Overlapping Community Detection using Flow based Rough Set Theory and Firefly Algorithm

Report submitted in fulfillment of the requirements
for the B.Tech Project of

Third Year B.Tech.

by

Vempalli Mugenna Gari Madhava Reddy
Sai Kiran Anumalla

Under the guidance of

Dr.Amrita Chaturvedi

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India
June 2020

Dedicated to

My parents, teachers,.....

Declaration

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.

2. The work has not been submitted for any project.

3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi

Date:16.06.2020

Vempalli Mugenna Gari Madhava Reddy
Sai Kiran Anumalla
B.Tech Student
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

*This is to certify that the work contained in this report entitled "**Overlapping Community Detection using Flow based Rough Set Theory and Firefly Algorithm**" being submitted by **Vempalli Mugenne Gari Madhava Reddy** and **Sai Kiran Anumalla** (**Roll No. 17075056,17075051**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of my supervision.*

Place: IIT (BHU) Varanasi
Date:16.06.2020

Dr.Amrita Chaturvedi
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Acknowledgments

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and our institute. We would like to extend our sincere thanks to all of them. We are highly indebted to our project supervisor Dr. Amrita Chaturvedi for her guidance and constant supervision as well as for providing necessary information regarding the project and also for her support in completing the project.

Place: IIT (BHU) Varanasi

Date: 16.06.2020

Vempalli Mugenna Gari Madhava Reddy

Sai Kiran Anumalla

Abstract

Community detection in complex network has become a vital step to understand the structure and dynamics of networks in various fields. However, traditional node clustering and relatively new proposed link clustering methods have inherent drawbacks to discover overlapping communities. Node clustering is inadequate to capture the pervasive overlaps, while link clustering is often criticized due to the high computational cost and ambiguous definition of communities. So, overlapping community detection is still a formidable challenge. In this work, we propose a Overlapping Community Detection using Flow based Rough Set Theory and Firefly Algorithm which mainly focuses on detecting various communities for a given graph. We use features of the nodes in graphs to determine the flow based similarities and tolerance classes for each nodes.We use these classes to determine the communities using ranking system.

Contents

Chapter 1

Introduction

1.1 Overlapping Communities

With the development of complex network in various fields including biological organisms and human society, community detection has become a vital step to understand the structure and dynamics of networks. Although no common definition of community has been agreed upon, it is widely accepted that a community should have more internal than external connections.

However, many real networks have communities with pervasive overlaps. For example, a person belongs to more than one social group such as family group and friend group. So, these objects should be divided into multiple groups, which are known as overlapping nodes. The aim of overlapping community detection is to discover such overlapping nodes and communities.

Most existing optimization-based community detection algorithms are only applicable to disjoint community structure. However, it has been shown that in most real-world networks, a node may belong to multiple communities implying overlapping community structure.

1.2 Firefly Algorithm

In modern numerical optimization problem like, NP-hard problems, biologically inspired algorithms are most powerful.[1] Firefly algorithm is one of the recent nature-inspired, metaheuristic algorithms developed by Yang in 2008. this metaheuristic algorithm is embraced to enhance the performance of the supervised feature selection method.

In our project of Overlapping community detection we take advantage of above firefly algorithm.

1.3 Tolerance rough sets

The tolerance rough set (TRS), which was developed on the basis of rough set theory, was found to handle continuous attributes effectively. A number of researchers have addressed applications of TRS to pattern classification. In a traditional TRS, the tolerance classes are determined using a tolerance relation, which is commonly defined by a simple distance measure [47] which indicates the proximity of any two patterns distributed in feature space.

1.4 Flow based tolerance rough sets

flow-based tolerance rough set (i.e., FTRS), incorporates flows among patterns into the similarity measure[2]. A flow represents the intensity of preference for one pattern over another pattern.

FTRS not only deals with continuous attributes but also uses a flow-based similarity measure that considers preference information among patterns using pairwise comparisons.The process of a FTRS-based classifier (FTRSC) consists of the deter-

mination of a net flow, a flow-based tolerance class, a FTRS, and a class label for each pattern.

we employ this flow-based rough set theory[2] to exploit it's advantages over traditional Tolerance rough sets (TRS).

1.5 Objective functions

Objective function (i.e., fitness function) quantifies the optimality of a solution.intra and inter objective functions are such objective functions.

These two functions have the potential to balance each other's tendency to increase or decrease the number of communities, which enables the use of a representation that does not fix the number of communities.Modularity is combination of these functions which is used to get optimal solutions for MOCD(Multi Objective Community Detection).

we utilize the advantage of these objective functions in our project to get optimal communities all for overlapping community detection.

Chapter 2

Related Research

2.1 Categories

In the past few years, many different approaches, such as hierarchical clustering, spectral clustering and optimization based algorithms have been proposed to uncover community structure in networks. These methods restrict a node to belonging to only one community and therefore result in some computational advantages. However, for real networks having complex overlapping community structures, these methods are obviously inadequate in identifying communities with overlaps. For this reason, overlapping community detection has drawn lots of attention. Generally speaking, existing overlapping community detection approaches could be divided into two categories: node based algorithms (node clustering) and link based algorithms (link clustering).

2.2 Researches

There are few methods that deals to get overlapping communities in complex networks like symmetric binary factorization [3],multi-objective evolutionary algorithms [4] and some efficient algorithms[5].

symmetric binary factorization model allows us not only to assign community memberships explicitly to nodes, but also to distinguish outliers from overlapping nodes[3]. In addition it is evaluated with modified partition density for quality of community structures and to determine the most appropriate number of communities.

Most existing multi-objective optimization-based algorithms[6] are only applicable to disjoint community structure, which means that a node in the network only belongs to a single community, this MOEA based approach deals with it to find overlapping communities in complex networks.

The multi-objective evolutionary algorithm approach[4] is based on based on the framework of non-dominated sorting genetic algorithm based on MOEA utilizing link-based adjacency representation of overlapping community structure and a population initialization method based on local expansion by proposing MOEA-OCD.

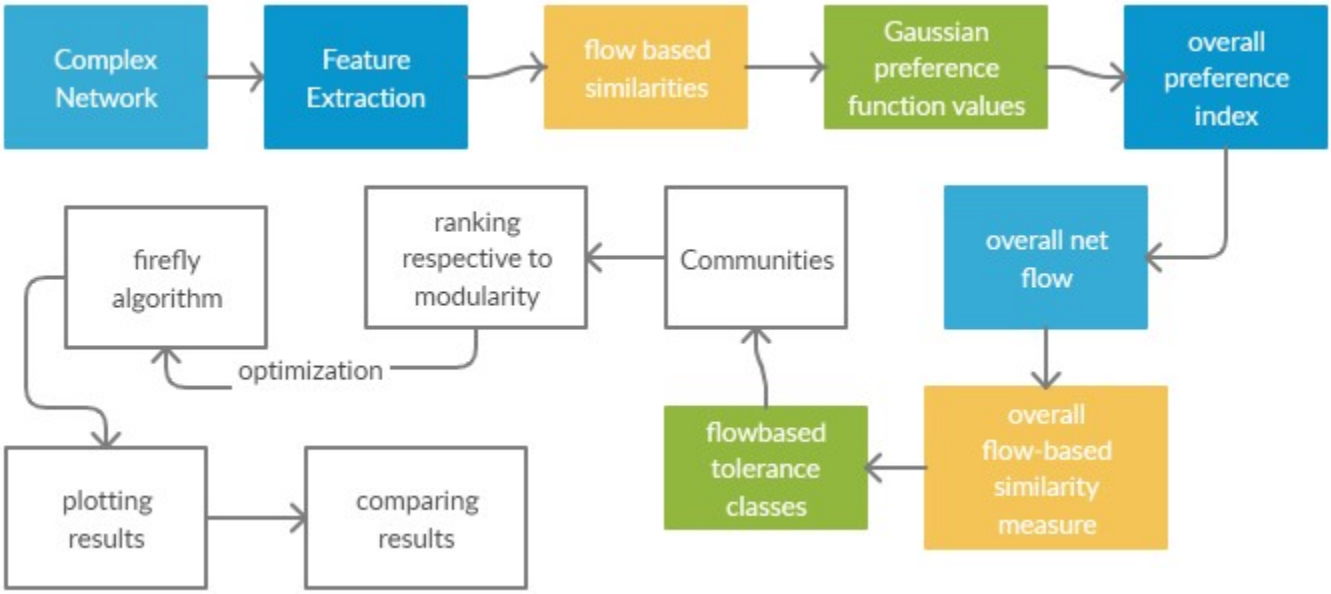
The MOEA-OCD algorithm achieves the tradeoffs between the two objective functions and requires no prior information about the network such as the number and sizes of communities.

There are some algorithms where their key strategy is to mine a node with the closest relations with the community and assign it to the community[5].Some researches also involved in using Multi-objective enhanced firefly algorithms for community detection in complex networks[7] but we try to detect overlapping communities using flow based rough set theory and firefly algorithm hoping to get optimal results than it's predecessors.

Chapter 3

Methodology

3.1 Workflow



Above block diagram represents our workflow for overall project and the colored blocks represents our current progress of the project.

3.2 Feature Extraction

Initially we extract the following features for each node in our complex network and store these feature vectors in a matrix for further processing.

- Direct Reachable Nodes
- Node Ranking - pagerank
- betweenness centrality
- Clustering

3.3 Flow Based Tolerance Rough Set

We will calculate the flow-based similarities of each node with the help of the feature vector that we obtained for each node in our complex network.

we apply below similarity formula to get flow based similarities

$$H(s_k) = 1 - e^{-d_k^2/2\sigma_k^2}$$

where k greater than 0 is a preference parameter that can be determined by decision-makers and dk = xik xjk [1]. Partial preference index pk(xi, xj) = H(sk) for xik xjk, where pk(xi, xj) [0,1] is a measure of the intensity of the preference for xi over xj for attribute k. pk(xi, xj) = 0 when xik ¡ xjk , pk(xi, xj) represents the leaving flow from xi to xj, whereas pk(xj, xi) represents the entering flow from xj to xi for attribute k. When pk(xi, xj) greater than 0, pk(xj, xi) = 0, and vice versa.

The net leaving flow + phi k (xi) for attribute k is defined by summing the intensity of the preference for xi over the training patterns in T for attribute k as follows:

$$\phi_k^+ = \frac{1}{|T|}\sum_{x_j \in T} p_k(\mathbf{x}_i, \mathbf{x}_j) \qquad [2]$$

similarly - phi k (xi)

The net flow k(xi) for attribute k is difference between phi k + and phi k -

$$\phi_k = \phi_k^+(\mathbf{x}_i) - \phi_k^-(\mathbf{x}_i)$$

An overall preference index p(xi, xj) can be further derived using the weighted average of pk(xi, xj) (1 less than equal to k less than equal to n)

$$\phi^+(\mathbf{x}_i) = \frac{1}{|T|}\sum_{x_j \in T} p(\mathbf{x}_i, \mathbf{x}_j)$$

$$\phi^-(\mathbf{x}_i) = \frac{1}{|T|}\sum_{x_j \in T} p(\mathbf{x}_j, \mathbf{x}_i)$$

A flow-based similarity measure Sf a(xi, xj) with respect to Rf a (where Rf a is a

flow-based tolerance relation with respect to attribute a) can be defined as

$$S_A^f(\mathbf{x}_i, \mathbf{x}_j) = \left|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\right|$$

The same definition can be used for all attributes. The relation between Rf a and Sf a(xi, xj) is as follows:

$$\mathbf{x}_i R_A^f \mathbf{x}_j \Leftrightarrow S_A^f(\mathbf{x}_i, \mathbf{x}_j) \leq \tau_A^f$$

an overall flow-based similarity measure Sf A(xi, xj) could be defined

A flow based tolerance class FTC(xi) of xi can be generated by considering those patterns that have a flow-based tolerance relation with xi as:

$$FTC(\mathbf{x}_i) = \{\mathbf{x}_j \in U | \mathbf{x}_i R_A^f \mathbf{x}_j\}.$$

3.4 Ranking

The first objective function minimizes 1 minus the intra-link strength of a partition, and it is called intra objective.

$$intra(C) = 1 - \sum_{c \in C} \frac{|E(c)|}{m} \quad [6]$$

The second objective function minimizes the inter-link strength of a partition, and it is called inter objective.

$$inter(C) = \sum_{c \in C} \left(\frac{\sum_{v \in c} deg(v)}{2m} \right)^2$$

According to the two definitions, we can deduce that

Q(C) = 1 - intra(C) - inter(C)

Highest ranked community has minimizes the values both intra(C) and inter(C) and are Ranked according to Q(C).these ranking is a measure of overlapping of the communities. firefly algorithm will be implemented to get optimized communities.

Chapter 4

Conclusion

4.1 progress

In this work we propose the flow based rough set theory and firefly algorithm. here initially feature vectors of the nodes are extracted and then they are send to calculate the similarities of nodes in complex network which are flow based[2] and we tried to find the flow based tolerance classes of each node from flow based similarity measure and threshold and by the use of gaussian preference functions.

After finding the subsets of the tolerance classes to get the communities we will find scores of the relative communities which reflects considering the number of internal links and external links of a community through objective functions which represents these properties.Modularity[6] representing objective functions is decided as scoring parameter.more the modularity more the score of those respective communities.

4.2 Future work

these communities from above process will be ranked according to their scores of modularity from objective functions which represents overlapping of the communities.

high score represents the overlapping nature of respective communities.

further the firefly algorithm[1] is applied to get optimal weights for these communities objective function in neural network which optimizes our overlapping community detection problem.

the results will be plotted to get the actual statistical analysis report of the project and will be compared to various research results to get the performance idea of current project expecting to achieve more than it's predecessors.
This is the line of work for our project and development continues in the next semester to achieve state-of-art results.

Bibliography

[1] G. Jothi *et al.*, "Hybrid tolerance rough set–firefly based supervised feature selection for mri brain tumor image classification," *Applied Soft Computing*, vol. 46, pp. 639–651, 2016.

[2] Y.-C. Hu, "Flow-based tolerance rough sets for pattern classification," *Applied Soft Computing*, vol. 27, pp. 322–331, 2015.

[3] Z.-Y. Zhang, Y. Wang, and Y.-Y. Ahn, "Overlapping community detection in complex networks using symmetric binary matrix factorization," *Physical Review E*, vol. 87, no. 6, p. 062803, 2013.

[4] Z. Yuxin, L. Shenghong, and J. Feng, "Overlapping community detection in complex networks using multi-objective evolutionary algorithm," *Computational and Applied Mathematics*, vol. 36, no. 1, pp. 749–768, 2017.

[5] D. Chen, Y. Fu, and M. Shang, "An efficient algorithm for overlapping community detection in complex networks," in *2009 WRI Global Congress on Intelligent Systems*, vol. 1. IEEE, 2009, pp. 244–247.

[6] C. Shi, Z. Yan, Y. Cai, and B. Wu, "Multi-objective community detection in complex networks," *Applied Soft Computing*, vol. 12, no. 2, pp. 850–859, 2012.

[7] B. Amiri, L. Hossain, J. W. Crawford, and R. T. Wigand, "Community detection in complex networks: Multi–objective enhanced firefly algorithm," *Knowledge-Based Systems*, vol. 46, pp. 1–11, 2013.