An Improvement Approach for Reducing Dimensionality of Data with Matrix Decomposition in Data Mining

Sasan Jamshidzadeh[1], Javad Hosseinkhani[2]

[1,2]Department of Computer Engineering/ Islamic Azad University, Zahedan Branch, Iran

Summary
Many approaches have been developed for dimensionality reduction. These approaches can broadly be categorized into supervised and unsupervised methods. In case of supervised dimensionality reduction, for any input vector the target value is known, which can be a class label also. In a supervised approach, objective is to select a subset of features that has adequate accuracy to predict the target value. Since there is no correct answer in case of an unsupervised approach, the feature selection cannot be accurately stablished. Most of the unsupervised approaches are designed based on finding a subset of features and variables, in such a way that the part of the major structure and configuration of the dataset is preserved. This thesis studied Singular Value Decomposition (SVD) and its application in the unsupervised feature selection. Modification have been made on this method has been analyzed and evaluated. The result of experiment conducted on real data shows the good performance of the feature selection methods based on Singular Value Decomposition.

Key Words:
Dimensionality of Data, Data Mining, Matrix Decomposition dimensionality reduction

1. Introduction

Analysis of large complex networks, such as social network, World Wide Web, have drawn great interests in various research communities. This topic is important because those communities often play special roles in the network systems. Detecting the community structure in a complex network helps better understand the network system and thus has practical applications.

Many methods and algorithms have been developed for Community Detection (CD) [1], [2]. Most contemporary community detection algorithms choose a cost function, such as modularity Q [3] and ”cut” [4] function, which measures the quality of community partitions first, and then optimize this function through searching the solution space. These algorithms can be regarded as single-objective methods. That is, a single objective function is designed beforehand and the algorithm returns a single solution as results. Although these single-objective approaches achieve great successes in both artificial and real networks, they have some fundamental drawbacks. For example, they often cause a fundamental discrepancy that different algorithms may produce distinct solutions for the same network. Moreover, these approaches have the resolution resolution limit problem [6], that is, modularity optimization fails to find small communities in large networks.

In order to alleviate disadvantages in single-objective community detection algorithms, a natural approach may be to consider community detection as a multi objective optimization problem. Moreover, some evolutionary multi-objective community detection algorithms have been developed recently [6] [7]. These algorithms simultaneously optimize multi-objectives and return a set of optimal solutions. These multi-objective methods have preliminarily shown their advantages in detecting more accurate community structures. However, it is still an unsolved issue to make best use of these optimal solutions. These solutions are generated by the different trade-offs of the objectives, and they reveal different community structures from different perspectives. It is promising to detect more accurate and comprehensive structures through exploiting the tradeoffs among these optimal solutions. Communities are usually unknown in complex network and they are often unequal in size or density and so we can say that finding communities in complex network are small but important functions and these networks have the hierarchical structure [8]. Community recognition problem can be considered almost as an optimization problem which proposed by Gervan and Newman in 2002 and a large number of studies on evolutionary methods such as GA, SA and mutual evolutionary algorithms and also their solution have been made in which this happened to make its solvation be an interesting subject to other researchers [9]. Many issues in single-object optimization is based on community recognition and the difference is to be determined on the basis of objective functions [10]. On the other hand, considering community recognition problem as an optimization problem, an objective function is required so modularity Q can be used as an stopping Criterion in GN [11, 12]. Single-objective optimization algorithms are only one criterion optimization and optimization may be inappropriate when the criteria are likely failure from another point of view, many of them need to learn some information number of communities in which are often unknown to actual networks and this causes the

Manuscript received December 5, 2016
Manuscript revised December 20, 2016

community detection based on Single Object algorithms to have such these shortcomings [12, 13, 14].
To overcome these shortcomings in this paper we considered community detection problem as a multi objective optimization problem and introduced a new hybrid multi-objective algorithm based on harmony search algorithm and chaotic local search. The rest of this paper is organized as follows. Section II summarizes the existing multi-objective community detection algorithm. Section III shows the considered system model and the requirements of community detection algorithm are introduced .This scheme is introduced in section III. Section IV describes the results analysis. Finally, section V summarizes the paper and layout future research.

2. Related Works

During the past decade, to understand and utilize the information in complex networks, the research on analyzing community structure has drawn a great deal of attention, and various kinds of algorithms have been proposed. Methods for detecting and deriving communities from social networks by using communication graph structure of network is divided into four groups: communities based on node, communities based on group, communities based on network and hierarchical communities[15].
Community diagnostic criteria based on nods in this way each node in the group must have the properties such as reciprocity and availability and in order to achieve this, algorithms such as category, Clique, K-Clique, K-Cub, K-Clan and K-Plex are used. Quasi-Clique uses for community-based groups because links within the group instead of taking count as one group are fully considered. Dividing graph method in bracket communities based on network have been used in a way that the social graph divided into two sub-graphs so that the number of edges in each of the following graphs is minimized. Then each of the following diagrams divide into two sub-graph with minimum cost and we keep doing this operation as long as the number of sub-graphs reaches a certain amount. Algorithms such as Kernighan-Lin and clustering algorithm are used for segmentation. Algorithms such as hierarchical clustering algorithm or Gervan-Newman algorithms are used to recognize hierarchical societies.
Using Bayesian Generative Model [16], social networks and communities are divided based on topics and to detect communities on social networks, based on the theme, list and link contents, interest of the people obtain deductive and then using data mining methods, we are able to categorize people in different communities. The model assumes that people in a community debate on a subject which is interesting for every individual and a debate topic determines that society. Works in the field of text mining, includes PLSI [17], LDA [18] and the AT [19].
In [20] a framework based on studying user's activity in interactive websites like social networks, is presented. In [21] a sender-receiver-subject model is used to find discussed topics in social networks. In [22] a method based on OLAP-Style is applied to bracket graph regarding similarity of properties. In this method nodes in a community have common properties values. In all of these methods, communication between members is ignored. In fact to detect communities both structure of communities and topics in social networks are important.
Lancichinetti et al. [23] presented the Order Statistics Local Optimization Method, the first method capable to detect clusters in network accounting for edge directions, edge weights, overlapping communities, hierarchies, and community dynamics. It is based on the local optimization of a fitness function expressing the statistical significance of clusters with respect to random fluctuations, which is estimated with tools of Extreme and Order Statistics. OSLOM can be used alone or as a refinement procedure of partitions/covers delivered by other techniques.
Meo et al. [24] proposed a strategy to improve existing community detection algorithm by adding a pre-processing step where the edges are measured with respect to the center and w.r.t the network topology. In this way, the centrality of a cross-border reflects their contribution to making arbitrary graph, for example, spreading the message on the network, as short as it possible. This strategy is able to effectively complete information about the network topology and can be used as an additional tool to improve the community's recognition. Calculation of the edge by a few random walks of bounded length on the network would be done. Although the above algorithms and measures for Community Detection obtained a good performance in the type of networks they were target to, they were all designed for either separated or overlapping communities. In general, because of the limitation of representations, almost all available methods based on EAs were designed for detecting separated communities only.

3. Proposed Algorithm

This study discuss about a method in which proposes a way to extract community of social networks and in the other hand, analysis the relationship between users and focuses on contextual information of social networks . The method presented in this study consists of 5 modules, which include data social network dataset module, pre-processing and text data modeling module, social issue clustering module, division of social network users module and link analysis module. Figure I have shown

Proposed method's algorithm based on clustering and analysis of communication graph's structure.

```
Procedure BeeColonyCommunityDetection
Input:  M, nPop, MaxIt,BestSolution, ParetoFront
Output: CP; // the community partition

1. Begin
2. l ←0;// l is the iteration number
3. ∀v ∈ V, COM(l) ←— v ; // initialize the Community of all
vertices
4.    For g=1:N
5.       Compute FitnessFunc;
6.    End
7.  Do
8.    l ← l+1;
9.    For j = 1: nPop
10.     Assign a rank to each individual and sort;
11.     Create a new population of offspring;
12.     Combine the parents and offspring;
13.       For g=1:N
14.         Compute FitnessFunc;
15.       End
16.    End
17.   CP ← COM(l);
18. Until (l= MaxIt)
19. Q ← Compute the value of the modularity for CP;
20. CP ← COM;
```

Figure I. Proposed Method'S Algorithm

This module prepares and cleans data input from social activities like email. Derived data after preparation steps consists of:

- Communication matrix between individuals with 1 and 0 values.
- Dataset of the number emails sent and received by users.
- Dataset of connection between users and emails.
- Dataset of number of email, subject of email and body of email.

To derive unique key words these steps should be passed:

A. Numbers and special characters should be removed and all letters should be lower case

B. Stop words like and, or, the and … should be removed.

C. Removing words with less than α character and more than β characters as specified by table 1.

D. Obtaining root of derived words with Porter's root finder algorithm.

E. Obtaining repetition count of each word in all emails.

F. Discarding words with less than μ times repeat and more than £ times repeat.

G. Removing words with different forms and same meaning by using WordNet database.

Members participating in any matter submitted or received text and e-mail are defined as participants. Email client can be identified in the CC and BCC. Relationship between any subject and text of each cluster is known from output of the clustering algorithm and the relationship between each user and every text subject is defined by preparation module. Each local cluster is an undirected graph from its original graph which has been showed by a matrix. Rows and columns of the matrix indicated which users of each cluster are connected and the contents of each cell show the relationship degree between two users of each cluster.

4. Experiments

This section discusses details of used database, evaluate effects of applied method in quality of detected communities and show obtained communities from social networks by using proposed method of this article.

For each data set, the algorithm runs individually for 100 times, because the effectiveness of randomized algorithms is largely dependent on the generation of initial solutions. This is done to test its effectiveness and each time the initial solutions runs it has been selected randomly. To assess the quality of the proposed community detection method a normalized mutual information (NMI) [25] and agreements Modularity [26] is used.

Real World Networks:

The Zackary's Karate Club network was generated by Zachary, who studied the friendship of 34 members of a karate club over a period of two years [27].

The Bottlenose Dolphins network: A network of 62 bottlenose dolphins. The network split naturally into two large groups where the number of ties was 159[28].

The American College Football network: The network consists of 115 nodes and 616 edges grouped in 12 teams [3].

Political books written by V. Krebs said: nodes were indicated in 105 books on U.S politics that have been collected from Amazon.com and books on the subject almost identical pair, often have been bought by the same buyer. Books were divided by Newman [24] according to their political level (conservative or liberal), but except a few which do not show any clear affiliation [22].

The e-print Arxiv: It is a network of 9000 scientific paper and their citations (9000 nodes and 24000 links) [28].

Table I. Modularity result obtained by the two algorithms on Zackary's Karate Club data

Durability	Run Time	Modularity	No. community	Methods
0.8903	9.0029	1.9578	6	Girvan-Newman Method
0.8798	46.209	1.8071	8	Proposed Method

5. Conclusion & Future Works

This paper presents an Algorithm to Detect Communities in Social Networks using Multi-Objective Evolutionary Algorithm based on the recently developed algorithm that

was conceptualized using the musical process of searching for a perfect state of harmony, and the local chaos search algorithm. In this study, a genetic algorithm based on fuzzy clustering to optimize modularity is laid on the social network. In the proposed algorithm, the Basic modularity which set as basis work and using cut-off values which contains the highest density and lowest resolution between the weights of clusters on a network transactions, to explore the communities. Generating communities of high-quality, low computational complexity and desired levels of access to communities are two of the proposed method's benefits .The time complexity of the proposed algorithm in this area is O(L*N) that significantly reduced.

References

[1] G.Palla, I.Dereyi, I.Farkas and T.Vicsek, ”Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society,” Nature, 2015, 435(7043): 814-818.

[2] L.Danon, A.Diaaz-Guilera, J.Duch and A.Arenas, ”Comparing Community Structure Identification,” Journal of Statistical Mechanics: Theory and Experiments, 2005.

[3] M.E.J.Newman, M.Girvan, ”Finding and Evaluating Community Structure in Networks,” Physics Review, E 2004, 69:026113.

[4] A.Pothen, H. Sinmon, and K-P. Liou, ”Partitioning Sparse Matrices with Eigenvectors of Graphs,” SIAM J. Matrix Anal App., 1990, 11:430-452.

[5] S.Fortunato and M.Barthelemy, ”Resolution Limit in Community Detection,” Proceedings of the National Academy of Sciences, 2007,104(1):36-41.

[6] C. Shi, C. Zhong, Zhenyu Yan, et al., ”A Multi-Objective Optimization Approach for CommunityDetection,” CEC2010.

[7] C. Pizzuti, ”A Multi-objective Genetic Algorithm for Community Detection in Networks,” ICTAI09 379-386. http://wwwpersonal. umich.edu/ mejn/netdata.

[8] T. B. S. de Oliveira and L. Zhao, "Complex Network Community Detection Based on Swarm Aggregation," 2008, pp. 604-608.

[9] 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.

[10] A. Ferligoj and V. Batagelj, "Direct multicriteria clustering algorithms," Journal of Classification, vol. 9, pp. 43-61, 1992.

[11] M. Tasgin, A. Herdagdelen, and H. Bingol, "Community detection in complex networks using genetic algorithms," Arxiv preprint arXiv:0711.0491, 2007.

[12] J. Liu and T. Liu, "Detecting community structure in complex networks using simulated annealing with k-means algorithms," Physica A: Statistical Mechanics and its Applications, vol. 389, pp. 2300-2309, 2010.

[13] A. Gog, D. Dumitrescu, and B. Hirsbrunner, "Community detection in complex networks using collaborative evolutionary algorithms," Advances in Artificial Life, pp. 886-894, 2007.

[14] J. Liu, W. Zhong, H. A. Abbass, and D. G. Green, "Separated and overlapping community detection in complex networks using multiobjective Evolutionary Algorithms," 2010, pp. 1-7.

[15] Charu Aggarwal and Haixun Wang, " Managing and mining graph data," Springer, vol. vol. 40, 2010.

[16] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," in Proceedings of the 15th international conference on World Wide Web, 2006, pp. 173-182.

[17] T. Hofmann, "Probabilistic latent semantic indexing," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50-57.

[18] D. M. Blei, A .Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[19] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 306-315.

[20] J. Zeng, S. Zhang, and C. Wu, "A framework for WWW user activity analysis based on user interest," Knowledge-Based Systems, vol. 21, pp . 905-910 ,2008.

[21] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," Computer Science Department Faculty Publication Series, p. 3, 2005.

[22] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 567-580.

[23] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS one 6(4):e18961

[24] Meo PD, Ferrara E, Fiumara G, Provetti A (2013) Enhancing community detection using a network weighting strategy. Inf Sci 222:648–668

[25] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," Journal of Statistical Mechanics: Theory and Experiment, vol. 2005, p. P09008, 2005.

[26] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol. 69, p. 026113, 2004.

[27] W. W. Zachary, "An information flow model for conflict and fission in small groups," Journal of anthropological research, pp. 452-473, 1977.

[28] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "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.