 187-197.

[27] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123.

[28] Scott Kirkpatrick, D. Gelatt Jr, Mario P. Vecchi. Optimization by Simmulated Annealing [J]. science, 1983, 42(3): 671-680.

[29] Seifi M, Guillaume J-L. Community cores in evolving networks [C] //Proceedings of the 21st International Conference on World Wide Web. Lyon, France: ACM. 2012: 1173-1180.

[30] Fern X Z, Lin W. Cluster Ensemble Selection [J]. Statistical Analysis & Data Mining the Asa Data Science Journal, 2008, 1(3): 128-141.

[31] Kanawati R. Ensemble Selection for Community Detection in Complex Networks [M]. Berlin: Springer Publishing, 2015.

[32] Huang F L, Huang M X, Yuan C A, et al. Spectral clustering ensemble algorithm for discovering overlapping communities in social networks [J]. Kongzhi Yu Juece/control & Decision, 2014, 29(4): 713-718.

[33] Strehl A, Ghosh J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions [J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.

[34] Newman M E J, Girvan M. Finding and Evaluating Community Structure in Networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69(2 Pt 2): 026113.

[35] Yang J, Leskovec J. Defining and Evaluating Network Communities based on Ground-truth [J]. Knowledge & Information Systems, 2012, 42(1): 181-213.

[36] Toussaint G T. The relative neighbourhood graph of a finite planar set [J]. Pattern Recognition, 12(4): 261-268.

[37] Murtagh, F. A Survey of Recent Advances in Hierarchical Clustering Algorithms [J]. Computer Journal, 1983, 26(4): 354-359.

[38] Zachary W W. An Information Flow Model for Conflict and Fission in Small Groups [J]. Journal of Anthropological Research, 1976, 33(4): 452-473.

[39] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations [J]. 2003, 54(4): 396-405.

作者简介

江逸楠(1988—),女,湖北人,工程师,主要研究方向为复杂网络分析、社会网络计算等;

E-mail: jiang_yinan@126.com

刘家琛(1997—),男,河南人,助理工程师,主要研究方向为社会网络分析;

王亚珅(1989—),男,山东人,高级工程师,主要研究方向包括自然语言处理、社交网络分析、知识图谱等;

张欣海(1975—),男,辽宁人,研究员级高级工程师,主要研究方向为大数据与人工智能应用技术;

张 博(1982—),男,山东人,高级工程师,主要研究方向为信息系统与大数据应用。