Community Detection Using Revised Medoid-Shift Based on KNN

Jiakang Li[1], Xiaokang Peng[1], Jie Hou[1], Wei Ke[2], and Yonggang Lu[1(✉)]

[1] School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, Gansu, China
ylu@lzu.edu.cn

[2] Gansu New Vispower Technology Co. Ltd, No. 1689 Yanbei Road, Lanzhou 730000, Gansu, China

Abstract. Community detection becomes an important problem with the booming of social networks. The Medoid-Shift algorithm preserves the benefits of Mean-Shift and can be applied to problems based on distance matrix, such as community detection. One drawback of the Medoid-Shift algorithm is that there may be no data points within the neighborhood region defined by a distance parameter. To deal with the problem, a new algorithm called Revised Medoid-Shift (RMS) is proposed. During the process of finding the next medoid, the RMS algorithm is based on a neighborhood defined by KNN, while the original Medoid-Shift is based on a neighborhood defined by a distance parameter. Since the neighborhood defined by KNN is more stable than the one defined by the distance parameter in terms of the number of data points within the neighborhood, the RMS algorithm may converge more smoothly. The RMS algorithm is tested on two kinds of datasets including community datasets with known ground truth partition and community datasets without ground truth partition respectively. The experiment results show that the proposed RMS algorithm generally produces better results than Medoid-Shift and some state-of-the-art together with most classic community detection algorithms on different kinds of community detection datasets.

Keywords: Clustering · Medoid-Shift · Community Detection · KNN

1 Introduction

Social networks have become ubiquitous in our day-to-day lives through such platforms as Facebook, Twitter, and Instagram. These networks can be modeled as graphs, with nodes representing individuals and edges representing their interconnections. Within these complex graphs, certain subgraphs exhibit particularly high density, where individuals are more closely interconnected than elsewhere. These subgraphs are commonly referred to as communities [1].

In recent years, a plethora of community detection algorithms have been proposed to mine hidden information within networks [2]. These algorithms are generally categorized into two types: overlapping and non-overlapping methods. To address the problem of community detection in network analysis, researchers have employed a variety

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
D.-S. Huang et al. (Eds.): ICIC 2023, LNAI 14089, pp. 345–353, 2023.
https://doi.org/10.1007/978-981-99-4752-2_29

of approaches, including hierarchical divisive, hierarchical agglomerative, and random walk-based methods [3], among others. To evaluate the performance of these algorithms, researchers have proposed various detection metrics. Among these, modularity [4] is a critical metric used to assess the quality of the community generated by different methods. A higher modularity value indicates better community creation [5]. Modularity measures the degree to which nodes within a community are more densely connected than nodes outside that community. Additionally, Normalized Mutual Information (NMI) [6] is a crucial evaluation metric when ground truth partitions exist for the dataset. The higher the NMI, the better the match with the ground truth partition.

Community detection poses a formidable challenge due to the complexity and scale of network structures. Notably, contemporary complex networks primarily comprise graphs, a form of structured data lacking coordinates that precludes the direct utilization of coordinate-based algorithms, such as the Mean-Shift algorithm, for community detection [7]. While the Medoid-Shift [8] algorithm proposed subsequently can address distance matrix-based issues, its application to community detection problems remains largely unexplored. Additionally, the Medoid-Shift algorithm may encounter a critical challenge of no data points existing within the neighborhood region defined by its distance parameter, leading to suboptimal performance on community detection problems.

Therefore, to address the challenges above, this paper has proposed a new community detection algorithm named RMS, which extracts the characteristics from both k-nearest neighbors (KNN) [9] and Medoid-Shift while focusing on detecting the non-overlapping community. In contrast to the traditional Medoid-Shift algorithm, our proposed method employs a modified approach for determining the neighborhood of a given node. Specifically, we have defined a parameter k for RMS borrowing the idea of KNN. The parameter k is used to define the medoid's neighborhood, rather than using a distance parameter as in the conventional Medoid-Shift algorithm. This modification effectively mitigates the issue of unstable number of data points within the defined neighborhood region, which is a known limitation of the original Medoid-Shift method. Moreover, during the shifting of the medoid, the RMS algorithm calculates similarities between each point p in the neighborhood of the current medoid and the KNN of p. Because the KNN of p is not related with the current medoid, the stableness of the shifting process is enhanced compared to the original Medoid-Shift method.

The remaining portions of this paper are organized as follows: Sect. 2 discusses related works in community detection. Section 3 discusses the process of the RMS algorithm and its data pre-processing in detail. In Sect. 4, we present the experimental results and analyze them in detail. Section 5 discusses the conclusion and future work.

2 Related Work

The research on uncovering the community structure of the real social network has been a hot research topic, which has also spawned many community detection algorithms. This section first introduces classical algorithms, KNN-based algorithms, and distance matrix-based algorithms. Then it introduces Medoid-Shift.

2.1 Classical Algorithms in Community Detection

In 2004, Newman and Girvan developed the Girvan Newman algorithm [10] which is a well-known method for discovering communities. The algorithm employs divisive hierarchical clustering and iteratively removes edges with the highest betweenness score to partition the network into subgroups. In 2010, Louvain [11] introduced a community detection algorithm that focuses on optimizing modularity. The algorithm aims to maximize the modularity of the entire network through an iterative process of optimizing the partition of the network into communities. Besides these two, there have been numerous community detection algorithms proposed during the past two decades, each with its characteristics and advantages. For instance, module-based optimization [12] algorithm, spectral clustering [13] algorithm, hierarchical clustering [11] algorithm, label propagation [14] algorithm, and information theory-based algorithm [15], which have become classical algorithms in community detection.

2.2 Medoid-Shift Algorithm

Derived from the idea of Mean-Shift, Medoid-Shift is very similar with Mean-Shift in the theory and process. They both calculate the shift toward regions of greater data density, find the cluster centers by iteration, and calculate the number of clusters automatically. The biggest difference is that Mean-Shift shifts to a location according to the Mean-Shift vector, while Medoid-Shift shifts to a certain point in the neighborhood. Furthermore, Medoid-Shift can be directly applied to distance-based community detection problems, but Mean-Shift can not.

Brief Introduction of Medoid-Shift Core Algorithm. Given an N × N symmetric matrix D(i, j) which is the distance between i and j starting from the point i, an index of point j is calculated as follows:

$$S(i,j) = \sum_{k=1}^{N} D(j,k)\phi(D(i,k)) \tag{1}$$

The next point to shift from i is point j with the minimum value in S(i, j), 1<=j<=n. By iteratively computing the next point from all the current points, tree traversal can be used to find the unique roots for all the points [8]. The number of clusters is the number of unique roots, and the label of each point can be derived directly from its corresponding unique root.

3 The Proposed RMS Algorithm

The RMS algorithm proposed in this study differs from the Medoid-Shift algorithm in several aspects. Firstly, the neighborhood in Medoid-Shift is defined by a distance parameter, whereas in RMS, it is defined by KNN. Secondly, in Medoid-Shift, distances between all the points in the current medoid's neighborhood are calculated when selecting the next point, while in RMS, similarities between each point p in the neighborhood of the current medoid and the KNN of p are computed for this purpose.

In the remainder of Sect. 3, the proposed RMS data preprocessing method is firstly introduced. Then we introduce the core algorithm of RMS and analyze the advantages of RMS compared to the Medoid-Shift algorithm.

3.1 The Definition of Graph in Social Network

Given a social network N, it can be denoted as a graph G(V, E), where $u \in$ V is a node, $e \in$ V is another node, $(u, e) \in$ E is an edge that measures some form of relationship (intimacy) between u and e. The number of edges and nodes are denoted by |E| and |V| respectively.

3.2 Data Preprocessing of the RMS Algorithm

In this study, a similarity matrix is utilized instead of a distance matrix to represent the weights of the graph. Specifically, the similarity matrix, denoted as SimM(i, j), captures the similarity between nodes i and j. In particular, when dealing with weighted graphs, the weight between two points is typically represented by their similarity value. However, for unweighted graphs, it is not appropriate to use the same weight scheme as in the case of weighted graphs. To address this issue, a novel method is proposed for computing the weights. The similarity between two points i and j in the unweighted graph is defined as:

$$SimM\,(i, j) = |N\,(i) \cap N\,(j)| \tag{2}$$

where N(i) represents the list of points connected to node I.

3.3 Core Process of the RMS Algorithm

The basic process of the RMS algorithm involves three main steps: Finding the KNN and similarity sum for each point, medoid clustering, and label assignment.

Step 1: Finding the KNN and Similarity Sum for Each Point. We define a property "Similarity Sum" represented as DL(i) for point i, which is the sum of the similarity between the point i and its k-nearest neighbors represented as KNN(i), where k is treated as a parameter for the RMS algorithm:

$$DL(i) = \sum_{p \in KNN\,(i)} SimM\,(i, p) \tag{3}$$

Before iterating, the "Similarity Sum" is calculated for each data point.

Step 2: Medoid Clustering. After the computation of the "Similarity Sum" and KNN, the proposed RMS algorithm initializes all points as the initial medoids. For each medoid i in SetA, the algorithm selects the next medoid from the $k + 1$ points which include the point i and its KNN. Specifically, the point with the largest "Similarity Sum" is chosen as the next medoid after point i, and is added to the new medoid set called SetB. This process continues for all the points in SetA. After each iteration, the algorithm compares

the previous medoid set, SetA, with the newly obtained medoid set, SetB. If the two sets are identical, the iteration stops; otherwise, SetB is assigned to SetA, SetB is cleared up, and a new iteration starts. Once the iteration stops, the points in SetA are returned as the cluster centers, and the next medoid of each point is stored in a list called the next medoid.

Step 3: Label Assignment. After getting the cluster centers and the next medoid of each point, the data points which converge into the same cluster center are assigned to the same cluster.

4 Experimental Results

Section 4 begins by describing the experimental setup, followed by an introduction of evaluation metrics and an overview of the comparative methods used in the study. The discussion is then divided into two parts: weighted graphs without ground truth, and unweighted graphs with ground truth. For each part, the section provides an introduction to the datasets used, the process of parameter tuning, the results of the comparative experiments, and a detailed discussion.

4.1 Overall Experiment Setup

Instead of utilizing the distance matrix, our approach employs the similarity matrix as the input for all algorithms. Notably, the datasets with known ground truth partitions consist solely of unweighted graphs, while the datasets without ground truth partitions consist solely of weighted graphs. To evaluate the results of the experiments on the datasets without ground truth partitions, modularity is used. While NMI is used for the datasets with ground truth partitions.

4.2 Evaluation Metrics

Similar to other community detection researches, we utilize the following metrics to assess the effectiveness of algorithms.

Normalized Mutual Information. The NMI value serves as an additional assessment parameter for community detection with ground truth partition. Generally, a higher NMI value is indicative of a closer partition to a real partition and is widely accepted.

Modularity Q. The accuracy of community detection in a complex network is conventionally evaluated by the modularity function Q, which is widely recognized as a standard.

4.3 Comparative Methods

This study employs both classical and state-of-the-art algorithms that focus on non-overlapping communities as comparative experiments on both weighted and unweighted

graph data. Compared to other methods, the RMS approach demonstrates its advantages by achieving higher modularity scores in weighted graphs and higher normalized mutual information values than most classical and some state-of-the-art methods. The algorithms used for comparative experimentation in this article are: Medoid-Shift [8], SCD [16], GEMSEC [17], EdMot [18], Girvan Newman [10]. To implement Medoid-Shift, we use $\Phi(D(i, j)) = \exp(-D(i, j)/2)$.

For all the algorithms including RMS, the parameter setting that gives the best performance is selected in the experiments.

4.4 Experimental Results for the Datasets Without Ground Truth Partition

Given the datasets have no ground truth partition, NMI is not suitable for evaluation. Thus modularity is adopted to assess algorithms in this section.

Dataset Description. The performance of the RMS algorithm is assessed on five real-world datasets, including Cell Phone Calls [19], Enron Email [20], Les Miserable network [21], and US airports [22]. We have not used datasets that are larger than these, because they are computationally hard for RMS to solve.

Implementation. We have implemented the proposed RMS algorithm and other comparative methods using Python. The comparative methods include Medoid-Shift, GEMSEC, Girvan Newman, SCD, and Edmot. Some of the algorithms' source codes are publicly available in Python packages such as karateclub [23] and networkx. All experiments are conducted on a Windows machine with 200GB of memory.

Experiment Results and Discussion. The results of modularity for the Medoid-Shift, RMS, Label Propagation algorithm, NNSED, SCD, and Girvan Newman algorithm are shown in the Table 1 below. Also, the result value is evaluated using modularity and the numerical value in the parenthesis denotes the number of clusters.

Table 1. The modularity and the number of clusters (in parenthesis) for datasets without ground truth partition.

Algorithm	Cell Phones	Enron Email	Lesmis	USA Airport
GEMSEC	0.4021 (36)	0.4682 (7)	0.5001 (6)	0.1774 (25)
SCD	0.3010 (170)	0.5339 (17)	0.4499(33)	0.0681 (259)
EdMot	0.6388 (20)	0.6356 (7)	0.5648 (8)	0.2847 (11)
Girvan Newman	0.5208 (15)	0.2365 (46)	0.4776 (12)	0.0169 (300)
Medoid-Shift	0.2365 (248)	0.2802 (90)	0.2973 (29)	0.1385 (281)
RMS	0.5379 (78)	0.5650 (14)	0.4271 (7)	0.1413 (40)

Based on Table 1, it is evident that the proposed RMS algorithm consistently outperforms most other methods, except for Edmot. For instance, in the Enron Email Dataset, the RMS algorithm improves modularity by 0.33, 0.03, and 0.1 compared to Girvan

Newman, SCD, and GEMSEC, respectively. This indicates that the RMS strategy is effective in datasets without ground truth partition. Furthermore, in comparison to the Medoid-Shift algorithm, the RMS algorithm has improved modularity by 0.3, 0.28, 0.13, and 0.01 on four different datasets. It indicates that the RMS algorithm improves modularity based on KNN instead of distance parameter, and in general, it is a successful modification from Medoid-Shift that has adapted to community detection effectively.

It is worth noting that the modularity scores for all methods are relatively low in the USA airports dataset. This can be attributed to the high sparsity of the dataset, which makes it difficult for graph-based methods to effectively explore the underlying graph structure.

4.5 Experimental Results for the Datasets with Ground Truth Partition

Given that the datasets have ground truth partition, NMI is the perfect evaluation metric under such circumstance compared to using modularity. Thus we adopt the NMI to assess all of the algorithm in this section.

Dataset Description. We have collected the unweighted datasets with ground truth partition, which are American Football Network [21], Dolphins Social Network [21] and American Kreb's Book [24].

Experiment Result and Discussion. The results are shown below in the Table 2. The result value is evaluated using NMI and the numerical value in the parenthesis denotes the number of clusters.

Table 2. The NMI and the number of clusters (in parenthesis) for datasets with ground truth.

Algorithm	Dolphins Social Network	American Football Network	American Kreb's book
GEMSEC	0.6288 (6)	0.6690 (115)	0.4260 (2)
SCD	0.4817 (25)	0.8664 (14)	0.3423 (28)
EdMot	0.8353 (4)	0.7553 (7)	0.3939 (3)
Girvan Newman	0.7560 (2)	0.6684 (114)	0.4353 (5)
Medoid-Shift	0.5857 (6)	0.7367 (57)	0.4258 (2)
RMS	0.7846 (3)	0.7768 (19)	0.4840 (2)

Table 2 shows that the proposed RMS method consistently outperforms most classical and state-of-the-art methods in the Dolphins Social Network and American Kreb's book datasets. In comparison to GEMSEC, SCD, and Girvan Newman, the RMS method achieves improved NMI scores of 0.16, 0.3, 0.01, and 0.03, respectively, on the Dolphins Social Network. This suggests that the RMS algorithm is effective even in datasets with a ground truth partition. Additionally, it is important to note that the RMS algorithm performs better than the Medoid-Shift algorithm in the Dolphins Social Network, American

Football Network, and American Kreb's book datasets, with improved NMI scores of 0.2, 0.04, and 0.06, respectively. These results demonstrate that the RMS algorithm is a significant improvement over the original Medoid-Shift algorithm.

5 Conclusion and Future Work

This paper introduces a novel community detection algorithm called RMS, which builds upon the Medoid-Shift algorithm and incorporates the concept of KNN to map the social network into a distance matrix. This approach addresses the limitations of using the Mean-Shift algorithm directly for community detection and achieves superior performance compared to other traditional algorithms. The study highlights several key insights: (1) The Medoid-Shift algorithm can be extended beyond mode-seeking problems to tackle community detection challenges; (2) The proposed RMS algorithm has the ability to automatically determine the optimal number of communities/clusters; (3) KNN provides a more effective way of defining neighborhood regions compared to using a radius parameter. As a part of our future work, we plan to:

- Incorporate kernel density estimation as part of distance matrix calculation.
- We will also explore more applications of the RMS clustering algorithm besides community detection.

Acknowledgments. This work is supported by Gansu Haizhi Characteristic Demonstration Project (No. GSHZTS 2022–2), and the Gansu Provincial Science and Technology Major Special Innovation Consortium Project (Project No. 21ZD3GA002), the name of the innovation consortium is Gansu Province Green and Smart Highway Transportation Innovation Consortium, and the project name is Gansu Province Green and Smart Highway Key Technology Research and Demonstration.

References

1. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)
2. Wang, C., Tang, W., Sun, B., Fang, J., Wang, Y.: Review on community detection algorithms in social networks. In: 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), pp. 551–555. IEEE, Nanjing (2015) doi: https://doi.org/10.1109/PIC.2015.7489908
3. Singh, D., Garg, R.: NI-Louvain: A novel algorithm to detect overlapping communities with influence analysis. J. King Saud Univ. Comput. Inform. Sci. **34**(9), 7765–7774 (2021)
4. Newman, M.E.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**(3), 036104 (2006)
5. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
6. Duncan, T.E.: On the calculation of mutual information. SIAM J. Appl. Math. **19**(1), 215–220 (1970)
7. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)

8. Sheikh, Y.A., Khan, E.A., Kanade, T.: Mode-seeking by medoidshifts. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE, Rio de Janeiro (2007) doi: https://doi.org/10.1109/ICCV.2007.4408978
9. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: KNN Model-Based Approach in Classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) OTM 2003. LNCS, vol. 2888, pp. 986–996. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39964-3_62
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
11. Blondel, V., Guillaume, J., Lambiotte, R., Mech, E.: Fast unfolding of communities in large networks. J. Stat. Mech **2008**, P10008 (2008)
12. Lee, J., Gross, S.P., Lee, J.: Modularity optimization by conformational space annealing. Phys. Rev. E **85**(5), 056702 (2012)
13. Shen, H.W., Cheng, X.Q.: Spectral methods for the detection of network community structure: a comparative analysis. J Stat Mech Theory Exp **10**, P10020 (2010)
14. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E **76**(3), 036106 (2003)
15. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci **105**(4), 1118–1123 (2008)
16. Prat-Pérez, A., Dominguez-Sal, D., Larriba-Pey, J.-L.: High quality, scalable and parallel community detection for large real graphs. In: Proceedings of the 23rd international conference on World wide web, pp. 225–236 (2014)
17. Rozemberczki, B., Davies, R., Sarkar, R., Sutton, C. A.: Gemsec: graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, pp. 65–72 (2019)
18. Li, P.-Z., Huang, L., Wang, C.-D., Lai, J.-H.: EdMot: an edge enhancement approach for motif-aware community detection. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 479–487. Anchorage, Alaska (2019)
19. VAST challenge (2008). http://www.cs.umd.edu/hcil/VASTchallenge08/. Last accessed 11 Apr 2023
20. Benson, A.R., Abebe, R., Schaub, M.T., Jadbabaie, A., Kleinberg, J.: Simplicial closure and higher-order link prediction. Proc. Natl. Acad. Sci. **115**(48), E11221–E11230 (2018)
21. The Network Data Repository with Interactive Graph Analytics and Visualization. AAAI. https://networkrepository.com (2015). Last accessed 8 Apr 2023
22. US airports dataset page. https://toreopsahl.com/datasets/#usairports. Last accessed 9 Apr 2023
23. Rozemberczki, B., Kiss, O., Sarkar, R.: Karate Club: an API oriented open-source python framework for unsupervised learning on graphs. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), pp. 3125–3132 (2020)
24. Political book dataset page. http://www.orgnet.com/divided.html. Last accessed 10 Apr 2023