Treatment Trajectories Graph Compression Algorithm Based on Cliques

Svetozar Milykh1 and Sergey Kovalchuk
ITMO University, Saint Petersburg, Russia

Abstract. Learning treatment methods and disease progression is a significant part of medicine. Graph representation of data provides wide area for visualization and optimization of structure. Present work is dedicated to suggest method of data processing for increasing information interpretability. Graph compression algorithm based on maximum clique search is applied to data set with acute coronary syndrome treatment trajectories. Results of compression are studied using graph entropy measures.

Keywords. Graph compression, treatment trajectory, graph entropy

1. Introduction

Medicine is one of the most significant areas of technologies development nowadays. Several areas such as disease outbreaks modeling, pathogens evolutionary and exploring, illness spreading prediction often use computational methods to find better solutions.

Moreover, assistance systems are used to speed up and ease analysis of disease course, make information more complete according to already accumulated data. For example, in work [1] machine learning techniques are applied for type of hepatitis determination.

During the analysis of treatment trajectories we deal with heterogeneous objects. Graph structure is one of possible variants to work with such data set [2]. Objects can be distanced from each other with sequence of edges and similar values will be grouped together. The more data is considered the larger structure becomes. Consequently large structure turns more difficult for visualization and future analysis. Present work is intended to propose method of optimization graph structure in term of interpretability using graph compression technique.

2. Methods

2.1. Initial Graph Structure

The graph used in this paper can be classified as similarity graph. In order to construct such object metric function should be defined. Further, graph can be constructed with help of threshold value - value that is used to determine if edge between value is exists.
Combination of metric function, threshold value and graph structure will produce object where existing edge means that connected elements is similar enough to be considered.

This type of structure is often applied for NNS (nearest neighbour search) problem and used in areas such as template classification, morphology and computer vision, web graphs compression.

2.2. Compression Algorithm

The idea of suggested algorithm is based on finding cliques in graph and replacing them with representative vertex. Common schema of algorithm in shown in Figure 1.

![Figure 1. Common algorithm schema](image)

Initial set of edges is usually filterer with threshold value. Then process of finding and compressing cliques is starting. Maximum clique search problem can be solved with any existing method. In current work evolutionary approach was applied. Algorithm finishes when no cliques with appropriate size is found in structure.

3. Results

3.1. Experiment Data

Given data contains information about treatment trajectory of acute coronary syndrome. Each episode is described with sequence of letter each of that is corresponding to certain department. Departments that are used in initial data is following:

1. A: Admission department 1
2. D: Admission department 2
3. E: Cardiology department
4. I: Surgery
5. F: Reanimation
6. N: Surgery (coronary angiography)

For example, we can have elements like this: AFIFD, AEFID, AFED.
All 3500 items in dataset was initially divided into 7 clusters according to similarity of pathways.

Another important thing that need to be defined is similarity metric. According to initial data format, normalized Levenshtein distance was chosen as metric between vertices.

With all defined values we can construct graph of episodes, where every vertex is treatment trajectory and edges between elements contain information about their similarity.

![Graph of episodes example](image)

**Figure 2.** Example of graph of episodes with threshold value 0.5

### 3.2. Compression of Structure

During the compression operation several important details should be taken into account:

1. Mono and multi clustering compression
2. Threshold value effect

First experiment with initial data was performed in order to determine number of cliques that can be found with or without restriction on embedding elements from different clusters. Results are shown in Table 1.

| Threshold value | Mono cluster | Multi cluster |
|-----------------|--------------|---------------|
| 0.5             | 0            | 17            |
| 0.55            | 0            | 26            |
| 0.6             | 2            | 27            |
| 0.65            | 3            | 35            |
| 0.7             | 7            | 37            |
| 0.75            | 10           | 42            |
| 0.8             | 11           | 33            |
| 0.85            | 6            | 10            |
| 0.9             | 0            | 4             |

As expected, quantity of cliques with elements from same cluster is significantly lower that multi cluster variant. The main information that can be found from calculated values is following: varying threshold value can improve or worsen results of compression, because this has direct influence on number of found cliques.
Example of graph compression was illustrated using graph with 30 vertices and 75% of edges. Results are shown on Figure 3 and Figure 4. Figure 3(a) and 4(a) shows initial graph structure where color of vertex means cluster property, figure 3(b) and 4(b) shows compressed graph results.

![Figure 3. Graph with single cluster compression example](image)

![Figure 4. Graph with 4 clusters compression example](image)

The difference between single and multi cluster compression is based on clique structure. Embedding elements with different properties should be carefully handled or just prohibited as it was done in our work.

Another important moment is that during compression we do not break initial graph connectivity because of edge keeping strategy. Example of clique compression is shown on Figure 5. Group of vertices 3, 4, 5 was replaced with representative vertex 7 with all 3 edges kept.

![Figure 5. General clique compression example](image)

Overall compression ratio of algorithm also was measured for different threshold values. Results of this experiment are present in Table 2.
Table 2. Compression ratio depending on

| Threshold value | Number of compression | Compression ratio |
|-----------------|-----------------------|------------------|
| 0.5             | 20                    | 85               |
| 0.6             | 25                    | 76               |
| 0.7             | 11                    | 33               |
| 0.8             | 3                     | 8                |
| 0.9             | 0                     | 0                |

This table acknowledges that threshold value for initial graph construction can vary compression process.

4. Analysis

Interpretation of results is important part of work with graph. Often this structure is used for visualization of initial data and simplification the receipt of new conclusions during analysis. In paper [3] authors describe block structure simplification and use tests for people who is not involved in graphs during everyday work. This experiment showed that for 59% of people simplified structure is preferable to initial one.

In our work we suggest numerical method of interpretability check. Graph entropy measure shows the level of uncertainty in structure. It means decreasing entropy value during compression algorithm work will show growing level of certainty in structure. This parameter can be measured differently [4], but in this paper we will use two variant: parametric entropy (PE) [5] and network entropy (NE) [6].

Experiments was performed on initial data with different threshold value and calculated results are shown in Figure 6.

![Figure 6](image_url)

**Figure 6.** Entropy dynamics during graph compression

According to the pictures we can say that parametric entropy has more consistent decreasing tendency, but overall values are getting smaller and certainty of structure is growing.
5. Conclusion and Future Works

Convolution of treatment trajectories is worse learning because analysis of collected data about treatment methods and possible disease progression can simplify and make therapy more predictable and accurate.

Our work propose structure and compression algorithm description for graph based structure that can be optimized in term of interpretability. Shown results can be applied not only for medical data, but for every data set, where similarity or distance metric between elements can be calculated.

Question of interpretability also can be considered from different angles because automatic tests have both advantages and disadvantages comparing to human tests. Nevertheless, purpose of decreasing information under consideration is reached and can be used in other related optimization works.

Acknowledgement

This research is financially supported by The Russian Science Foundation, Agreement #19-11-00326.

References

[1] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning ma-chine for diagnosis of hepatitis disease. Applied Soft Computing Journal, Aug. 2013; 13(8): 3429–3438, doi: 10.1016/j.asoc.2013.03.008.
[2] Metsker O, et al. Modelling and Analysis of Complex Patient-Treatment Process Using Graph-Miner Toolbox. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNCS Jun. 2019; 11540: 674–680, doi: 10.1007/978-3-030-22750-0-65.
[3] Dwyer T, Riche NH, Marriott K, Mears C. Edge Compression Techniques for Visualization of Dense Directed Graphs. IEEE Transactions on Visualiz. 2013; 19(12): 2596–2605.
[4] Dehmer M, Mowshowitz A. A history of graph entropy measures. Information Sciences Jan. 2011; 181(1): 57–78, doi: 10.1016/j.ins.2010.08.041.
[5] Dehmer M. Information processing in complex networks: Graph entropy and information functionals. Applied Mathematics and Computation 15 July 2008; 201(1–2): 82-94. doi: 10.1016/j.amc.2007.12.010.
[6] v Solé R, Valverde S. Information Theory of Complex Networks: on evolution and architectural constraints. Lecture Nots in Physics. LNP 2004; 650:189-207. Available: http://citeseer.nj.nec.com/450707.html.