



Increased diagnostic accuracy in social networking communities using clustering method

Authors

Farzaneh Javid

Department of Computer Engineering/ Islamic Azad University, Damavand
Branch, Iran

frzz19866@gmail.com

Damavand, Iran

Javad Hosseinkhani

Department of Computer Engineering/ Islamic Azad University, damavand
Branch, Iran

jhkhani@gmail.com

Zahedan, Iran

Abstract

Clustering is one of the best ways to work with data provided. Clustering is one of the learning branches unsupervised and is an automated process in which samples are divided into categories whose members are similar to each other. A hierarchical clustering is a kind of traditional algorithm which follows the structure of society in social networks and on the basis of similarity or strength of the connection between each node; the entire network is divided into several subgroups. The advantage of community detection methods is based on the hierarchical clustering in which methods data is not necessary to be available already about the optimum number of communities and algorithms will be able to determine the optimum number of communities. The proposed method is a method based on hierarchical clustering which is able to identify overlapping communities. The main purpose of the community detection is to put similar nodes in a cluster with each other. Therefore, the nodes in a cluster share the same characteristics and the relationship between nodes within each cluster is denser from the connection of nodes between different clusters. The proposed method has been used for clustering the concept of the centrality of interstitial ridge. As well as to identify nodes that belong to two different societies, the criterion of the centrality of a node is used. Also the performance of the method proposed was compared with two methods previously proposed in terms of two criterions, Modularity and NMI, to detect communities. Simulation results showed the proper functioning of the proposed method.

Key Words

Web Mining, Social Networks, Clustering, Link Analysis and Accuracy

I. INTRODUCTION

A social network is a social structure made up of different groups of individual or organization. In the world of communications, social networking can be a useful context for producing and sharing ideas and an important factor in personal and social development [1]. The Internet has created a new kind of social connections that could create a new world parallel to the real world. The media have a wide audience that can actually provide them to consider a new concept of time and space. Thanks to these networks practically intervals are removed in time and space and transferring information, data or capital is done with incredible speed and wide level. Thanks to these virtual networks that create new concept against social networks, the possibility of simultaneous communication between people in different parts of the world is created in a new form [2]. Today, social networking has become a rich source of heterogeneous data. The analysis of this data can be used in interpreting social behavior of people and discover unknown information between users within the network. "The discovery of society" as one of the major challenges in the field of social network analysis, has been widely studied. Social networks, in addition to the graph structure, containing useful information within the network users, who could use this information to improve the quality of the discovery of the communities [3]. Although multiple methods for community detection algorithm is presented in social networks. However, these methods still have problems. The most important challenges in this regard are as follows:

The optimum number of clusters: optimum number of communities in many community recognition algorithms must be set by the user. On the other hand, in many applications the user may not know the exact number of communities. Thereby determining the appropriate number of clusters is considered one of the main challenges in this algorithm. Thus providing a method which is optimally able to determine the number of communities, is the challenges and objectives of providing a new community detection algorithm [4].

Initialize the community centers: The majority of the community detection algorithms act like this: first after determining the primary centers, then they try to update and improve these primary centers [5]. As a result, community detection algorithm depends on primary centers of any communities very much. Thus providing a method for initializing communities and the centers of each community should be targeted and it should be one of the paramount objectives of providing new methods for community detection [6].

Control data features: control the shape of community and final clusters is the number of data existing in each cluster, also the data density in each cluster is another problem in current community detection algorithm. Thereby providing a detection method to solve the existing problems of the major challenges in the field of data clustering is considered. In this study we have tried to provide a new community detection method based on hierarchical clustering algorithm to solve current problems and challenges [7].

II. RELATED WORKS

An issue-based community detection method has been proposed which combines both social issues and clustering analysis of the link. First, by using a clustering algorithm, all social issues

are divided into some groups of issue. Then network members in clusters of social issues in which involved, are divided. Due to differences in bond strength between nodes, link analysis is done in each local cluster to detect local communities.

A new hierarchical algorithm is provided to detect the community. The approach of this proposed method based on the density fell between each pair of father and son nodes in the dendrogram, on the basis that by decreasing higher density, the likelihood that children can form an independent society is more [3,4].

Therefore, based on the maximum flow of minimum cut, a new algorithm that can automatically find an optimal set of local communities is proposed [8]. An algorithm which is proposed based on simple publication to detect communities which require no prior know ridge. The main idea of offering different types of clusters through an appropriate hierarchical delicate clustering method [10]. A new method called Conclude is suggested. This clustering method has both the accuracy of global methods and local scalability approaches. This method creates a random walk without backtracking over the limit to calculate the importance of each ridges linked in a network, for example, calculates the center of the ridge [10]. Central ridge allows each node to be written on the points of a Euclidean space and calculates the distance between each pair of vertices; then, the distance used for clustering networks. In the proposed method for hierarchical clustering, the centrality of interstitial ridge and for overlapping communities, the centrality of nodes is calculated [11,12].

III. PROPOSED METHOD

The centrality criterion detects the role of a node. In the proposed algorithm for detection of the community, two parameters of the centrality of interstitial ridge and exchange messages (dynamically checked) are used. In general, the proposed approach will be as follows:

For all ridges existing in different communities, the centrality criterion is calculated:

A ridge with minimum criterion of the centrality ridge is detected and if the amount of this center ridge is less than a predefined threshold, that ridge would be known as the ridge separating of that community and the community nodes are divided into two communities.

For all community nodes, the criterion of centrality is calculated. Nodes, whose criterion of node centrality is more than a certain amount, are known as overlapping nodes which should be separated in this community and are attributed to both communities.

The proposed method for evaluating the performance of the two criteria NMI and Modularity is used.

Later in this section, first we address to connections related to calculating these two criteria. Normalized mutual information criterion is known by the acronym NMI. It is one of the known criteria for evaluating the performance of a community detection method in social networks.

This program was run 10 times and the results were averaged to be able to get the optimum performance.

The results based on performance evaluation and Modularity is the NMI.

TABLE1: REVENUE NMI CONCLUSION TO ASSESS

<i>NMI</i>	<i>data collection</i>
0.712.	karate
0.728	dolphin

As you can see in Table 1, performance evaluation output of NMI is shown. In fact, the runtime shows community detection time. Since pressing Run button of counter, runtime starts and then stops when all community detections are made.

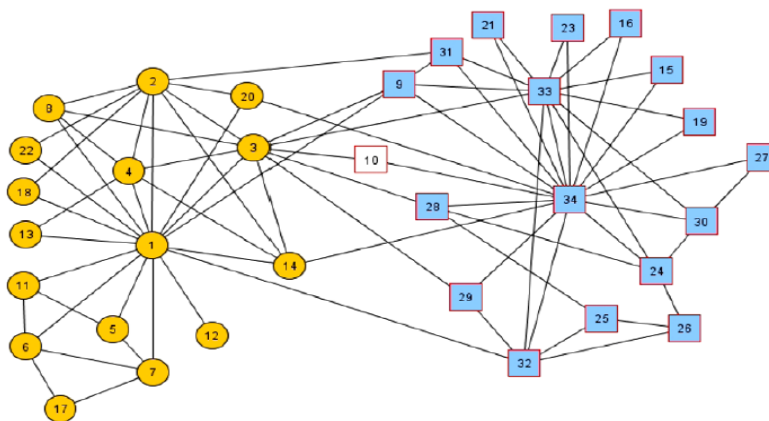


FIGURE 1: TWO OVERLAPPING COMMUNITIES DETECTED IN DATA COLLECTION KARATE

TABLE 2: MODULARITY PERFORMANCE EVALUATION OUTPUT

Modularity	<i>data collection</i>
0.689	karate
0.694	dolphin

As you can see in Table 2 modularity performance evaluation output is shown. In Figures 1 and 2 overlapping communities detected in data sets Karate and Dolphin are shown.

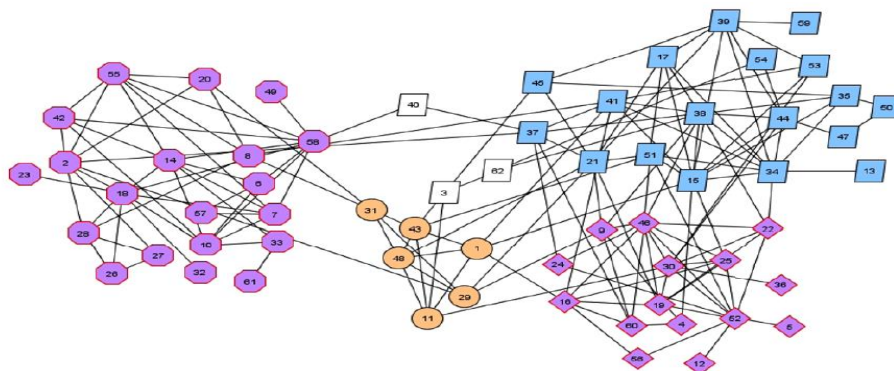


FIGURE 2: OVERLAPPING COMMUNITIES DETECTED IN DATA COLLECTION DOLPHIN

In this section, the performance of proposed method is compared with the previous methods in the fields of community detection. Table 3 shows comparing results of the proposed method with other methods based on the Modularity criteria.

TABLE 3: COMPARE THE PROPOSED METHOD AND PREVIOUS PERFORMANCE OF THE BENCHMARK MODULARITY

SUGGESTED METHOD	SWARM	CHEN	LOUVIAN	GN	NETWORK
0.689	0.619	0.656	0.627	0.612	KARATE
0.694	0.683	0.671	0.667	0.637	DOLPHIN

As this data shows that the proposed method has better performance than the methods were compared. In Figure 3 and 4, the performance of different detection methods was compared in terms of both Modularity and NMI criteria.

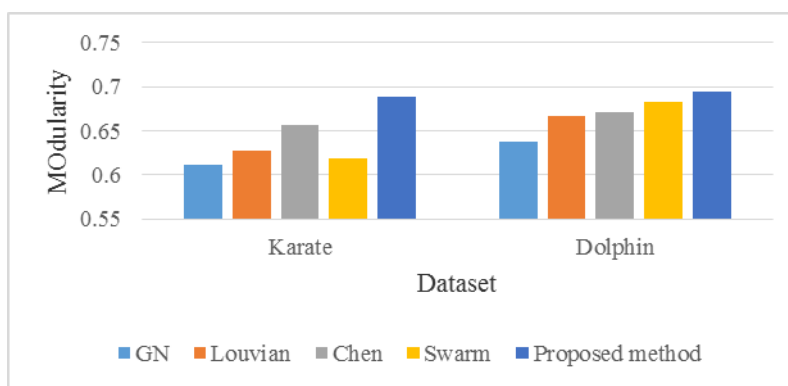


FIGURE 3: COMPARISON OF DIFFERENT METHODS OF CRITERIA MODULARITY

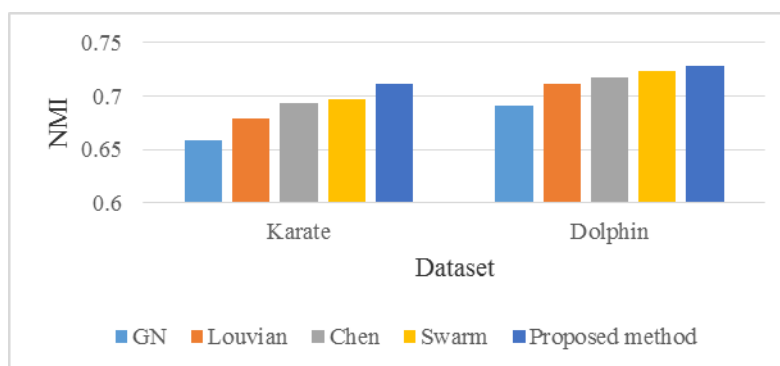


FIGURE 4: COMPARISON OF DIFFERENT METHODS OF NMI CRITERIA

According to the data tables 2 and 3 as well as graphical form of community detection, it can be concluded that the proposed method using NMI and modularity criteria, is able to detect the community with a high quality and has better advantages and performances compared to previous methods.

IV. CONCLUSION

Social network consist of the accumulation of a number of factors and their relationships. The goal of community detection is auto-discovery communities based on the structure or graph features. Optimization module, as one of the methods used in detecting communities, is widely used and has a quality function for evaluating the quality of communities found in the graph. Modularity offered, is one of the best known quality functions based on this idea, communities are a group of nodes in a graph that have dense connections. Therefore, the structure of society can be shown based on density under the graphs in the original graph. In this research, the combination of both concepts, centrality of the intermediate nodes and central interstitial ridge is used for the detection of hierarchical communities. In fact, two concepts of centrality node and centrality of ridge have been combined to detect overlapping communities. The performance of proposed method was evaluated and its accuracy was compared with other methods already provided. And in this study two dataset were used, two criteria NMI and Modularity were used to evaluate the performance and the results show that the proposed method has better performance than other methods. Advantages of proposed method can be stated as follows:

1. Due to the nature of hierarchical clustering, automated and optimized number of communities in the process of clustering is determined. So the methods in which communities can be manually selected, they have priority.
2. The use of criterion of the centrality of a node makes it possible to identify overlapping nodes and detect overlapping communities. In fact, by identifying the nodes that have higher centrality and reputation.

One of the limitations of the proposed method is that due to the use of centrality ridge criterion, it is possible to create limitation for large and high-density graphs. To solve this problem, it is suggested that this method be implemented in two phases for large social networks.

REFERENCES

- [1] Aggarwal, C.C., 2011. "An introduction to social network data analytics." Springer.
- [2] Hosseinkhani, J., Chuprat, S., and Taherdoost, H., (2012a). Criminal Network Mining by Web Structure and Content Mining. *11th WSEAS International Conference on Information Security and Privacy (ISP '12), Prague, Czech Republic September 24-26*.
- [3] Yoshida, K., 2014. "Memory Management for Big Data Mining – Cache Hit Rate Estimation of LessFU." *Procedia Technology*, 17: p. 114-121.
- [4] Freeman, L.C., 2014. "The development of social network analysis.: Empirical Press Vancouver."
- [5] Alhajj, R. and J. Rokne, *Graph-Theory*, 2014, "in *Encyclopedia of Social Network Analysis and Mining*", Springer New York. p. 661-661.
- [6] Newman, M., 2009, "Networks: an introduction. Oxford University Press."
- [7] Opsahl, T., F. Agneessens, and J. Skvoretz, 2010. "Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*," 32(3): p. 245-251.
- [8] Newman, M., 2008. "The mathematics of networks. *The new palgrave encyclopedia of economics*," 2: p. 1-12.
- [9] Al Falahi, K., N. Mavridis, and Y. Atif, 2012. "Social Networks and Recommender Systems: A World of Current and Future Synergies, in *Computational Social Networks*. "Springer. p. 445-465.
- [10] Dangalchev, C., 2006. "Residual closeness in networks. *Physica A: Statistical Mechanics and its Applications*," 365(2): p. 556-564.
- [11] Zhao, P. and C.-Q. Zhang, 2011. "A new clustering method and its application in social networks. *Pattern Recognition Letters*," 32(15): p. 2109-2118.
- [12] Jain, A.K., 2015. "Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*," 31(8): p. 651-666.