Neurocomputing 279 (2018) 54–62

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Effective community division based on improved spectral clustering

Yi Xu [a], Zhi Zhuang [a], Weimin Li [a,*], Xiaokang Zhou [b]

[a] School of Computer Engineering and Technology, Shanghai University, Shanghai, China
[b] Faculty of Data Science, Shiga University, Hikone, Japan

ARTICLE INFO

Article history:
Received 6 October 2016
Revised 2 March 2017
Accepted 18 June 2017
Available online 21 November 2017

Keywords:
Spectral clustering
Attribute and relationship
Community division
Particle swarm optimization (PSO)
Simulated Annealing (SA)

ABSTRACT

Not only does attribute of nodes affect the effectiveness and efficiency of community division, but also the relationship of them has a great impact on it. Clusters of arbitrary shape can be identified by the Spectral Clustering (SC). However, k-means clustering used in SC still could result in local optima, and the parameters in Radial Basis Function need to be determined by trial and error. In order to make such algorithm better fit into community division of social network, we try to merge attribute and relationship of node and optimize the ability of spectral clustering to get the global solution, thus a new community clustering algorithm called Spectral Clustering Based on Simulated Annealing and Particle swarm optimization (SCBSP) is proposed. The proposed algorithm is adapted to social networking division. In related experiments, the proposed algorithm, which enhances the global searching ability, has better global convergence and makes better performance in community division than original spectral clustering.

1. Introduction

Recently, with the rapid development of Internet technology, big data generated by social networks such as Twitter and Facebook are urgently needed to be explored and analyzed. Due to the large scale of social networks, it is important to turn them into smaller ones through community division methods. Thus, the effectiveness and efficiency of community division has become an important issue in this regard [1]. Dividing social network by employing effective clustering algorithms can have a wide range of application in our real-life world. Social networks have underlying clusters according to individuals' attribute and relationship. Usually, people in a certain cluster have similar interests and personality. Therefore, the common interest in a certain cluster can be deduced by analyzing a small sample in it, and then companies can recommend their products with higher accuracy and lower risk. In the field of education, educators can adapt their educational method to different level of students classified by such algorithms. In criminal psychology, if there are several criminals in a cluster, police can pay more attention to other individuals in this cluster, who may also have high potential of committing a crime. What's more, an effective clustering algorithm can be applied to other fields such as image processing and computer vision.

There are many kinds of clustering algorithms used to detect community. Classical algorithms are popular. For example, k-means, k-medoids and K-Harmonic [2] are based on the node attributes, while Fast-Newman algorithm [3] was proposed on the basis of the modularity focuses on relationship. Hierarchical clustering is another kind of algorithm to divide the community, such as Louvain algorithm for large networks [4], Girvan Newman algorithm [5] and link community [6]. Scott and Smyth combined both Q modularity and spectral clustering together through Q-Laplacian matrix and local greedy heuristic search [7]. Shihua Zhang raised a new modularity function based on generalizing Q modularity and fuzzy c-means clustering [8]. In addition, there are some other community clustering algorithms: Density-based spatial clustering which has a wide range of application in urban planning and marketing [9], graph-skeleton based clustering [10] and Structural Clustering Algorithm for Networks (SCAN) based on the edge density and Clustering Centrality [11].

Most of the clustering algorithms separate the attribute and relationship of nodes while clustering a complex graph. But in fact, they both affect the result of community division. For example, in co-authorship network, author's researching direction and current partnership both affect the frequency and probability of cooperation. And in social network, similar interest and friendship make two users close to each other. Therefore, more exact information can be obtained by taking attribute and relationship of node into account together. To some degree, it can be important for social network research. Actually, there have been some studies in this field [12]. On the basis of these factors, we aimed to improve a clustering algorithm adapted to this kind of community network. Thus, spectral clustering comes into our mind.

* Corresponding author.
E-mail addresses: wmli@shu.edu.cn (W. Li), zhou@biwako.shiga-u.ac.jp (X. Zhou).

https://doi.org/10.1016/j.neucom.2017.06.085
0925-2312/

In this paper, we aim at improving Spectral Clustering (SC) to increase its efficiency in community division of social network. Intensified research on spectral clustering leads to explosive development and improvement over the past several years [13], which is easy to implement and reasonably fast. However, traditional SC is sensitive to the initial data. What's more, after feature decomposition, traditional SC chooses k-means clustering to cluster with the eigenvectors-matrix, while k-means clustering is easy to converge to a local optimal solution. Taking these problems above into account, and given the rapid convergence of Particle Swarm optimization (PSO) and good ability to search the global optimal solution of Simulated Annealing (SA), a new algorithm called Spectral Clustering Based on the Simulated annealing and Particle swarm optimization (SCBSP) was proposed. Actually, the proposed algorithm is adapted to the community division of social network. From experiments which are going to be explained in detail later, SCBSP really do make a step forward in community division of real-life social networks such as Sina Microblog and Facebook.

In our study, we first improve the traditional spectral clustering itself. We implement traditional spectral clustering and apply it to cluster simple real-life social network, in which the performance of traditional SC is not very satisfying. To some degree it is because k-means algorithm is easy to reach a local optima. One of the good replacements of k-means is PSO, which has high convergence speed and good performance in low-dimensional vector space, combined with SA, which has the ability of finding global optima. SCBSP proposed in this paper is based on merging these two algorithms with SC.

Next, we revise the step of preprocessing data in order to make our improved SC better fit into community division. When doing community division, it is important to take both attribute of nodes and relationship between them into consideration. However, traditional similarity matrix just cares about the differences between the attributes of nodes, and it ignores the relationship between nodes in the situation of community division. Thus we define a new similarity matrix which merges attribute and relationship.

Finally, we conduct several experiments on both randomly generated data and real-life social network to evaluate our method. The experiment results show that the method has both high efficiency and high accuracy.

The rest of this paper is organized as follows: Section 2 introduces related work; Section 3 explains our proposed algorithm in detail which begins with a brief review on SC, PSO and SA; The experimental results on random generated data and real-life data from famous social networks are presented in Section 4, and Section 5 concludes this paper.

2. Related work

Social networks are ubiquitous, and researchers have investigated a growing number of data generated by social networks. Yet, most existing measuring methods do not take both attribute and relationship of nodes into consideration, thus they do not fully capture the richness of the information contained in the data [14,15]. Most methods focus on improving the k-means, k-medoids, Newman and Girvan, Density-based methods and so on.

The classical partitioning methods for clustering are k-means and k-medoids, they are easy to implement but are based on complex mathematical theory. These classical algorithms are foundation of many other clustering algorithms. In k-means algorithm, clusters are represented by a mean value and object exchanging stops if the average distance from objects to their cluster's mean value converges to a minimum value [16]. K-medoids algorithm represents each cluster by an actual object in it. However, as is known to all, the original k-means proposed by James MacQueen is easy to converge to a local optimal solution and sensitive to the initial data [14]. In k-harmonic means, harmonic means function which applies distance from the data point to all clustering centers is used to solve the problem that clustering result is sensitive to initial value instead of the minimum distance [2]. Although the problem about initial data is solved, another problem still exists. On the basis, the k-harmonic means was improved by the Simulated Annealing called K-Harmonic Means Clustering with Simulated Annealing [17]. In the K-Harmonic Means Clustering, simulated annealing is used to search the global solution whenever a new result is obtained by k-harmonic means.

In the classical algorithms, the number of clusters must be selected by researchers and is usually tried for times by tests. So, researchers sought for methods which can automatically choose number of clusters. Unlike the k-means that needs to know the number of clusters first of all, an objective function for graph clustering was raised by Newman and Girvan called Q function. In this method, the number of clusters can be automatic selected, which avoid trying cluster numbers for several times before getting a better result. In other word, Q function has higher value when combined with good clustering method. While because of the high complexity of Girvan-Newman algorithm, Newman proposed another algorithm based on the Q function called Fast-Newman algorithm [3,18]. What's more, because spectral clustering is popular for its process of recursively splitting the graph, Scott et al combined both Q function and spectral clustering together. From the experimental results, the two novel algorithms proposed by Scott are efficient and effective [7]. The first algorithm directly searches for global maximum of Q by performing eigenvector decomposition on a matrix called Q-Laplacian matrix, while Newman's algorithm improves the maximum of Q by local iteration. The second algorithm uses a local greedy heuristic search which is similar to Newman's method. At the same time, for the popularity of spectral clustering, a lot of researcher has explored the algorithm, such as the Shi and Malik algorithm, the Kannan, Vempala and Vetta algorithm, the Ng, Jordan and Weiss algorithm [19,20,14] and other more efforts [21–24]. In addition, the ideal of Q function is also applied to identification of overlapping community structure. Shihua Zhang combined a new modularity function based on generalizing Q function and fuzzy c-means clustering [8].

Other researchers sought for methods different from the classical algorithms, they proposed algorithms such as density-based algorithms based on the structure of data. To some degree, density-based methods are the same as Girvan-Newman methods, users don't need to know the number of clusters at the beginning. Density-based methods are more sensitive to initial parameters than Girvan-Newman methods, but they can identify the noise in the data. When several objects are close to each other, they form dense clusters, and they are separated from each other by regions with low density of objects. Thus dense clusters can be detected by finding such low-density regions. DBSCAN (Density-based spatial clustering of application with noise) is the most representative method in this field. In addition, density-based methods can also be applied to spatial clustering, such as clustering in Geo-Social Networks [16]. Because most traditional density-based methods have difficulty in detecting communities of arbitrary and sometimes extreme shape, for example evenly distributed nodes, and they usually have difficulty identifying central nodes and outlying ones, gSkeletonClu (graph-skeleton based clustering) was proposed for community division [10]. The gSkeletonClu algorithm projects the network to its Core-Connected Maximal Spanning Tree (CCMST) and finds its core nodes to cluster this network. This novel algorithm can also avoid the problem of the chaining effect and resolution limit faced by typical MST-based clustering algorithms and modularity-based algorithms.

Recently, much ink has been spilled onto new methods like spectral clustering [13]. These methods focus on improving the