А. Д. ФАТИН, Е. Ю. ПАВЛЕНКО, И. С. ЕРЕМЕНКО
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МЕТОДЫ ГРАФОВОЙ КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ ИММУНИЗАЦИИ КИБЕРФИЗИЧЕСКИХ СИСТЕМ

В рамках доклада представлены подробный разбор и анализ пяти наиболее перспективных алгоритмов кластеризации графов в контексте задач иммунизации, адаптации и детектирования скрытых связей в киберфизических системах. Приведенные особенности и способы реализации рассматриваемых методов в дальнейшем позволяют провести численный анализ и сравнение эффективности существующих методов и, с учетом полученной информации, синтез нового метода, базирующегося на особенностях и преимуществах проанализированных методов.

Введение. В настоящее время задача кластеризации сетей является одной из основополагающих при решении вопросов иммунизации киберфизических систем, детектирования ботнетов и скрытых кластеров, а также при решении задач адаптивности, отказоустойчивости, надежности и так далее [1–3].

В задачах иммунизации киберфизических систем алгоритмы кластеризации применяются обычно для первичного разделения зараженных узлов на группы согласно их приоритету. Альтернативным вариантом применения алгоритмов кластеризации является изоляция зараженных узлов путем выделения новых кластеров, включающих в себя как зараженные, так и потенциально зараженные узлы, с последующей изоляцией всего кластера.

Касаемо задач детектирования ботнетов и скрытых кластеров, алгоритмы кластеризации используются при решении задачи нахождения скрытых связей между узлами, которые удается установить путем неявной кластеризации графа социальных связей или графа сетевых запросов, а также анализа сетевой нагрузки уже выделенных графов.

Говоря о задачах адаптивности, отказоустойчивости, надежного функционирования киберфизических систем, обычно рассматривают классическую кластеризацию системы на зоны ответственности или на зоны потенциальной надежности, то есть на так называемые «точки возможного отказа».

Под кластеризацией, согласно [4], понимается классификация или ряд моделей разделения вершин и/или ребер графа по ряду признаков и/или свойств на определенные группы, называемые кластерами.

Хотя задача эффективной кластеризации графов на данный момент является открытой, существует достаточное множество алгоритмов, потенциально являющихся лидерами в своих направлениях.

Предлагаемый доклад посвящен анализу, обобщению и систематизации существующей и актуальной на данный момент информации о методах кластеризации компьютерных сетей в задачах иммунизации, детектирования скрытых кластеров, адаптивности, отказоустойчивости и надежности, а также созданию теоретического базиса для последующей реализации собственных численных тестов и синтезу нового метода кластеризации графов в задачах иммунизации киберфизических систем.

Обоснование выбора рассматриваемых методов кластеризации. Среди исследованных алгоритмов кластеризации наибольший интерес представляют следующие:

1. Алгоритм кластеризации распределенной сети.
2. Алгоритм быстрого обнаружения центральных узлов.
3. Алгоритм обнаружения латентных состояний сети.
4. Алгоритм вариационного обучения с совместным встраиванием.
5. Тензорное разложение для кластеризации многослойных сетей.

Выбор данных алгоритмов и методов кластеризации графов обусловлен их широкой применимостью в ряде рассматриваемых задач (иммунизация киберфизических систем, детектирование скрытых связей, адаптивный анализ, анализ устойчивости и надежности), а также наибольшей

эффективностью по сходимости, скорости численного моделирования, точности и корректности кластеризации, а также рядом прочих уникальных параметров, рассматриваемых далее.

Алгоритм кластеризации распределенной сети. Алгоритм кластеризации распределенной сети ANCA [5] наиболее интересен следующим рядом преимуществ:

1. Более низкая энтропия по сравнению с SA-Cluster [6].
2. Не требует изменения структуры графа.
3. Работа на неориентированном, взвешенном или невзвешенном графе.

Описанные выше преимущества алгоритма достигаются за счет объединения топологической структуры и информации об атрибутах графа, а именно за счет отказа от изменения топологии в пользу добавления семян (наборов вершин), которые используются для описания характеристик каждой вершины графа. В конечном итоге, опираясь на набор дополнительных атрибутов в виде семян, проводится разбиение вершин графа на кластеры с использованием метода *k*-means [7].

Алгоритм быстрого обнаружения центральных узлов. Алгоритм кластеризации сети, основанный на быстром обнаружении центрального узла или же Graph clustering algorithm based on fast detection of central node (CFCN) [8], решает следующий ряд задач, которые редко можно встретить в иных алгоритмах сразу всем набором, а именно:

1. Вопрос об определении центра кластеризации.
2. Вопрос о стратегии кластеризации нецентральных узлов.
3. Вопрос определения расстояния и плотности узла.

К дополнительным преимуществам рассматриваемого алгоритма можно отнести следующие моменты:

1. Самостоятельное определение количества кластеров в сети.
2. Зависимость количества кластеров только от топологии сети.
3. Высокая скорость нахождения центра кластеризации.

Принцип работы алгоритма сводится к разделению графа на непересекающиеся подграфы с плотными краевыми связями внутри и разреженными краевыми связями между собой. В каждом из выделенных подграфов определяется центр кластера, по отношению к каждому из которых далее разделяются некластерные центральные узлы, создавая конечные кластеры.

Алгоритм обнаружения латентных состояний сети. Алгоритм кластеризации сети для обнаружения латентных состояний и точек изменения или же alternating direction method of multipliers (ADMM) [9] сводится к расширению системы выпуклой кластеризации на данные о нескольких сетях, использованию нормы матрицы Шаттена в штрафе слияния для корректировки межкластерной изменчивости, а также к использованию преимуществ сильной сходимости алгоритма.

Таким образом, реализуя описанный выше подход, удается добиться следующих результатов преимуществ:

1. Объединение краевой и спектральной информации и, как результат, повышенная устойчивость к наличию шумов.
2. Отсутствие необходимости в выборе заранее заданного количества кластеров.
3. Полное отсутствие генеративной модели для графа и, как результат, отказ от суммарной статистики.
4. Возможность работать как с направленными, так и ненаправленными графами.
5. Возможность контроля характера различий между центроидами кластеров.
6. Возможность работы с несколькими сетями: определения точек изменения или обнаружения скрытых временных состояний графов.

Алгоритм вариационного обучения с совместным встраиванием. В работе [10] рассматривается вариационная модель обучения с совместной сверткой для кластеризации приписанных сетей (VCLANC). Принцип работы метода сводится к использованию двойных вариационных автокодировщиков для последующего одновременного встраивания узлов и атрибутов в единое латентное пространство и последующее восстановление взаимного родства между узлами и атрибутами.

В качестве преимуществ использования данного метода следует выделить следующие особенности:

1. Превосходство по эффективности решения задач кластеризации над алгоритмами SDCN [11], NEC [12] и CAN [13].
2. Превосходство над SDCN по скорости решения задачи кластеризации.
3. Объединение сети и атрибутов в едином семантическом пространстве – возможность использования родства узлов и атрибутов.

Тензорное разложение для кластеризации многослойных сетей. Метод кластеризации многослойных сетей на основе центроиды (CMNC) [14], использующий мультилинейное разложение по рангам, позволяет рассмотреть задачу кластеризации с иной стороны, а именно предоставляет возможность разделить нерелевантные связи на группы сетей и одновременно выявить кластерную структуру в каждой из рассматриваемых групп.

Принцип работы данного метода сводится к анализу тензора третьего порядка как суммы низких мультилинейных ранговых членов, каждый из которых может быть записан как внешнее произведение матрицы ранга и вектора.

Данный подход позволяет достичь следующих преимуществ в сравнении с другими методами кластеризации, а именно:

1. Отказ от использования гиперпараметров – возможность реализации обучения без наблюдения.
2. Возможность ускорения задачи оптимизации за счет использования нелинейной системы наименьших квадратов (NLS).
3. Превышение эффективности классических алгоритмов (SymNMF [15], SC [16], CTSC [17], PairCRSC [18], CentCRSC [18] и NONCLUS [19]) за счет использования взаимодополняющих сетей.
4. Весьма высокая устойчивость к шуму и нерелевантным данным.
5. Возможность разделения шума на группы других сетей и использование полученных сетей для корректировки изначально полученных данных.
6. Учет предварительно известных данных и дополнительных параметров анализируемых графов.

Заключение. В работе были рассмотрены и проанализированы 5 наиболее актуальных методов кластеризации компьютерных сетей; были найдены преимущества и недостатки алгоритмов, выявлены их области применения.

Основная цель настоящего исследования заключалась в изучении, классификации и сравнении методов кластеризации компьютерных сетей. Данное исследование позволит упростить дальнейшую работу с данным материалом другим исследователям, а также обеспечит создание теоретико-практического базиса для собственной реализации алгоритма кластеризации компьютерных сетей.

В качестве дальнейшего вектора развития данной темы авторами рассматривается возможность численного сравнения эффективности выбранных методов и алгоритмов на синтетических и реальных данных, а также возможность реализации своего собственного с учетом выявленных преимуществ и способов их достижения в контексте задачи иммунизации киберфизических систем.

Исследование выполнено в рамках гранта Президента РФ для государственной поддержки молодых российских ученых – кандидатов наук МК-3861.2022.1.6.

ЛИТЕРАТУРА

1. Probabilistic partition of unity networks: clustering based deep approximation / N. Trask et al. 2021, arXiv:2107.03066. DOI: 10.48550/arXiv.2107.03066.
2. Adaptation of Vehicular Ad hoc Network Clustering Protocol for Smart Transportation / M. Ahmad et al. *Computers, Materials & Continua*. Vol.67. No. 2. 2021. pp.1353-1368. DOI: 10.32604/cmc.2021.014237.
3. Adaptive Network Automata Modelling of Complex Networks / A. Muscoloni, U. Michieli, C. V. Cannistraci. Preprints 2020, 2020120808. DOI: 10.20944/preprints202012.0808.v1.
4. Кластеризация данных методом растущего нейронного газа / А. В. Чернов и др. *Инженерный вестник Дона*. 2020. №7. С. 1-17.

5. ANCA : Attributed Network Clustering Algorithm / Issam Falih et al. COMPLEX NETWORKS 2017: Complex Networks & Their Applications VI, 2017. pp 241–252.
6. Clustering large attributed graphs: An efficient incremental approach. / Zhou Y. et al. *Proceedings of IEEE International Conference on Data Mining* (ICDM), 2010. pp. 689–698. DOI: 10.1109/ICDM.2010.41.
7. Some methods for classification and analysis of multivariate observations / MacQueen et al. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1961. Vol. 1. No. 14. pp. 281–297.
8. **Ziruo J., Fuqiang Q**. Network Clustering Algorithm Based on Fast Detection of Central Node / Jia Ziruo, Qi Fuqiang. *Scientific Programming*, 2022. pp 1-5. DOI: 10.1155/2022/4905190.
9. Network Clustering for Latent State and Changepoint Detection / Madeline Navarro et al. // arXiv - CS - Social and Information Networks, 2021. DOI: arxiv-2111.01273.
10. Variational Co-embedding Learning for Attributed Network Clustering / Shuiqiao Yang et al. // arXiv - CS - Machine Learning (IF). 2021. DOI: arxiv-2104.07295.
11. Structural deep clustering network / D. Bo et al. *Proceedings of The Web Conference 2020*, ser. WWW ’20, 2020. pp. 1400–1410.
12. Network embedding for community detection in attributed networks / H. Sun et al. *ACM Transactions on Knowledge Discovery from Data* (TKDD), 2020. Vol. 14, No. 3, pp. 1–25.
13. Co-Embedding Attributed Networks / Z. Meng et al. *the Twelfth ACM International Conference*, 2019. pp. 393–401.
14. Tensor Decomposition for Multilayer Networks Clustering / Zitai Chen et al. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 33(01). pp. 3371-3378. DOI: 10.1609/aaai.v33i01.33013371.
15. Symmetric nonnegative matrix factorization for graph clustering/ Kuang D et al. *SDM*, 2012. pp. 106–117. DOI: 10.1137/1.9781611972825.10.
16. **von Luxburg U**. A tutorial on spectral clustering / Ulrike von Luxburg. *Statistics and Computing*, 2007. No. 17(4). pp. 395–416. DOI: 10.48550/arXiv.0711.0189.
17. **Abhishek Kumar, Hal Daumé III**. A co-training approach for multi-view spectral clustering. *Proceedings of the 28th International Conference on Machine Learning* (ICML), 2011. pp. 393–400.
18. Co-regularized multi-view spectral clustering / Kumar A et al. *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011. Pp. 1413–1421.
19. Flexible and robust multi-network clustering / Ni J. et al. *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. pp. 835–844. DOI: 10.1145/2783258.2783262.

A.D.Fatin, E.Yu.Pavlenko, I.S.Eremenko (Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg)

Graph clustering methods in problems of immunization of cyber-physical systems

The paper presents a detailed analysis of the five most promising graph clustering algorithms in the context of immunization, adaptation and detection of hidden links in cyber-physical systems. The above features and ways of implementing the considered methods in the future allow for a numerical analysis and comparison of the effectiveness of existing methods and, taking into account the information received, the synthesis of a new method based on the features and advantages of the analyzed methods.

Авторы готовы представить текст на английском языке для сборника материалов мультиконференции, который будет подан для индексирования в Scopus.