ImprovingResults ofMixture Model Based Graph Clustering Methods Using Evolutionary Algorithms

Ghasem Elyasi, Parham Moradi and Fardin Akhlaghian
Department of Electrical and Computer Engineering
University Of Kurdistan
Sanandaj-Iran
ghasem.elyasi@gmail.com,p.moradi@uok.ac.ir, F.akhlaghian@uok.ac.ir

***Abstract*—In the recent years there has been an interest within the physics community in the properties of networks of many types. Graph clustering is the process of identifying the network structure in terms of grouping the vertices of a graph into clusters taking into consideration the edge structure of the graph that in such a way there should be many edges within each cluster and relatively few between the clusters. Based on high computational cost, the classical algorithms will slow much since data size in real application increases rapidly. In such a situation, model based graph clustering algorithms are an efficient alternative to classical ones. The performance of the model based graph clustering algorithms depends on the correct initial parameter setting. We areproposedan evolutionary algorithm to find proper values for the model based graph clustering algorithms. The proposed method is tested on both simulated and real data sets and gave improving results in comparison with random parameter setting.**

Keywords-Mixture model; Genetic Algorithm; Random Graphs; Graph clustering

I. INTRODUCTION

Recently, graph clustering algorithms are applied to discover structural properties of networks of different types including World Wide Web, social, biological, email, citation and transportation networks and software.Nodes in real networks are structured into communities or groups. The process of finding these groups of nodesiscalled graph clustering. Many different approaches have been proposed for community detection in networks. These approaches can be divided in two main classes. The first class of proposed approaches are algorithmic which is based on spectral graph theory, optimization of centrality measure or other approaches. The second category of proposed approaches is model-based or based on random graph model.Model-based approachesrely on a statistical model of network edges and vertices.

Graph nodes in the random graph model are given and edges can be considered as random variables. The most famous random graph model is Erdös–Rényimodel where each pair of nodes is connected with the probabilityof$p$. But this model does not present most of the real networks properties such as degree distribution and clustering coefficient.The alternative modelisErdös–Rényimixture model [1-4]which is proposed to overcome the limitation of the previous random graph models in real networksmodeling. This model is associated with EM algorithm and allows capturing structure of a network and particularly detect communities,assumednodes are spread over an unknown number of hiddencommunities,which have a low density of edges between groups and a high density of edges within each group[5]. Unfortunately, the performance of the model depends on initial values of corresponding parameters.An evolutionary algorithmis proposed to estimate proper initial values for the model to improve the results of themodelin this paper. The experiments of different simulated and real networks present the proposed algorithm improves the performance of the main Erdös–Rényi mixture model.

The rest of the paper followed by. In the next section Erdös–Rényi mixture model theory is presented. In the third section the proposed evolutionary algorithm is presented and the results of itdescribed in the fourth section. And it is concluded in the fifth section.

A. Notation

We consider an undirected graph without self-loop with n vertices and define the variable X_{ij}whichequals 1if vertices i and j are connected and to zero otherwise. In our case $X_{ii}= 0$ why graphs are without self-loop. However, what we propose in the following method can be generalized to directed graphs ($X_{ij}\neq X_{ji}$) with self-loops ($X_{ii}\neq0$). We also denote K_i the degree of vertex i, i.e. the number of edges connecting it:

$$K_i = \Sigma_{j\neq i} X_{ij} \quad (1)$$

II. ERDÖS–RÉNYI MIXTURE MODEL

TheErdös–Rényimixture model proposed by Daudin[1]contains a mixture of distributions assumes vertices are partitioned into Q classes with prior probabilities $\{\alpha_1,\ldots,\alpha_Q\}$. Moreover, a sequence of independent hidden variables $\{Z_{iq}\}$ (with$\sum_q Z_{iq}=1$) which are exists to indicate label of vertices to classes such that:

$$\alpha_q = \Pr\{Z_{iq} = 1\} = \Pr\{i\in q\}, \qquad \text{with } \Sigma_q\alpha_q = 1 \quad (2)$$

Then we denote $\pi_{ql}$the probability for a vertex from class q to be connected to a vertex from class l.

To estimate the model parameters a variationalapproachbased on EM [1] algorithm has been proposed[6]. This approach is used to perform an approximation of maximum likelihood on the parameterswhichaims at optimizing a lower bound of the following likelihood:

$$\log\mathcal{L}(\mathcal{X}, \mathcal{Z}) = \Sigma_i\Sigma_q Z_{iq}\log\alpha_q + \tfrac{1}{2}\Sigma_{i\neq j}\Sigma_{q,l} Z_{iq}Z_{jl}\log b(X_{ij}; \pi_{ql}) \quad (3)$$

where $b(X; \pi) = \pi^x(1-\pi)^{1-x}$, $\mathcal{X} = \{X_{ij}\}_{i,j=1,\ldots,n}$,

$$\mathcal{Z} = \{Z_{iq}\}_{i=1,\ldots,n}^{q=1,\ldots,Q}$$

$\mathcal{X}$describeset of all edges and $\mathcal{Z}$ set of all indicator variables for vertices,For more details,referto[1]. The estimation parts of the EM algorithm starts with some initial values $\{t_{iq}(0)\}$that$t_{iq}$is an approximation of $\Pr\{Z_{iq} = 1|\mathcal{X}\}$ and parameters t_i, α and π are iteratively updated are followed:

$$(\alpha^{(h+1)}, \pi^{(h+1)}) = \arg\max_{(\alpha,\pi)} \mathcal{J}(R_{\mathcal{X}}; \{t_i^{(h)}\}, \alpha, \pi)$$
$$(t^{(h+1)}) = \arg\max_{(t_i)} \mathcal{J}(R_{\mathcal{X}}; \{t_i\}, \alpha^{(h+1)}, \pi^{(h+1)}) \quad (4)$$

where $\mathcal{J}(R_{\mathcal{X}}; \{t_i^{(h)}\}, \alpha, \pi)$ is an estimation of $\log\mathcal{L}(\mathcal{X})$ with parameters $t_i^{(h)}$, α, π.

This EM algorithm is a local optimization algorithm and its performance depends strongly on the proper setting of its initial parameter values[5]. Therefore, improper initial values of the parameter,can acquire poor results consequently.

III. PROPOSEDALGORITHM

Asitmentioned, all local optimization algorithms are very sensitive about initial values of their parameters. Unsuitable initial values of MixNet [2] parametersacquirelow accuracy detection of the correct community of each vertex. We assume each vertex of the given graph has a cluster.

In this section an evolutionary approach is proposed tosetinitial values of the Erdös–Rényi mixture model parameters. Following the previous section, these parameters are π, $\alpha$and t with the following limitations.

$$\Sigma_{l=1,\ldots,Q}\, t_{il} = 1 \quad (5)$$

$$\Sigma_{l=1,\ldots,Q}\, \pi_{ql} = 1 \quad (6)$$

$$\Sigma_{l=1,\ldots,Q}\, \alpha_l = 1 \quad (7)$$

[1] Expectation-Maximization

[2] Other name of Erdös–Rényi mixture model

A. Representing the Chromosom

Given the variational parameters $\{t_i\}$, the values of parameters α and π are

$$\alpha_q = 1/n\ \Sigma_i t_{iq}\ ,\ \pi_{ql} = \Sigma_{i\neq j} t_{iq}t_{jl}X_{ij} / \Sigma_{i\neq j} t_{iq}t_{jl} \quad (8)$$

This equation is derived by the Eq. 6 and 7.According to definition $t$is a 2-Dimension matrix. The rows are corresponding to vertices and the columns are corresponding to clusters andt_{iq}shows the probability of the vertex *i*belongincluster*q*. Therefore,thechromosome structure is used to estimate t is 2D, see Fig. 1.Moreover, this representationmustsatisfyingEq.5.

B. Fitness Function

The ultimate goal of the algorithm is selecting the best initial values of parameters that get maximum accuracy detection of the correct cluster of each vertex for a given graph.Therefore, the evaluation of current selected values of the parameters should be based on the correct clustered vertices of a given graph. According to likelihood definition, likelihood is a function of the parameters of a statistical model and the likelihood of a set of parameters values given some observed data equal the probability of those observed outcomes given those parameter values[7, 8].We have used complete likelihood (Eq. 3) as fitness function of the proposed genetic algorithm. High value of the fitness shows a good chromosome.

C. Genetic operators

1) Reproduction: This opreator is performed with probabilityP_r.A chromosome is randomly selected from the current population and copied into the new population without any modification.

*2) Crossover:*Thisopreator is performed with probabilityP_c.To cope with 2D representaion of chromosomes, the uniform crossover[9] of the conventional GAisadapted. First,two different random parents (chromosomes) from the current population selected.We consider each chromosome contains two parts, these two parts can be have non-equal size.However two correspondingparts in each parent must have same size, see Fig. 2. Then weexchnage the first part of a parent and the first part of other parent, see Fig. 3.

3) Mutation: the mutation operator of the proposed approach replace the values of some genes with new values from the values space.

*a) Mutation in row values level:*This opreator is performed with probability P_{mv}.First, one random chromosome from current population will be selected. Then,one random row (one veretx) fromthe chromosome will beselected and willreplace its values using a random probability number (0 < number < 1), see Fig. 4.Moreover, the sum of the values of each row is equal 1 (Eq. 5).

*b) Mutation in rows level:*In order to acquire more variety in the population, we can using another type of

mutation. We call it mutation in level of rows. After selection a chromosome randomly, two different random rows (vertices in our case) from the chromosome will be selected. Then, exchange corresponding values of them, see Fig. 5.This opreator is performed with probabilityP_{mr}. You can see the evolutionary algorithm toghther in the mixture algorithm[1] in Fig. 6.

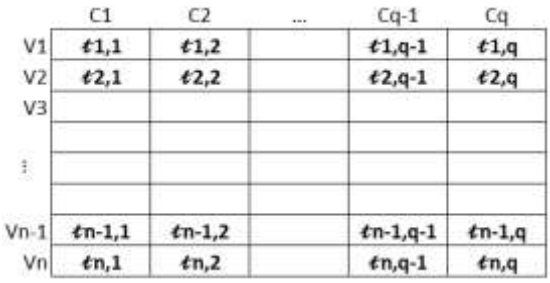

Figure 1. Representing the chromosome with n vertices and q clusters.

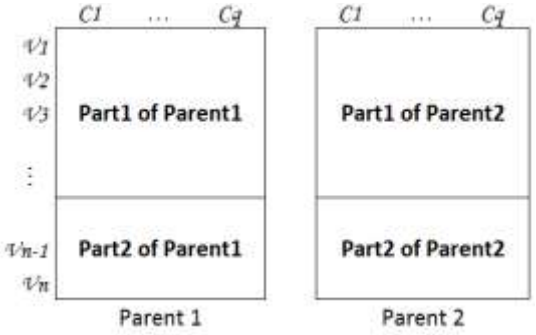

Figure 2. Two parts of each parent before performing the crossover oprator. Each two parts in same chromosome can be have non-equal size, however two corresponding parts in each chromosome must have the same size. In this figure Size(Part1 of Parent1) = Size (Part1 of Parent2), Size(Part2 of Parent1) = Size (Part2 of Parent2).

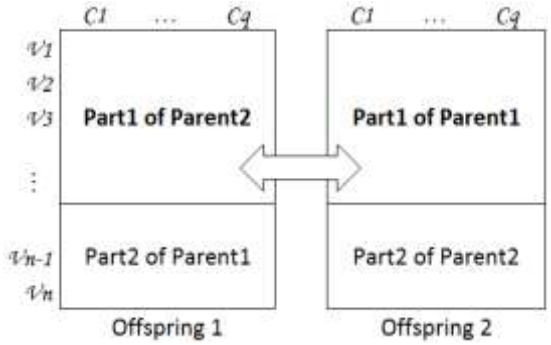

Figure 3. Two parts of each offspring after performed the crossover oprator. First two corresponding parts of the parents exchange to produce two offsprings.

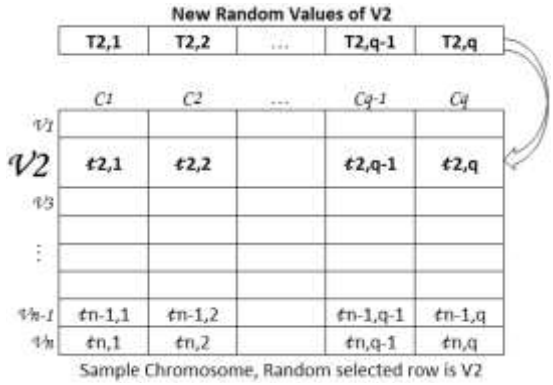

Figure 4. Mutation in row values level. Replaceall values of row byusing a random probability number which each value is a probability between 0 and 1. Also the sum of the values of each row is equal 1.

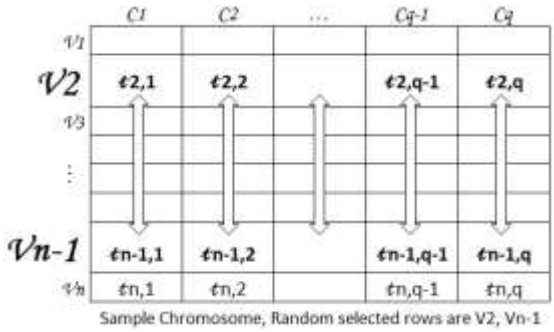

Figure 5. Mutation in rows level. Exchange corresponding values of two different random selected rows of the parent to produce the offspring.

IV. RESULTS

In this section, we applythe proposed method to two real networks to assess how well the method discovers node clusters. We consider the "karate club" network which studied in [10]. The network is a social network of friendships between 34 members of a karate club at a U.S university as it described by Wayne Zachary in 1977. The vertices of the network spread on four communities, it is made up of n = 34 vertices and the total number of edges is 78, refer to [10-12] for more details.

As the first experiment we applied the proposed method in this paper withMixNet compared their clustering results when only used MixNet. Table 1 showsthe method together with the MixNet make better fits than only using the MixNet. Generally, the number of vertices which clustered in correct clusters has increased using the GA. As you can see in the Table 1, in the karate club network, the clustering accuracy is 0.6176 using proposed method while the accuracy is 0.4117 only using MixNet.

The best fitness of each generation in proposed GA[3] has increased ascending order, Fig. 7. Moreover, the maximum likelihood of the MixNet model has got a

[3]Genetic Algorithm

Data: Connectivity Matrix **X** with **N** vertices
/* Random Initialization of the parameters and the hidden variable */
foreach chromosome **in** firstPopulation **do**
$\theta^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_Q^{(0)}, \pi_{11}^{(0)}, \ldots, \pi_{QQ}^{(0)}, \tau_{11}^{(0)}, \ldots, \tau_{NQ}^{(0)})$
/* Computation of chromosome fitness according to Eq. 3 */
fitness = $(Rx; \{t_i^{(h)}\}, \alpha, \pi)$
while GenerationCounter <= MaxGeneration **do**
while ChromosomeCounter <= PopulationSize **do**
/* select one of genetic operators of part C, section III randomly */
If Crossover is selected **then**
/* to perform crossover according to Fig. 2, Fig. 3 on two random selected parents and produce two offsprings */
ChromosomeCounter += 2
If Mutation1 is selected **then**
/* to perform mutation on row values according to Fig. 4 on one random selected parent and produce one offspring */
ChromosomeCounter += 1
If Mutation2 is selected **then**
/* to perform mutation on row level according to Fig. 5 on one random selected parent and produce one offspring */
ChromosomeCounter += 1
If Reproduction is selected **then**
/* select a parent randomly and consider it as new offspring */
ChromosomeCounter += 1
/* Computation of offsprings fitness according to Eq. 3 */
fitness = $\mathcal{J}(Rx; \{t_i^{(h)}\}, \alpha, \pi)$
Result: Select best chromosome of last generation as parameters initial values and initial posterior probabilities t_{iq}, now using these values as mixture model initial parameters values.

Figure 6. Proposed Evaloutionary Algorithm Pseudocode.

TABLE I. TO COMPARE CLUSTERING ACCURACY. AS YOU CAN SEE, OUR PROPOSED METHOD REACH BETTER RESULTS THAN THE ERDÖS–RÉNYI MIXTURE

	GA and Erdös–Rényi mixture	**Erdös–Rényi mixture**
Number of Vertices	***Correct Clustered***	***Correct Clustered***
34	21	14
Accuracy	0.6176	0. 4117

TABLE II. AVERAGE OF COMPLETE AND INCOMPLETE LIKELIHOOD TOGETHER ENTROPY AVERAGE OF THE MIXTURE MODEL WHICH OBTAINED BY RUNNING THE ALGORITHM 100 TIMES ON KARATE-CLUB NETWORK.

	Erdös–Rényi mixture-GA	**Erdös–Rényi mixture-Random**
log -Complete	-173.7508606	-180.1669568
log -Incomplete	-173.030522	-179.0590337
log -Entropy	-0.720338621	-1.107923149

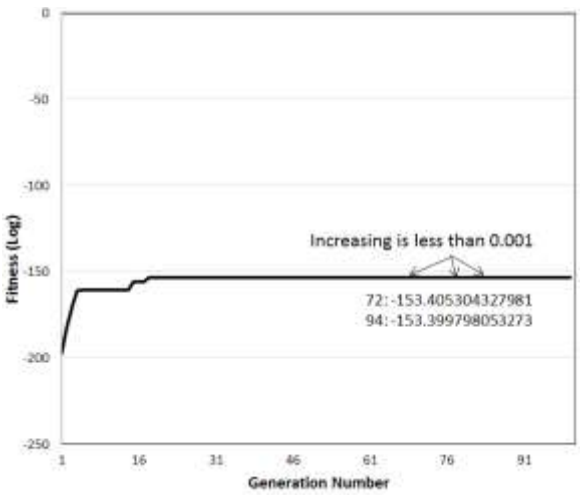

Figure 7. The best fitness of each generation.The generations, bigger than 20, improving is less than 0.001at most.

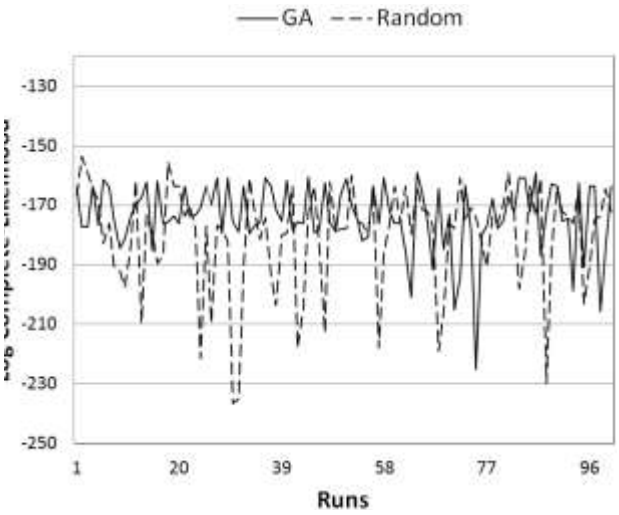

Figure 8. The Complete Likelihood of the mixture model with 100 times runs.

significant improvement using our proposed GA, i.e. now the MixNet model make a better fit on the network, see Table 2.

We have shown complete likelihood of the mixture model with running 100 times in Fig. 8. We have considered two methods, first one used GA and second one used the random methods to initial the parameters of the models. As you can see complete likelihood variations using GA is smoothly.

TABLE III. AVERAGE OF COMPLETE AND INCOMPLETE LIKELIHOOD TOGETHER ENTROPY AVERAGE OF THE MIXTURE MODEL WHICH OBTAINED BY RUNNING THE ALGORITHM 100 TIMES ON DOLPHINS NETWORK.

	Erdös–Rényi mixture-GA	**Erdös–Rényi mixture-Random**
log -Complete	-462.6148157	-473.1966331
log -Incomplete	-461.3496899	-471.965267
log -Entropy	-1.165125806	-1.23136613

We apply the proposed method to other networks, such as "Dolphin social network", which is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand[13].This graph includes 62 vertices and 159 edges. We acquired improvements in the results of the Erdös–Rényi mixture mode.You can see its results in Table 3.

V. CONCLUSION

As you notice in this paper a local optimization algorithm need suitable initial values for its parameters.The proposed evolutionary algorithm estimate initial values of the parameters of theErdös–Rényi mixture model. This model classifies the vertices of a network. The proposed method tries to achieve better values from the values space. These resulted values can improve clustering accuracy of the model. Inthe feature, we plan to investigate Online clustering algorithms to find better clustering algorithms.

REFERENCES

[1] J. J. Daudin, et al., "A mixture model for random graphs," Statistics and computing, vol. 18, pp. 173-183, 2008.

[2] O. Frank and F. Harary, "Cluster inference by using transitivity indices in empirical graphs," Journal of the American Statistical Association, pp. 835-840, 1982.

[3] T. Jaakkola, "Advanced mean field methods: theory and practice. chapter Tutorial on variational approximation methods," ed: MIT Press, 2000.

[4] M. Newman and E. Leicht, "Mixture models and exploratory analysis in networks," Proceedings of the National Academy of Sciences, vol. 104, p. 9564, 2007.

[5] H. Zanghi, et al., "Fast online graph clustering via Erdős–Rényi mixture," Pattern Recognition, vol. 41, pp. 3592-3599, 2008.

[6] A. P. Dempster, et al., "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), pp. 1-38, 1977.

[7] A. Hald, A History of Mathematical Statistics from 1750 to 1930 vol. 2: Wiley New York, 1998.

[8] J. W. Pratt, "FY Edgeworth and RA Fisher on the efficiency of maximum likelihood estimation," The Annals of Statistics, vol. 4, pp. 501-514, 1976.

[9] M. Mitchell, "An introduction to genetic algorithms, 1996," PHI Pvt. Ltd., New Delhi, 1996.

[10] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences, vol. 99, p. 7821, 2002.

[11] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," Physical Review E, vol. 69, p. 066133, 2004.

[12] V. Nicosia, et al., "Extending the definition of modularity to directed graphs with overlapping communities," Journal of Statistical Mechanics: Theory and Experiment, vol. 2009, p. P03024, 2009.

[13] D. Lusseau, et al., "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behavioral Ecology and Sociobiology, vol. 54, pp. 396-405, 2003.