基础科研训练结题报告

1. 训练目的

1.1 训练题目

本次科研训练的训练方向为数据挖掘与机器学习算法的研究。具体的训练题目为“引入博弈论的社团发现算法”。该算法在一般基于标签传递的动力学算法的基础上引入了博弈论的有关概念，对节点间的动力学进行了更加详细与贴合实际的描述，被用于解决普通标签传递算法中参数调整与高重叠密度下社团发现的准确性不足的问题。

1.2 研究现状

1.2.1 复杂网络中的社团

复杂网络中的社团概念来自于社交网络中的社团概念，在社交网络中的几个典型社团如图 1 所示。

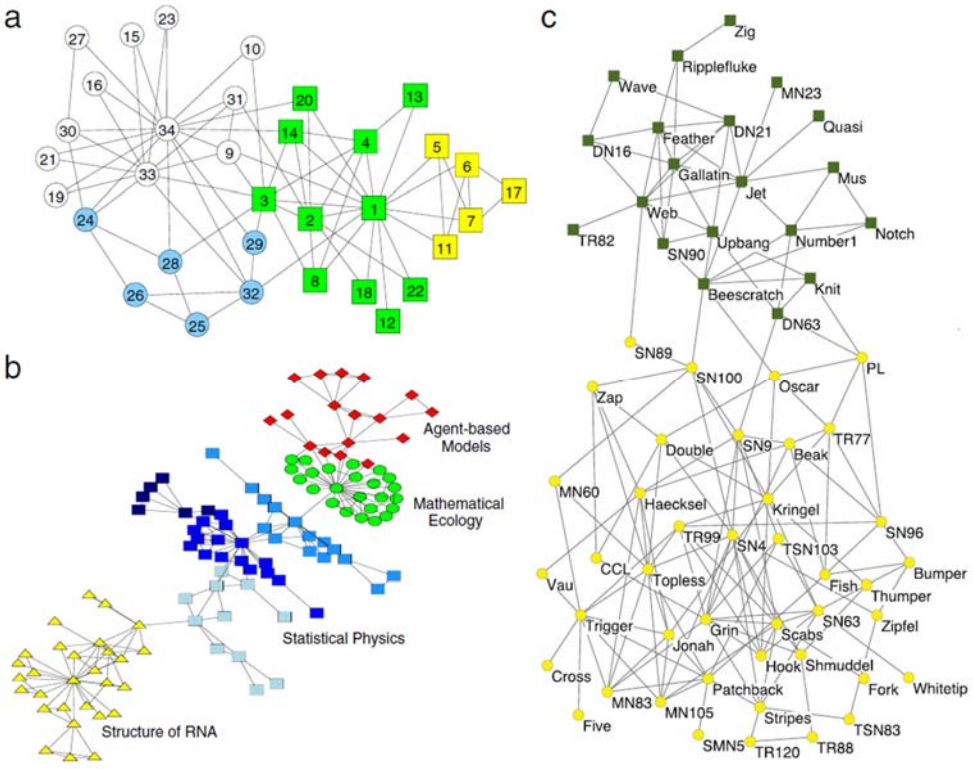

图 1 社交网络中的几个典型社团的划分：(a) Zachary 的空手道俱乐部，图中的划分代表了四个社团[1];(b) Santa Fe Institute 的科学家合作网络，图中的划分代表了七个社团[2];(c) Lusseau 的海豚网络，图中的划分代表了两个社团[3]。

对与社团本身的定义，学界并没有统一认可的方式，一个基本的社团划分原则是社团内的边的数目应该比连接社团内外的节点的边的数目要多[4]。

1.2.2 重叠社团

与传统的社团不同，在考虑重叠社团的情况下，图中的每一个节点 i 都可以从属于多个社团，对于每一个所属的社团 c 具有一个从属因子 α_{ic}，满足[5]

$$\begin{aligned} 0 \le \alpha_{ic} \le 1 \\ \sum_{c \in C} \alpha_{ic} = 1 \end{aligned} \tag{1}$$

重叠社团的引入与现实网络的实际相吻合，在现实的许多社交网络中，社团都体现了重叠的特性[6,7]。

1.2.3 重叠社团的发现算法

目前，学界对于重叠社团的发现算法可以分为以下几类：

(1) 派系过滤算法（CPM）。这种算法基于的假设是社团由重叠的全联通子图构成，通过搜索相邻的派系进行社团发现。典型的算法如 CFinder[8] 、CPMw[9] 、SCP[10] 等。

(2) 边划分法。边划分法的思想是将边而不是点进行社团发现，之后如果某个结点的不同边属于不同的边社团，那么这个点就从属于多个社团。典型的算法如 Link[11] 、CDAEO[12] 、Extended-Infomap[13] 等。

(3) 局部扩张和优化法。这种算法基于局部的效益函数来衡量稠密连接的结点集合的质量。典型的算法如 CIS[14] 、LFM[15] 、OSLOM[16] 、UEOC[17] 、iLCD[18] 、EAGLE[19] 等。

(4) 模糊发现法。模糊发现法对每个结点和社团间的联系程度进行量化。用到这个思想的算法有 FOG[20] 、SSDE[21] 、OSBM[22] 等。

(5) 动力学算法。基于动力学或者仿生思想的算法主要有 COPRA[23] 、SLPA[24,25] 、Game[26] 等。

(6) 其他种类。性能较好的有 CONGA[27] 、CONGO[28] 等。

这些算法中，在低重叠密度下，SLPA、OSLOM、Game 和 COPRA 有较好的性能；在高重叠密度下，SLPA 和 Game 拥有较好的性能[5]。

1.2.4 重叠社团发现算法的衡量指标

与普通社团发现算法的衡量指标不同，重叠社团发现算法必须考虑社团的重叠性，因此对算法性能的衡量指标也必须做对应的更改。衡量指标共分为两类，有真实划分信息情况下的衡量指标与缺少真实划分信息的衡量指标。有真实划分信息的衡量指标主要有 NMI[29] 与 Omega Index[30]，而缺少真实划分信息的衡量指标为 Overlapping Modularity E: Q_{ov}^{E} [31] 与 Overlapping Modularity Ni: Q_{ov}^{Ni} [32]。

2. 训练过程

本次训练主要分为三个阶段进行：

第一阶段：基本知识的学习。在本阶段中，对数据挖掘和机器学习相关领域的基本知识进行系统地学习。对必要的工具进行学习掌握。

第二阶段：阅读本领域有关文献，形成对领域的理解，确定研究方向，找出领域中存在的问题，确定自己的研究课题。在之前提到，本次训练的主要方向为基于同步和动力学的聚类分析。本阶段的安排包括自主阅读有关文献，与实验室其他同学老师进行讨论，并参与实验室的组会学习。通过大量的阅读与总结文献资料，总结当前领域的突出问题，作为自己后一阶段的研究课题。

第三阶段为根据前一阶段确定的研究课题，展开研究，改进现有的算法并将新的设计思路引入算法中，形成具有一定创新性并能解决领域前沿问题的新算法，加以实现并在各种数据集上进行测试，做进一步改进，最后形成有关研究成果，登刊发表。

3. 训练内容

3.1 第一阶段

第一阶段主要学习有关领域的基本知识，掌握基本工具，具体内容如下。

在数据挖掘领域，主要的参考资料为 Han 等编著的《Data Mining: Concepts and Techniques》[33]。学习的内容包括以下几个方面：

(1) 数据预处理部分：数据清理、数据聚合、数据变换、数据离散化等。

(2) 频繁模式挖掘部分：Apriori 算法、FP-Growth 算法等。

(3) 分类部分：决策树算法、贝叶斯学习、规则集学习、神经网络、支持向量机、懒惰学习等。

(4) 聚类部分：k-means 算法、k-medoids 算法、DBSCAN 算法、EM 算法、SCAN 算法等。

在机器学习领域，主要参考资料为 Mitchell 的《Machine Learning》[34]。学习

的内容包括以下几个方面：

(1) 概念学习：Candidate-Elimination 算法等。

(2) 决策树学习：ID3 算法等。

(3) 神经网络：梯度下降法、BP 法等。

(4) 贝叶斯学习：朴素贝叶斯分类器、贝叶斯网络、EM 算法等。

(5) 基于实例的学习：k-邻居算法、懒惰算法等。

(6) 遗传算法：GA 算法等。

(7) 规则集学习：Sequential-Covering 算法、CN2 算法、FOIL 算法等。

(8) 分析学习：Prolog-EBG 算法等。

(9) 混合推理和分析学习：KBANN 算法、FOCL 算法等。

(10)增强学习：Q 学习算法等。

在工具的学习上面，主要是对 Java 编程语言的学习。在语法方面，主要参考的资料为 Horstmann 的《Core Java Volume I - Fundamentals》[35]。在算法方面主要的参考资料为 Sedgewick 的《Algorithms》[36]。

3.2 第二阶段

第二阶段为阅读本领域有关文献，形成对领域的理解，确定研究方向，并找出领域前沿的问题以此确定自己的研究课题。

本阶段首先是对动力学的文献进行广泛的阅读[37-49] 。在阅读的基础上，确定了研究的方向为复杂网络中的重叠社团发现。确定了研究方向后，先是详细阅读了具体方向的文献综述[5,50]，了解到当前的重叠社团发现算法的性能，无论是从时间复杂度还是各项结果指标来看，都是基于动力学的算法如 COPRA 和 SLPA 等，因此将自己的研究范围进一步缩小，专注 COPRA 和 SLPA 的进一步改进提升上。

对 SLPA 的后续改进算法 Fast[51]，COPRA 的后续改进算法 BMLPA[52]、AntCBO[53]、PLPAC[54]、PGLPA[55] 等进行了阅读后，注意到虽然这些后续改进算法都在原有的 SLPA 与 COPRA 算法上进行了一定的改进，但是对最终结果的影响都较小，且都没有解决在高重叠密度下性能较低的问题。不仅如此，这些算法在实际使用上，都涉及参数调整问题，即对于不同的数据集，算法中的特定参数需要进行手工调整以得到较好结果，这给非专业领域人士的人使用带来了极大的不变。因此，可以考虑通过对 SLPA 和 COPRA 算法进行进一步改进，对结点间的动力学进行更为完整的建模，从而解决参数调整以及高重叠密度下性能较低的问题。

为了实现这一目标，注意到基于博弈论的算法 Game 以及其改进算法 OCEO[56]、Co-Game[57]，虽然这些算法本身性能不如基于动力学的 COPRA 和 SLPA，但是其

算法对于重叠密度并不敏感，因此考虑将它们中的某些思想以及博弈论本身的概念引入结点间相互作用的动力学中，提升算法的性能。从而得到了本次的研究课题：引入博弈论的社团发现算法。

3.3 第三阶段

第三阶段为完成具体的研究课题，包括算法的设计、实现、测试与改进。

对于结点间的相互作用，考虑将节点本身视为一些参与博弈的个体。在相互作用时，参与博弈的个体总是考虑本次作用是否能给自己带来收益，并总是趋向进行让自己收益最大化的行为。

为了对这种博弈行为建模，分别考虑个体的收益函数与损失函数。

收益函数定义为节点的局部 Overlapping Modularity Ni，即

$$q_i(c) = F(\alpha_{i,c})(\sum_{j \in N_i} F(\alpha_{j,c}) - \frac{k_i}{2m}\frac{\Phi^2(c)\Psi(c)}{n^2}) \tag{2}$$

其中

$$\Phi(c) = \sum_{s \in c} F(\alpha_{s,c}), \Psi(c) = \sum_{s \in c} F(\alpha_{s,c}) k_s, F(\alpha_{i,c}) = \frac{1}{1+e^{-f(\alpha_{i,c})}}, f(x) = -30 + 60x \tag{3}$$

k_i 为第 i 个节点的度，m 为网络边的总数，n 为网络节点的总数，$\alpha_{i,c}$ 为结点 i 对与社团 c 的从属因子，N_i 为节点 i 的邻居的集合。

通过让个体间作用时考虑这一优化函数，可以贪心地实现整个网络向着增大 Overlapping Modularity Ni 的方向演化，从而提升社团划分的性能。

损失函数考虑为节点相互作用时，每次标签的传递都需要消耗一定的能量。为了模仿真实社交网络的实际情况，我们定义单位标签传递的能量消耗反相关与节点间标签集合的相似性。一种实现方式即为

$$\delta_{i \to j,c} = (1 - \alpha_{i,c})(1 - \alpha_{j,c}) \tag{4}$$

而对于每一对节点间，单次迭代的能量的总数为1。在每次相互作用时，节点将优先传递消耗能量少的标签，之后传递能量次少的标签，直到本次迭代的能量余额耗尽。

基于引入以上收益函数与损失函数进入标签传递的动力学模型中，我们具体设计了 GLPA（Game theory incorporated Label Propagation Algorithm）算法。这一算法的伪代码如下：

```
void GLPA()
{
        readData(inputNetworkData);
        initialize();
        while(loopFlag)
        {
                for(Node node:nodeSet)
                {
                        node.interact();
                }
                loopFlag.update();
        }
        postProcess();
        produceCommunity(outputFile);
}
```

分别考虑收益函数与损失函数的情况下，GLPA 在几个典型的数据集下的结果如表 1 所示。

Function\Dataset	Karate		Dolphins		Football	
	Q_{ov}^{Ni}	Q_{ov}^{E}	Q_{ov}^{Ni}	Q_{ov}^{E}	Q_{ov}^{Ni}	Q_{ov}^{E}
Gain Function	0.674	0.343	0.708	0.488	0.691	0.601
Loss Function	0.408	0.020	0.753	0.211	0.680	0.070

表 1 GLPA 对典型社交网络的社团划分结果

从表中可以看出，GLPA 算法的结果与 COPRA 和 SLPA 的结果是可以比拟的，但是仍然有许多值得改进之处。如对于采用损失函数时，Q_{ov}^{E} 的表现较为不理想，这是因为后处理时对于无关噪声对应的标签缺乏一定的处理。

4. 总结与收获

本次基础科研训练历时一年，从基本的算法与工具的学习掌握，到对领域前沿进行系统了解研究，到发现总结当前领域亟待解决的问题，从而形成自己的研究课题，最后到完成算法的设计与实现，并对算法的可行性进行测试，可谓艰辛。最终算法在多个数据集上与领域前沿的算法具有可比拟性，但还存在一些实际问题需要进一步研究解决。通过这次训练，实际的感受到了科研工作的复杂性，认识到了自己在学习能力以及思考能力上仍存在一定不足。得到的最宝贵的经验是在之后的科研中，前期不要太泛、太广的阅读，应当尽早确定小方向从而快速获得实质性进展。

参考文献

[1] Donetti L, Muñoz M A. Detecting network communities: a new systematic and efficient algorithm[J]. Journal of Statistical Mechanics Theory & Experiment, 2004, 2004(10):10012.

[2] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.

[3] Arenas A, Fernandez A, Gomez S. Analysis of the structure of complex networks at different resolution levels[J]. New Journal of Physics, 2008, 10(5): 053039.

[4] Fortunato S. Community detection in graphs[J]. Physics reports, 2010, 486(3): 75-174.

[5] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state-of-the-art and comparative study[J]. Acm computing surveys (csur), 2013, 45(4): 43.

[6] Kelley S, Goldberg M, Magdon-Ismail M, et al. Defining and discovering communities in social networks[M]. Handbook of Optimization in Complex Networks. Springer US, 2012: 139-168.

[7] Lee C, Reid F, McDaid A, et al. Detecting highly overlapping community structure by greedy clique expansion[J]. arXiv preprint arXiv:1002.1827, 2010.

[8] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[9] Farkas I, Ábel D, Palla G, et al. Weighted network modules[J]. New Journal of Physics, 2007, 9(6): 180.

[10] Kumpula J M, Kivelä M, Kaski K, et al. Sequential algorithm for fast clique percolation[J]. Physical Review E, 2008, 78(2): 026109.

[11] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks[J]. nature, 2010, 466(7307): 761-764.

[12] Wu Z, Lin Y, Wan H, et al. A fast and reasonable method for community detection with adjustable extent of overlapping[C]. Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on. IEEE, 2010: 376-379.

[13] Kim Y, Jeong H. Map equation for link communities[J]. Physical Review E, 2011, 84(2): 026110.

[14] Kelley S. The existence and discovery of overlapping communities in large-scale networks[D]. Rensselaer Polytechnic Institute, 2009.

[15] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.

[16] Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks[J]. PloS one, 2011, 6(4): e18961.

[17] Jin D, Yang B, Baquero C, et al. A Markov random walk under constraint for discovering overlapping communities in complex networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2011, 2011(05): P05031.

[18] Cazabet R, Amblard F, Hanachi C. Detection of overlapping communities in dynamical social networks[C]. Social Computing (SocialCom), 2010 IEEE Second International Conference on. IEEE, 2010: 309-314.

[19] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(8): 1706-1712.

[20] Davis G B, Carley K M. Clearing the FOG: Fuzzy, overlapping groups for social networks[J]. Social Networks, 2008, 30(3): 201-212.

[21] Magdon-Ismail M, Purnell J. Fast overlapping clustering of networks using sampled spectral distance embedding and gmms[J]. Rensselaer Polytechnic Inst., Tech. Rep, 2011.

[22] Latouche P, Birmelé E, Ambroise C. Overlapping stochastic block models with application to the french political blogosphere[J]. The Annals of Applied Statistics, 2011: 309-336.

[23] Gregory S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2010, 12(10): 103018.

[24] Xie J, Szymanski B K, Liu X. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[C]. Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. IEEE, 2011: 344-349.

[25] Xie J, Szymanski B K. Towards linear time overlapping community detection in social networks[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2012: 25-36.

[26] Chen W, Liu Z, Sun X, et al. A game-theoretic framework to identify overlapping communities in social networks[J]. Data Mining and Knowledge Discovery, 2010, 21(2): 224-240.

[27] Gregory S. An algorithm to find overlapping community structure in networks[J]. Knowledge discovery in databases: PKDD 2007, 2007: 91-102.

[28] Gregory S. A fast algorithm to find overlapping communities in networks[J]. Machine learning and knowledge discovery in databases, 2008: 408-423.

[29] Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis[J]. Physical review E, 2009, 80(5): 056117.

[30] Collins L M, Dent C W. Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions[J]. Multivariate Behavioral Research, 1988, 23(2): 231-242.

[31] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(8): 1706-1712.

[32] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009(03): P03024.

[33] Han J, Pei J, Kamber M. Data mining: concepts and techniques[M]. Elsevier, 2011.

[34] Mitchell M. Machine Learning[M]. McGraw-Hill, 1997.

[35] Horstmann C S, Cornell G. Core Java 2: Volume I, Fundamentals[M]. Pearson Education, 2002.

[36] Sedgewick R, Wayne K D. Algorithms[M]. Addison-Wesley Professional, 2011.

[37] Pikovsky A, Rosenblum M, Kurths J. Synchronization: a universal concept in nonlinear sciences[M]. Cambridge university press, 2003.

[38] Kuramoto Y. Chemical oscillations, waves, and turbulence[M]. Springer Science & Business Media, 2012.

[39] Böhm C, Plant C, Shao J, et al. Clustering by synchronization[C]. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 583-592.

[40] Shao J, Böhm C, Yang Q, et al. Synchronization based outlier detection[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010: 245-260.

[41] Shao J, Plant C, Yang Q, et al. Detection of arbitrarily oriented synchronized clusters in high-dimensional data[C]. 2011 IEEE 11th International Conference on Data Mining. IEEE, 2011: 607-616.

[42] Shao J, He X, Yang Q, et al. Robust synchronization-based graph clustering[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2013: 249-260.

[43] Shao J, He X, Böhm C, et al. Synchronization-inspired partitioning and hierarchical clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 893-905.

[44] Shao J, Ahmadi Z, Kramer S. Prototype-based learning on concept-drifting data streams[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 412-421.

[45] Shao J, Yang Q, Dang H V, et al. Scalable clustering by iterative partitioning and point attractor representation[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2016, 11(1): 5.

[46] Shao J, Han Z, Yang Q, et al. Community detection based on distance dynamics[C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1075-1084.

[47] Wu H, Mao J, Sun W, et al. Probabilistic Robust Route Recovery with Spatio-Temporal Dynamics[C]. The ACM SIGKDD International Conference. 2016:1915-1924.

[48] Zhong E, Fan W, Zhu Y, et al. Modeling the dynamics of composite social networks[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 937-945.

[49] Tsytsarau M, Palpanas T, Castellanos M. Dynamics of news events and social media reaction[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 901-910.

[50] Fortunato S. Community detection in graphs[J]. Physics reports, 2010, 486(3): 75-174.

[51] Elyasi M, Meybodi M, Rezvanian A, et al. A fast algorithm for overlapping community detection[C]. Information and Knowledge Technology (IKT), 2016 Eighth International Conference on. IEEE, 2016: 221-226.

[52] Wu Z H, Lin Y F, Gregory S, et al. Balanced multi-label propagation for overlapping community detection in social networks[J]. Journal of Computer Science and Technology, 2012, 27(3): 468-479.

[53] Zhou X, Liu Y, Zhang J, et al. An ant colony based algorithm for overlapping community detection in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2015, 427: 289-301.

[54] Chen N, Liu Y, Cheng J, et al. Parallelizing label propagation for overlapping community detection[C]. Behavioral, Economic and Socio-cultural Computing (BESC), 2016 International Conference on. IEEE, 2016: 1-7.

[55] Zhang Q, Qiu Q, Guo W, et al. A social community detection algorithm based on parallel grey label propagation[J]. Computer Networks, 2016, 107: 133-143.

[56] Havvaei E, Deo N. A Game-Theoretic Approach for Detection of Overlapping Communities in Dynamic Complex Networks[J]. arXiv preprint arXiv:1603.00509, 2016.

[57] Zhao X, Wu Y, Yan C, et al. An Algorithm Based on Game Theory for Detecting Overlapping Communities in Social Networks[C]. Advanced Cloud and Big Data (CBD), 2016 International Conference on. IEEE, 2016: 150-157.