Multiobjective Evolutionary Community Detection for Dynamic Networks

Francesco Folino
ICAR-CNR
Via P. Bucci 41C
87036 Rende(CS), Italy
ffolino@icar.cnr.it

Clara Pizzuti
ICAR-CNR
Via P. Bucci 41C
87036 Rende(CS), Italy
pizzuti@icar.cnr.it

ABSTRACT

A multiobjective genetic algorithm for detecting communities in dynamic networks, i.e., networks that evolve over time, is proposed. The approach leverages on the concept of evolutionary clustering, assuming that abrupt changes of community structure in short time periods are not desirable. The algorithm correctly detects communities and it is shown to be very competitive w.r.t. some state-of-the-art methods.

Categories and Subject Descriptors

H.2.8 [**Database Managment**]: Database Applications —*Data Mining*; I.2.2 [**Artificial Intelligence**]: Automatic Programming; I.5.3 [**Computing Methodologies**]: Pattern Recognition—*Clustering*

General Terms

Algorithms.

Keywords

Genetic Algorithms, Data Mining, Clustering, Dynamic Networks, Community Detection.

1. INTRODUCTION

The analysis of social network data is gaining an increasing interest because of the capacity of networks to represent the relationships among objects composing many real world systems. More recently a growing attention is focusing on *dynamic* networks, i.e. networks that evolve over time. Dynamic networks capture the modifications of interconnections over time, allowing to trace the changes of network structure at different time steps. Some methods (e.g., [2, 5, 8, 4]) employ the concept of *evolutionary clustering* [1] for catching the cluster evolution in temporal data. Evolutionary clustering groups data coming at different time steps to produce a sequence of clusterings by introducing a framework called *temporal smoothness*. It assumes that abrupt changes of clustering in a short time period are not desirable, thus it *smooths* each community over time.

In this paper we propose a multiobjective approach, named *DYN-MOGA*, to discover communities in dynamic networks by employing genetic algorithms. The detection of community structure with temporal smoothness is formulated as a *multiobjective optimization problem*. The first objective measures how well the clustering found represents the data at the current time. The second objective is the minimization of the temporal cost and it measures the distance between two clusterings at consecutive timesteps.

Experiments confirms the effectiveness of our algorithm w.r.t. some state-of-the-art approaches.

2. MULTIOBJECTIVE EVOLUTIONARY CLUSTERING

Let $\{1, \ldots, T\}$ be a finite set of time steps and $V = \{1, \ldots, n\}$ be a set of individuals or objects. A static network $\mathcal{N}^t$ at time t can be modeled as a graph $G^t = (V^t, E^t)$ where V^t is a set of objects, called nodes or vertices, and E^t is a set of links, called edges, that connect two elements of V^t at time t. Thus G^t is the graph representing a snapshot of the network $\mathcal{N}^t$ at time t. $V^t \subseteq V$ is a subset of individuals V observed at time t. An edge $(u^t, v^t) \in E^t$ if individuals u and v have interacted at time t.

A community (i.e., cluster) in a static network $\mathcal{N}^t$ is a group of vertices $V_i^t \subseteq V^t$ having a high density of edges inside the group, and a lower density of edges with the remaining nodes V^t/V_i^t. Let C^t denote the sub-graph representing a community. A clustering, or community structure, $\mathcal{CR}^t = \{C_1^t, \ldots C_k^t\}$ of a network $\mathcal{N}^t$ at time t is a partitioning of G^t in groups of nodes such that for each couple of communities C_i^t and $C_j^t \in \mathcal{CR}^t$, $V_i^t \cap V_j^t = \emptyset$. A dynamic network is a sequence $\mathcal{N} = \{\mathcal{N}^1, \ldots, \mathcal{N}^T\}$ of static networks, where each $\mathcal{N}^t$ is a snapshot of individuals and their interconnections at time t.

A multiobjective evolutionary clustering problem $(\Omega, \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_h)$ for a static network $\mathcal{N}^t$ can be defined as

$$min\ \mathcal{F}_i(\mathcal{CR}^t), \quad i = 1, \ldots, h \qquad subject\ to \quad \mathcal{CR}^t \in \Omega$$

where $\Omega = \{\mathcal{CR}_1^t, \ldots, \mathcal{CR}_k^t\}$ is the set of feasible clusterings of $\mathcal{N}^t$ at time stamp t, and $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_h\}$ is a set of h single criterion functions. Each $\mathcal{F}_i : \Omega \to \mathcal{R}$ is a different objective function that determines the feasibility of the obtained clustering. Since $\mathcal{F}$ is a vector of competing objectives that must be simultaneously optimized, there is not one unique solution to the problem. The set of all possible solutions are found through the use of *Pareto optimality theory* [3].

In this scenario, we are interested in optimizing the weighted cost function $\alpha \cdot \mathcal{SC} + (1 - \alpha) \cdot \mathcal{TC}$ having two competing ob-

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.
ACM 978-1-4503-0072-8/10/07.

jectives: *(i)* the snapshot cost $\mathcal{SC}$ and *(ii)* the temporal cost $\mathcal{TC}$. Notice that α is used to emphasize one of the two objectives.

Since $\mathcal{SC}$ measures how well a community structure C^t represents the data at time t, we need an objective function that maximizes the number of connections inside each community and minimizes the number of links between the communities. To this end we employ the *community score* introduced in [7] and proved very effective in detecting communities. The second objective must minimize the temporal cost $\mathcal{TC}$, thus we need a metric to measure how similar the community structure $\mathcal{CR}^t$ is w.r.t. the previous clustering $\mathcal{CR}^{t-1}$. To this end we employ the *Normalized Mutual Information (NMI)*, a well known entropy measure in information theory.

Basically, the *DYN-MOGA* algorithm works in this way. Given a dynamic network $\mathcal{N} = \{\mathcal{N}^1, \ldots, \mathcal{N}^T\}$ and the sequence of graphs $\mathcal{G} = \{G^1, \ldots, G^T\}$ modeling it, *DYN-MOGA* starts by partitioning the network $\mathcal{N}^1$ by means of the genetic algorithm that optimize only the first objective (i.e., the community score). For a given number of timestamps, the multiobjective genetic algorithm creates a population of random individuals whose size corresponds to the number of nodes in the current graph G^t. Then, for a fixed number of generations, it iteratively executes the following steps: (1) decode the individuals to generate the partitioning at time step t, (2) evaluate the objective values, (3) assign a rank to each individual according to the Pareto dominance and sorts them, (5) generate a new population of offspring, (6) combine parents and offspring and partition the new pool into fronts and, finally, (7) create a new population with individuals having lower rank.

At the end of each timestamp, *DYN-MOGA* returns all solutions contained in the Pareto front. The best solution is selected by leveraging on the *modularity* criterion introduced in [6].

3. EXPERIMENTAL RESULTS

In this section we study the effectiveness of our approach and compare the results obtained by *DYN-MOGA* w.r.t. the algorithms of Lin et al. [5] (named *FacetNet*) and Kim and Han [4] on synthetic networks for which the partitioning in communities is known.

We used the same dataset adopted by both Lin et al. [5] and Kim and Han [4]. It represents a dynamic network with a fixed number of communities (named SYN-FIX). The network consists of 128 nodes divided into four communities of 32 nodes each. Every node has an average degree of 16 and shares a number z_{in} of links with the nodes of its community, and z_{out} with the other nodes of the network. The dynamicity in the network is introduced by randomly selecting 3 nodes from each community and randomly assigning them to the the remaining ones. For the test purposes, we generated 10 different networks for 10 timestamps and run *DYN-MOGA* on them.

The quality of clustering has been assessed through the NMI that measures the similarity between the true partitions and the detected ones.

Figure 1 shows the average normalized mutual information, over the 10 networks for the 10 timestamps for SYN-FIX when the value of $z_{out} = 5$.

The figure clearly highlights the significantly better results obtained by *DYN-MOGA* w.r.t. both FacetNet and Kim-Han algorithms.

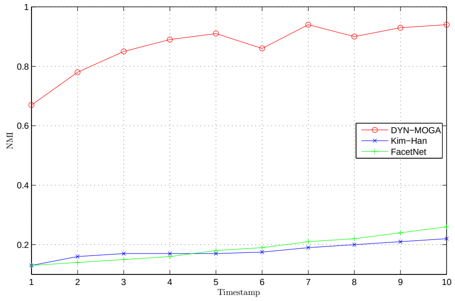

Figure 1: Normalized mutual information of clustering results for SYN-FIX when $z_{out} = 5$.

4. CONCLUSIONS

The paper presented a novel multiobjective genetic algorithm for detecting communities in dynamic networks. The algorithm optimizes the accuracy of the clustering obtained with respect to the data of the current time step, and the drift from one time step to the successive by providing a solution that represents the best trade-off between these two objectives. The approach has been shown to correctly detect communities on both synthetic and real datasets and to be very competitive w.r.t. state-of-the-art methods.

5. REFERENCES

[1] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proc. of 12th Int. Conf. on Knowledge Discovery and Data Mining (KDD'06)*, pages 554–560, 2006.

[2] Y. Chi, X. Song, D.Zhou, K.Hino, and B.L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proc. of 13th Int. Conf. on Knowledge Discovery and Data Mining (KDD'07)*, pages 153–162, 2007.

[3] M. Ehrgott. *Multicriteria Optimization*. Springer, Berlin, 2nd edition, 2005.

[4] M. S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. In *Proc. of 35th Int. Conf. on Very Large Data Bases (VLDB'09)*, pages 622–633, 2009.

[5] Y. Lin, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proc. of the 17th World Wide Web Conference (WWW'08)*, pages 685–694, 2008.

[6] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[7] C. Pizzuti. GA-NET: a genetic algorithm for community detection in social networks. In *Proc. of 10th Int. Conf. on Parallel Problem Solving from Nature (PPSN'08)*, pages 1081–1090, 2008.

[8] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proc. of 13th Int. Conf. on Knowledge Discovery and Data Mining (KDD'07)*, pages 677–685, 2007.