Metaheuristic Based Clustering Algorithms for Biological Hypergraphs

Boonyarit Changaival[1], Gregoire Danoy[1], Marek Ostaszewski[2], Kittichai Lavangnananda[3], Franck Leprevost[1] and Pascal Bouvry[1]

[1]Computer Science and Communications Research Unit, University of Luxembourg, Luxembourg City, Luxembourg
boonyarit.changaival@uni.lu, gregoire.danoy@uni.lu, franck.leprevost@uni.lu, pascal.bouvry@uni.lu

[2]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belval, Luxembourg
marek.ostaszewski@uni.lu

[3]School of Information Technology ,King Mongkut's University of Technology Thonburi,
Bangkok, Thailand
kitt@sit.kmutt.ac.th

Keywords: combinatorial optimization, linear programming.

1. Introduction

Hypergraphs are widely used for modeling and representing relationships between entities, one such field where their application is prolific is in bioinformatics [1] and [2]. In the present era of big data, sizes and complexity of these hypergraphs grow exponentially, it is impossible to process them manually or even visualize their interconnectivity superficially. A common approach to tackle their complexity is to cluster similar data nodes together in order to create a more comprehensible representation. This enables similarity discovery and hence, extract hidden knowledge within the hyper graphs. Several state-of-the-art algorithms have been proposed for partitioning and clustering of hypergraphs. Nevertheless, several issues remain unanswered, improvement to existing algorithms are possible, especially in scalability and clustering quality. This article presents a concise survey on hypergraph-clustering algorithms with the emphasis on knowledge-representation in systems biomedicine. It also suggests a novel approach to clustering quality by means of cluster-quality metrics which combines expert knowledge and measurable objective distances in existing biological ontology.

2. Hypergraph-based knowledge representation and clustering: the Parkinson Disease case study

Knowledge representation using graph formalism is often achieved by encoding entities as graph nodes and relations between them as graph entities. A research domain greatly benefiting from such an approach is systems biomedicine, where entire (patho)biological processes are perceived as systems, and information about them is encoded as graphs. However, graph representation is insufficient for complex relationships between multiple elements, for instance enzyme-catalyzed biochemical reactions. In such situations the notion of graph has to be generalized to a hypergraph. For instance, Open Biological Expression Language (OpenBEL) format [3], used for biomedical knowledge representation, is based implicitly on hypergraph formalism.

A hypergraph is defined as H = (V, X), where V is a set of all vertices in the graph and X is a non-empty subset of vertices in V which are named as a hyperedges. We illustrate our problem by the hypergraph created from curated Parkinson disease (PD) knowledge from literature. A biomedical resource, called PD map [4] is available in OpenBEL format. Each OpenBEL statement contains the ID of biological entities (proteins, genes, etc.) and their relationships (catalyticActivity, directlyIncreases, etc.). By connecting entities in the same statement based on their relationships and/or connecting different statements together, the result is a hypergraph where subsets in E are connected by an arc (in this case, for simplification, the orientation of the arcs are not considered and thus treated as edges). By utilizing the given information in each statement, we extend the notation of the hypergraph by attributes, defining as H = (V, X, E, A_V, A_E) where E is set of all edges (or arc), A_V is introduced as the set of all possible attributes of the vertices in V and A_E is the set of all possible attributes of the edges in E. The example graph is shown in figure 1. Each shaded area represent hyperedge and one hyperedge can be a part of another hyperedge or connect to another hyperedge. Let us highlight that the PD map hypergraph is sparse continues to grow at an alarming rate as more is known about the parkinson-disease is known. As commonly known, graph-clustering problem is NP -Hard [2] and [5], meta-heuristic is a suitable candidate for providing possible satisfactory solutions.

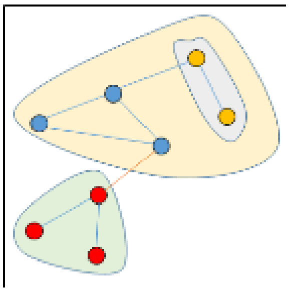

Figure 1 Hypergraph

3. Related Work

A hypergraph can be transformed into a collection of graphs, this allows set of vertices in the hypergraph be treated as set of vertices in an ordinary graph. This transformation makes it possible to apply graph algorithms in order to extract useful information. Edges can then be added to connect vertices in the same hyperedge resulting in a clique for each hyperedge.

Partitioning algorithms such as Kernighan-Lin (KL) algorithm and Fidduccia-Mattheyses (FM) as described in [7] are two of the most used vertex-swapping approaches used for hypergraph clustering, i.e. after applying the above-mentioned transformation from hypergraph to graph. These two algorithms have their application in VLSI circuits design. Moreover, general clustering algorithms like “K-mean”, “Minimum Cut” and “Spectral clustering” can also be applied [5]. This extends to other community-based clustering algorithm such as MetaFac [8], Louvain Algorithm [9], RG+[10], MCL [11] and K-Clique-Community [12]. Abovementioned algorithms cluster the graph based on its structure and leave out the node-based attributes of the graph, which may improve clustering quality. From the survey, several algorithms had been proposed such as k-SNAP [13] and SA-Cluster [14] for attributed graph clustering. Both algorithms reported a better result comparing to clustering algorithm that find clusters based on the structure of the graph alone [13,14]. There are also works that include the edge attribute in clustering process as can be seen in [6] where the authors applied clustering algorithm on the signed social network. One of the critical weakness of all of these algorithms is the computation complexity is such that they do not scale well when the graph is large. Most of the algorithms were tested against sparse graphs with about 1000-10000 nodes [5, 7, 10, 11, 12] which cannot be compared to the size of biomedical graphs since they are exceedingly larger than a thousand nodes. In fact, we have experimented on the Louvain Algorithm [9] and K-Clique-Community [12], only to find that they cannot handle Erdos-Renyi Random Graph with 10000 nodes and p = 0.025 - 0.1 [15] on a robust notebook (Intel core i7, 16GB RAM). This stresses once more the need of heuristic and metaheuristic algorithms.

As the clustering problem can be reduced to cluster quality metric optimization, mataheuristic algorithms can be applied easily. Examples of metaheuristic algorithms that can be used are Tabu Search and Simulated Annealing [7], Genetic Algorithm, Ant Colony Optimization, Particle Swarm Optimization and other hybrid algorithms [16, 17, 18, 19]. In [17], the authors compared existing metaheuristic clustering algorithms for wireless networks with some of the work used a graph to represent the network. These algorithms tried to solve the energy consumption problem to prolong the network lifetime in a dynamic environment where nodes move from one region to another. Most algorithms also adapted to dynamic environments by allowing a variable number of clusters.

More efforts have been put in to make the metaheuristic based clustering algorithm become more autonomous or parameterless to overcome the parameter problem in many state-of-the-art clustering algorithm. As in [18], the author proposed a genetic algorithm which employs different number of clusters per solution along with four fitness functions that are mostly based on structural aspects. This eliminate the burden of deciding number of cluster and eliminate the bias that may hinder the analysis. Another GA-based-clustering technique was proposed in [19]. The authors Evolutionary Algorithm resulted in a two-phase clustering algorithm where EA is used to refine the solutions before feeding to GA and improve the end result. Yet again, the fitness function is based on the graph structure (i.e. ratio of Intra- and Inter-connectivity). Even though many efforts have been focusing on clustering techniques, little has been done on improving the current quality metrics such as Modularity, Connectivity, and Coverage [10, 16].

In [20], the authors explore user-guided clustering. In this work, the seed clusters are selected based the consolidation of multiple criteria from multiple users. Then the local clustering algorithm is applied on each seed clusters. This idea appears to be particularly suited for the proposed use case.

4. Conclusion and research directions

There exist several algorithms for hypergraph clustering. They differ in terms of both problem complexity and objectives. Each algorithm has its own limitation and shortcomings. Therefore, improvement can be made and there are scenarios where existing algorithms cannot offer satisfactory solutions.

This article proposes an approach where hidden knowledge may be extracted by an attempt to cluster the Parkinson Disease map (i.e. the hypergraph). The approach comprises the following characteristics :

1. Meta-heuristic-based approach: Due to the NP -hard nature of clustering, meta-heuristic is a suitable approach to tackle the complexity and the large size of the Parkinson Disease map.
2. Node-based and edge-based knowledge extraction: The result of the clustering of the hypergraph of the Parkinson Disease map ought to enable knowledge extraction hidden in the nodes and edges.
3. Multi-objective approach: Assessing the clustering quality is a multi-objective problem which implies a hierarchical approach, a combination or a pareto-based approach for the various objectives and metrics.
4. Expert-driven search. As many possibilities of clusters as well as traversal paths may exist, expert knowledge has to be incorporated in order to find satisfactory solutions

We intend to design such algorithm as upcoming work and collaborate the assessment with biologists.

References

[1] C. Pizzuti and S. E. Rombo (2014). Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics. 30(10), 1343-1352.

[2] S. Parthasarathy, S. Tatikonda and D. Ucar (2010). A Survey of Graph Mining Techniques for Biological Datasets. Managing and Mining Graph Data, 40, 547-580.

[3] OpenBEL Wiki. (n.d.). Retrieved May 4, 2016, from http://wiki.openbel.org/display/home/Home

[4] K. A. Fujita, M. Ostaszewski, Y. Matsuoka, S. Ghosh, E. Glaab, C. Trefois, I. Crespo, T. M. Perumal, W. Jurkowski, P. M. A. Antony, N. Diederich, M. Buttini, A. Kodama, V. P. Satagopam, S. Eifes, A. del Sol, R. Schneider, H. Kitano, R. Balling (2013) Integrating pathways of Parkinson's disease in a molecular interaction map. Molecular Neurobiology, 49, 88–102.

[5] S. E. Schaeffer (2007). Survey: Graph clustering. Computer Science Review, 1(1), 27-64.

[6] K. Y. Chiang, C. J. Hsieh, N. Natarajan, I. S. Dhillon and A. Tewari (2014). Prediction and Clustering in Signed Networks: A Local to Global Perspective. Journal of Machine Learning Research, 15, 1177-1213

[7] A. Trifunović (2006). Parallel Algorithms for Hypergraph Partitioning (Doctoral Dissertation). Retrieved from https://www.doc.ic.ac.uk/~wjk/publications/trifunovic-2006.pdf

[8] Y. R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram and A. Kelliher (2009). MetaFac: community discovery via relational hypergraph factorization. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 527-536.

[9] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10008.

[10] M. Ovelgonne and A. Geyer-Schulz (2010). Cluster Core and Modularity Maximization. 2010 IEEE International Conference on Data Mining Workshops, 1204-1213.

[11] U. Brandes, M. Gaertler and D. Wagner (2003). Experiment on Graph Clustering Algorithms. Algorithms – ESA 2003, 2832, 568-579.

[12] G. Palla, I. Derényi, I. Farkas and T. Vicsek (2005). Nature, 435, 814-818.

[13] Y. Tian, R. A. Hankins and J. M. Patel (2008). Efficient Aggregation for Graph Summarization. Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 567-580.

[14] Y. Zhou, H. Cheng and J. X. Yu (2009). Graph Clustering Based on Structural/Attribute Similarities. Proceedings of the VLDB Endowment, 2(1), 718-729.

[15] A. A. Hagberg, D. A. Schult and P. J. Swart (2008). Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference, 11–15.

[16] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, A. C. Ponce Leon F. de Carvalho (2009). A Survey of Evolutionary Algorithms for Clustering. IEEE Transactions on Systems, Man, and Cybernetics, 39(2), 133-155.

[17] M. M. Moshizi, V. K. Bardsiri and E. Heydarabadipour (2015). The application of Meta-Heuristic based Clustering techniques in Wireless Sensor Network. International Journal of Control and Automation, 8(3), 319-328.

[18] G. Bello-Orgaz, H. D. Menéndez and D. Camacho (2012). Adaptive K-Means Algorithm for Overlapped Graph Clustering. International Journal of Neural Systems, 22(5), 1250018.

[19] J. Kohout and R. Neruda (2013). Two-Phase Genetic Algorithm for Social Network Graphs Clustering. 27th International Conference on Advanced Information Networking and Applications Workshops, 197-202.

[20] J. Cao, S. Wang, F. Qiao, H. Wang, F. Wang, P. S. Yu (2016). User-Guided Large Attributed Graph Clustering with Multiple Sparse Annotations. Advances in Knowledge Discovery and Data Mining, 9651, 127-138.