EPiC Series in Computing
Volume 69, 2020, Pages 327–333
Proceedings of 35th International Conference on Computers and Their Applications

Utility-Aware Graph Dimensionality Reduction Approach

Lamyaa J Al Omairi, Jemal Abawajy, and Morshed U. Chowdhury

School of Information Technology, Deakin University, Geelong, Victoria, Australia.

{lalomair, jemal.abawajy, and morshed.chowdhury} @deakin.edu.au

Abstract

In recent years graphs with massive nodes and edges have become widely used in various application fields, for example, social networks, web mining, traffic on transport, and more. Several researchers have shown that reducing the dimensions is very important in analyzing extensive graph data. They applied a variety of dimensionality reduction strategies, including linear methods or nonlinear methods. However, it is still not clear to what extent the information is lost or preserved when these techniques are applied to reduce the dimensions of large networks. In this study, we measured the utility of graph dimensionality reduction, and we proved when using the very recently suggested method, which is HDR to reduce dimensional for graph, the utility loss will be small compared with popular linear techniques, such as PCA, LDA, FA, and MDS. We measured the utility based on three essential network metrics: Average Clustering Coefficient (ACC), Average Path Length (APL), and Average Betweenness (ABW). The results showed that HDR achieved a lower rate of utility loss compared to other dimensionality reduction methods. We performed our experiments on the three undirected and unweighted graph datasets.

1 Introduction

Recently, the number of social network users has been increasing rapidly; the advent of these technologies has led to the availability of large amounts of data with increasing features. The social network data comprise several objects (nodes) and connections (links). The nodes represent users, and the links represent relationships (e.g., friends, family, etc.), financial exchange, weblinks, etc. The social networks with a lot of nodes and numerous features have evolved to be of very high dimensions.

The studies of social networks of the last decade have demonstrated the urgent need to reduce the dimensions of networks. These include the objective and dealing with dimension reduction methods such as linear methods and nonlinear methods. The dimension reduction methods offer a useful strategy for addressing several issues, for example, the computational complexity of social networks analysis

[1], security in Social networks [2], and the visualization [3]. However, it is still not clear to what extent information is lost or preserved when these methods are applied to reduce the dimensions of large networks. Moreover, in our knowledge, there is no unique framework available to compare and evaluate these methods despite the fact that this particular related issue of network dimensionality reduction and information preservation/loss is of great importance.

In this study, we measured the utility of network dimensionality reduction and compared it with the utility of the original social network. In our research, we compared several different linear dimension reduction techniques such as Principal Component Analysis (PCA) method [4], Factor Analysis (FA) [5], Linear Discriminant Analysis (LDA) [6], and Multidimensional Scaling (MDS) [7] to evaluate the ability to preserve the utility of the original network. It showed that the rate of utility loss rose when applying the classical dimensionality reduction methods. Thus, the development of utility loss rate reduction is required for a complex of real networks after reducing their dimensions. Therefore, in this paper, in order to preserve utility for graph dimensionality reduction, we applied the HDR method [8] to reduce the dimensional, which minimizes the information loss of data. We measured the utility based on three essential network metrics: Average Clustering Coefficient (ACC), Average Path Length (APL), and Average Betweenness (ABW). The main contributions of this work are:

- Identification of the effectiveness of the standard utility measurement of dimensionality reduction in preserving the structural features of social networks.
- Application of HDR method to reduce data dimensions and produce data that maintain much utility for network dimensionality reduction strategy.
- Performance of the HDR method with the other existing methods such as PCA, LDA, FA, and MDS.

We used Email, Dolphin, and Poolbook datasets for our simulation experiments in MATLAB 2016b. Our simulation results show that the HDR method can achieve the least loss of utility in graph dimensionality reduction.

This paper is organized as follows: Section 2 presents related work; Section 3 will be the problem overview; Section 4 describes the experimental setup; Section 5 presents and discusses our experimental results, and finally, Section 6 presents the conclusions drawn from the study.

2 Related work

The problem of dimensionality reduction for social network data that keeps the utility of the reduced data compared with an original social network graph is still challenging. The dimensionality reduction of the original social network should have minimum utility loss as possible. Zenil et al. [9] have done some work in discovering the rate at which information can be lost when reducing the complexity of large networks. They proposed an approach to address the complexity of biological networks and evaluate network dimensionality reduction processes by applying information-theoretic measures to observe global and local patterns.

They also compared three different network dimension reduction techniques. Although the results showed that the approach preserved different amounts of information from the original objects, however, when deleting more than half the edges, it led to significant inconsistencies and loss of information. On another hand, in [1], Vaclav and Ajith used matrix factorization methods to reduce the dimension of social network data in addition to dimensionality reduction for networks. Their objective was to measure the amount of information lost during the reduction by using their method. The results showed that the processing of a larger amount of data was allowed by ignoring some of the valuable information. While in [10], the authors proposed a privacy-preserving framework of non-linear dimensionality reduction. They used the non-metric multidimensional scale as a perturbation tool to hide the original data values, then they compared the accuracy between the original data and perturbed data. They measured utility as a weighted sum of differences between distances in the input space and the corresponding distances in the output space, for the purpose of measuring the size of information loss between data points before and after the transformation. However, the drawback of this approach was that it added noise, which would distort the distances between data points. Therefore, poor results would be obtained. M.Al-Ghalibi et al. [3] had considered the huge dimensionality for graph about time series social network construction and visualization. They proposed a dimensionality reduction approach by applying a mathematical model to avoid biased feature selection. The authors in [2] considered that reducing the dimensions of complex social networks is one of the factors to achieve security in Social networks, where they highlighted several non-linear dimensionality reductions. We observed in these works [2,3] that their approach was based on features, but the authors ignored to measure lost information.

3 Problem overview

3.1 Preliminaries

A. Graph

A graph is represented as an ordered pair $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$, where $\boldsymbol{V}=\{1,\dots,n\}$ is a finite set of vertices and the set $\boldsymbol{E}=\{\boldsymbol{e}_1,\dots.,\boldsymbol{e}_m\}$ encompasses the edges. Each edge $\boldsymbol{e}_i$ is characterized as a pair of connected vertices $(\boldsymbol{u},\boldsymbol{v})$. A graph $\boldsymbol{G}$ may be directed or undirected [11],

$$\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})=\begin{cases}(\boldsymbol{u},\boldsymbol{v})\in\boldsymbol{E}\leftrightarrow(\boldsymbol{v},\boldsymbol{u})\in\boldsymbol{E},\forall\boldsymbol{u},\boldsymbol{v}\in V & undircted\\ Otherwise & directed\end{cases}$$

B. Dimensionality reduction

Dimension reduction task is to convert the data from high dimensional space onto a low-dimensional space [12] and can be classified into feature selection and feature extraction [13]. The analytical procedures facilitating this reduction are called “dimensionality reduction techniques.” A significant number of algorithms have been introduced for dimensionality reduction by dividing into linear and non-linear methods. Non-linear dimensionality reduction methods are commonly used for non-linear data that need to be reduced before being processed. Examples of non-linear methods are Locally Linear Embedding (LLE), ISOMAP, etc. [14]. Linear dimensionality reduction methods deal with data sets that have a linear relationship; for example, linear methods are PCA, FA, LDA, etc. [15]. The dimensionality reduction problem can be explained thus: consider the original data $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_m\}$ in high dimensional space $\boldsymbol{R}^n$. Then, find a matrix $\boldsymbol{A}$ which is the number of components of data. Matrix $\boldsymbol{A}$ converts the original data points into a new set of data points $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\dots,\boldsymbol{y}_m\}$ in a low-dimensional space $\boldsymbol{R}^m$ $(m\ll n)$, such that $\boldsymbol{y}_i$ “represents”$\boldsymbol{x}_i$,

where: $\boldsymbol{y}_i=\boldsymbol{A}^T\boldsymbol{x}_i,$ (1).

C. Utility measurement

The utility measurement of any graph is to measure how much the structural attributes of the original graph are maintained. In this work, the main objective is to minimize network information loss after network dimensionality reduction. We used some of the common metrics to quantify the level of information loss after reducing the graph data, we will explain them in section 4.

3.2 Problem Description

Nowadays, massive graphs are common in many applications, such as social networks, chemical compounds, and energy networks. It is obvious all the examples have complex network structures as well as high dimensional data. In order to process massive graph data effectively, the first critical challenge is to reduce the dimensional space of the original data properly. Reducing the dimensions is very important in the processing of huge graph data. However, unknown what extent the utility loss of the resulting low-dimensionality representation of the original high-dimensional graph data. The challenge discussed in this study is what impacts graph dimensionality reduction on the utility of the graphs and how the HDR model improves the utility of reduced data compared to the classical linear dimensionality reduction methods. HDR focused on two linear methods, namely, the PCA method which is a classical method of feature extraction that has been extensively utilized in the area of machine learning. Another technique is the Neighborhood Preserving Projections (NPE) method. This linear method is used to reduce dimensionality, where the HDR algorithm is designed based on a combination of these two methods/ algorithms. The detail explanation of the HDR algorithm can be seen in [8].

4 Experimental setup

We performed our experiments on the three graphs dataset. The graphs in the dataset are undirected and unweighted.

- PolBooks: A network of books about US politics sold by the online bookseller Amazon.com. Edges in the network represent the frequent purchasing of the book by the same buyers. The data was compiled by V. Krebs (www.orgnet.com).
- Email: This is the email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Nodes are users, and each edge represents that at least one email was sent (http://konect.uni-koblenz.de/networks/arenas-email).
- Dolphin: The undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand, as compiled by Lusseau et al. (2003), and made available by Mark Newman (http://www-personal.umich.edu/~mejn/netdata/).

Table 1 lists some of the graph dataset's structural properties and utility measure of those graphs. In the first step, we measured the utility based on three essential network metrics that are frequently used to quantify the amount of information loss, which are

- Average Path Length (APL): This property represents the average number of steps along the shortest paths for all possible pairs of network nodes. We can calculate average path length $\boldsymbol{APL}$ of a graph by using the following formula:

$$\boldsymbol{APL} = \frac{\sum_{i\neq j} d(\boldsymbol{v}_i, \boldsymbol{v}_j)}{\boldsymbol{N}(\boldsymbol{N}-1)} \dots\dots\dots\dots\dots\dots\dots (2)$$

- Average Clustering Coefficient (ACC): One property of a graph is the clustering coefficient, which represents the tendency of vertices in a network to cluster together. It calculates as follows: Let $\boldsymbol{v}_i, \boldsymbol{v}_j \in \boldsymbol{V}$ and $\boldsymbol{e} \in \boldsymbol{E}$, $\boldsymbol{v}_i, \boldsymbol{v}_j$ are neighbors of $\boldsymbol{v}$ then the clustering coefficient of $\boldsymbol{v}$ is

$$\boldsymbol{CC(v)} = \frac{\boldsymbol{N_e(v_i, v_j)}}{\boldsymbol{N(v_i, v_j)}} \ldots\ldots\ldots\ldots..(3)$$

Where $\boldsymbol{N_e(v_i, v_j)}$ is the number of pairs of neighbors connected by edges and the number of all pairs of neighbors of $\boldsymbol{v}$. To compute the average clustering coefficient (clustering coefficient for a graph $\boldsymbol{G}$), simple average $\boldsymbol{CC(v)}$ for all $\boldsymbol{v} \in \boldsymbol{V}$. Therefore, to find the average clustering coefficient $\boldsymbol{C}$: let $\boldsymbol{N} = |\boldsymbol{V}|$ be the number of nodes:

$$\boldsymbol{ACC} = \frac{\sum_{i=1}^{n} \boldsymbol{CC(I)}}{\boldsymbol{N}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots.(4)$$

- Average Betweenness (ABW): This property represents the importance of a particular vertex in terms of the number of times. The vertex is included in the shortest paths between vertex pairs in the network. The number of shortest paths between u and w that go through vertex v.

$$\boldsymbol{BW(v)} = \sum_{\boldsymbol{u,w \in V}} \frac{\boldsymbol{O_v(u,w)}}{\boldsymbol{O(u,w)}} \ldots\ldots\ldots.(5)$$

Where $\boldsymbol{O_v(u,w)}$ is the number of shortest paths between $\boldsymbol{u}$ and $\boldsymbol{w}$ that go through $\boldsymbol{v}$, and $\boldsymbol{O(v,w)}$ is the number of shortest paths between vertices $\boldsymbol{u}$ and $\boldsymbol{w}$. Thus, the average betweenness $\boldsymbol{ABW}$ will be

$$\boldsymbol{ABW} = \frac{\sum_{i=1}^{n} \boldsymbol{BW}(I)}{\boldsymbol{N}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(6)$$

Then we reduced the dimensions of the graph by applying the popular linear dimensionality reduction methods PCA, FA, LDA, MDS, and HDR on the original graph. We calculated the corresponding measurements of the dimensionality reduction. We made a comparison in terms of how the graph data utility has changed after the dimensionality reduction and how the HDR method achieved the least loss of graph utility.

Data	Nodes	Edges	ABW	ACC	APL
Email	1133	5452	0.061	0.25	3.606
Dolphin	62	161	0.0198	0.302	3.35
Pollbook	105	441	0.0404	0.47	3.76

Table 1 Structural properties of a dataset

5 Results and discussions

In this section, the results are presented. Table 2 summarizes our results. Each triple (dataset, metrics to measure the utility loss, and dimensionality reduction strategy) defines a cell that contains two numbers. The first one is utility measurement based on three network metrics (ACC), (APL) and (ABW) by the experiments that applied the dimensionality reduction strategies. The second one (presented between brackets) is the percentage of utility loss. To calculate the utility loss in percentage after dimensionality reduction, we used the following formula:

$$100 - \left(\frac{\alpha}{\beta} \times 100\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(7)$$

where: α is the value after dimensionality reduction and β is the original value, for example, the average path length measurement for Email dataset at the original = 3.606 and when applied LDA = 3.02, therefore, the percentage of utility loss is $[100 - (\frac{3.02}{3.606} \times 100) = 16\%]$.

Data	Utility	Dimensionality reduction methods				
		PCA	FA	LDA	MDS	HDR
Email	ACC.	0.19 [24%]	0.15 [40%]	0.2 [20%]	0.19 [24%]	0.22 [12%]
	APL.	3.02 [16%]	3.07 [14%]	3.101 [14]	3.02 [16%]	3.21 [10%]
	ABW	0.05 [18%]	0 [100%]	0.041 [32]	0.05 [18]	0.051[16%]
Dolphin	ACC	0.24 [20%]	0 [100%]	0.2 [33%]	0.24 [20%]	0.262 [13%]
	APL	2.79 [17%]	2.4 [29%]	1.3 [22%]	2.79 [17%]	3.01 [10%]
	ABW	0.0173 [12%]	0[100%]	0.0141[28%]	0.0173[12%]	0.0183 [7%]
Pollbook	ACC	0.37 [21%]	0 [100%]	0.35 [25%]	0.37 [21%]	0.399 [15%]
	APL	2.98 [19%]	2.8 [25%]	3 [18%]	2.98 [19%]	3.11 [17%]
	ABW	0.0335 [17%]	0 [100%]	0.031 [23%]	0.033[17%]	0.035 [13%]

Table 2 Summary of results obtained

From Table 2, we can see that the HDR outperformed compared to all other methods when measuring the utility by ACC, APL, and ABW metrics; this is because of the Rayleigh Quotient, which gives a particularly powerful algorithm that was applied in the HDR method. The results were insignificant when the FA method was applied, where we observe that the data lost all its information in some cases. Regarding the MDS strategy, the results were similar to PCA because of using the same Euclidean distances [16], when the percentage of loss was between 17% to 24%. While when applied the LDA method, we find the results were varying, such as in the Email dataset when we measured the utility by ACC metric, we got 20% as the percentage of utility loss, on the other hand, utility loss of Dolphin dataset was 33%.

6 Conclusion

Frequent use of the large graph nowadays leads to the critical need to simplify graph analysis. The dimensionality reduction is an essential factor to address this and other issues; however, it is possible that the graph will lose some information because of the reduction. In this paper, we conducted a comparative study to measure the utility of the graph (information loss) by applying some of the common linear dimensionality reduction techniques. Our experimental results showed that when the graph data are reduced by the HDR method, the graph preserves most of its information compared to MDS, PCA, and LDA methods

References

[1] V. Snášel, Z. Horák, J. Kocibova, A. Abraham, Reducing social network dimensions using matrix factorization methods, 2009 International Conference on Advances in Social Network Analysis and Mining, IEEE, 2009, pp. 348-351.
[2] L. Jain, S. Jain, D. NSIT, A New Approach to Supervise Security in Social Network through Quantum Cryptography and Non-Linear Dimension Reduction Techniques, IJCSI (2010).
[3] M. Al-Ghalibi, A. Al-Azzawi, Time series social network visualization based on dimension reduction, (2018).
[4] I. Jolliffe, Principal component analysis, Springer2011.
[5] D.N. Lawley, A.E. Maxwell, Factor analysis as a statistical method, Journal of the Royal Statistical Society. Series D (The Statistician) 12(3) (1962) 209-229.
[6] Q. Gu, Z. Li, J. Han, Linear discriminant dimensionality reduction, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 549-564.
[7] T.F. Cox, M.A. Cox, Multidimensional scaling, CRC press2000.
[8] L. Al-Omairi, J. Abawajy, M.U. Chowdhury, T. Al-Quraishi, High-Dimensionality Graph Data Reduction Based on a Proposed New Algorithm, Proceedings of 32nd International Conference on, 2019, pp. 1-10.
[9] H. Zenil, N.A. Kiani, J. Tegnér, Quantifying loss of information in network-based dimensionality reduction techniques, Journal of Complex Networks 4(3) (2015) 342-362.
[10] K. Alotaibi, V.J. Rayward-Smith, W. Wang, B. de la Iglesia, Non-linear dimensionality reduction for privacy-preserving data classification, 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, 2012, pp. 694-701.
[11] J.L. Gross, J. Yellen, Graph theory and its applications, CRC press2005.
[12] C.O.S. Sorzano, J. Vargas, A.P. Montano, A survey of dimensionality reduction techniques, arXiv preprint arXiv:1403.2877 (2014).
[13] C. Ding, X. He, K-means clustering via principal component analysis, Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 29.
[14] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, N. Koudas, Non-linear dimensionality reduction techniques for classification and visualization, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 645-651.
[15] J.P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: Survey, insights, and generalizations, Journal of Machine Learning Research 16 (2015) 2859-2900.
[16] A. Ghodsi, Dimensionality reduction a short tutorial, Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada 37 (2006) 38.