Relation Analysis Through Graph Data Reduction Technique

**ABSTRACT:**

This talk focuses on the application of a graph technique for reducing the number of attributes in large data sets. We use a data set containing > 20,000 variables (i.e. Exposome). The unprecedented number of variables can seem overwhelming, however, we propose a novel method to reduce the size of large data sets. We also propose in our method to keep variables that are strongly related together and related to the conceptual modeling that we consider as appropriate to finding insightful viewpoints. Our method utilizes a graphing technique to pull out interesting variable relationships through the use of maximal clique graph algorithms.

Utilizing a subset of the Exposome, 12,000 variables, our method reduced the set of variables by approximate 100-fold. In addition, 22 para-cliques were found with 219 variables thus eliminating 11,781 variables. The 219 variables were reduced further by eliminating those that were not strongly related to the dependent variable for obesity. We found the method does not have the disadvantage of eventually producing false positives although false negatives may occur. Another advantage turned out to be reduced background noise.

Our approach is novel in that no a priori assumptions about the population or data are made. In addition, our approach enables the display of the network relationships in the data thus allowing the data itself to highlight which factors are relevant. In contrast, current statistical modeling is effective at displaying the relationship between closing neighboring factors, but not multi-modal factors.

Keywords: graph technique, data Reduction, Big Data, clustering

# Introduction

With the growth of the techniques of the data storage and the ability of the data analysis. We can make analyzing and prediction on the large dataset. For instance, customers’ preference analysis from the purchase history; health risk prediction from genomic information, etc. However, there is a significant challenge when we deal with a large data set, which is only few features are important to the result and so many noise disturb our analysis. This research focus on using graphing technique to make data reduction so that can make analysis easily and clearly. The specific technique we used are tradition maximal clique algorithm and it’s variant named paraclique algorithm. We chose a subset of exposome, which is reported related to the CVD in literature to implement and test paraclique method. For purpose of comparison, we also compare to the machine learning technique of dimension reduction. Besides introduction, the organization of this paper as follow: section 2 background review some previous researches related to the data reduction topic; and make introduction about the exposome data set; Section III demonstrate our method in detail, Section IV set up two experiments on the CVD dataset and Infant Mortality dataset respectively. Section V shows the results of the experiment and comparison. Section VI is the conclusion of this research and future works we can expected.

# Background

Big data is the combination of huge volume, high velocity and different varieties of data streams

originated from heterogeneous and autonomous data sources. [1] A dataset, called Exposome, contains greater than 20,000 variables spanning the years 1990 to 2010. The data files of the Exposome are geotagged to the county level yielding an unusual level of detail for potential data analysis. The variables can be grouped into 5 broad categories: Social indicators (poverty, crime, racial segregation, and unemployment); Built factors (work, play, learn); Natural environment (air, climate, water, land and pollutants); Health factors (mortality, behaviors and disease specific indicators); Policy items (governmental laws, ordinances and regulations). For analyzing this kind of large dataset, we need to use some techniques to identify those important features, which can not only to help visualization, but also can reduce the computation burden. In other words, we need to make data reduction.

Data reduction methods for big data vary from pure dimension reduction techniques to compression-based data reduction methods and algorithms for preprocessing cluster-level data deduplication, redundancy elimination, and implementation of complex network concepts. Dimension reduction techniques are useful to handle the heterogeneity and massiveness of big data by reducing million variable data into manageable size [2]. These techniques usually work at after the data collection phases. Similarly, cluster deduplication and redundancy elimination algorithms that remove duplicated data for efficient data processing and useful knowledge discovery are primarily post-data collection methods [3]. Recently, the network theory concepts have also been used for big data reduction [4]. Paraclique is a graph based method, which has both clustering and data reduction properties. In this research, we adopt paraclique algorithm as the data reduction method.

The term paraclique was first defined in 2006, where it was shown to have advantages in the amelioration of noise inherent in high throughput biological data. [5] A paraclique is a maximum clique plus additional nodes outside that with strong relationship to the members in that clique. Paraclique relies on a maximum clique algorithm to find highly dense cores and then extends the cores to include nodes omitted due to simply lack of few edges connected. **Jeremy J Jay, John D Eblen et al** compared graph-based techniques and conventional methods. They used Jaccard similarity to measure each cluster’s agreement with every GO and KEGG annotation set, and the highest Jaccard score was assigned to the cluster. [6] They concluded that the best performing algorithms were graph based such as paraclique. However, most previous researches applied paraclique algorithm on biologic related datasets, which have somewhat network properties such as gene network, protein network, etc. In addition, they have preference on analysis of clustering performance and ignore the aspect of data reduction of the paraclique algorithm. For the purpose of avoiding clique overlapping, in the process of paraclique clique formation, it removed those nodes that are not in the paraclique, which can achieve the effect of data reduction.We will define paraclique and discuss how paracliques can be implemented in III. Method.

# Method

Our method utilizes a graphing technique to pull out interesting variable relationships through the maximal clique graph algorithm. The basic idea of this method is to identify the densely subgraphs, which mean a set of variables with strongly relationship. In briefly, this method contains 3 steps: 1) Transform the attribute correlation matrix to an undirected weighted graph; 2) Use a commonly known graph algorithm, maximum clique, identify dense or fully-connected subgraphs; 3) Reduce the noise caused by overlapped maximal cliques through paraclique algorithm.

## Constructing Graph

The concept of graph is from mathematic, which represents a collection of nodes and edges. Graph based techniques require us to transform the dataset to the form of a graph. We can treat the variables of the dataset as the nodes of the graph, then build the relationship between each pair of nodes thereby statistical methods. There many ways for computing similarity of two nodes. Pearson correlation, for example, can be used to find linear dependence between two nodes with quantitative measurements across common conditions. Through Pearson correlation method, we can form a correlation matrix, which computing each pairwise variables. Then, we can construct an undirected weighted graph. For using graph algorithms, we can transform the graph to an undirected unweighted by selecting a threshold. When a pair of nodes has Pearson coefficient at or above the threshold, an edge is placed between the two nodes. Otherwise, no edge is placed. Thus, we will get an undirected unweighted graph after filtered by threshold. In this study, we choose 0.5 and 0.6 as our threshold so that we can not only compare the outputs of different threshold, but also adjust the effect of clustering. Moreover, we utilize p-Value and set it equals 0.05 to reduce the false positive.

## Clique

For identify the nodes with strong relationship, we need to extract those dense subgraphs. A clique is a densest possible subgraph. Each pair of its vertices is connected by an edge. A clique is maximum if it is a largest clique in a graph. A clique is maximal if it is not contained wholly within a larger clique. Finding a maximum clique is a classic NP-Complete problem in computer science, which means it cannot be solved in polynomial time. Many variations of backtracking algorithms for maximal clique enumeration have been published in the literature. To the best of our knowledge, all can be traced back to the algorithms of Bron and Kerbosch. [7]

## Paraclique

Clique is the densest graph, since it contains all edges and its density equals 1. However, it cannot achieve an ideal effect, since one node may belongs to several cliques so that it connects them and still maintains a large size subgraph. Paraclique is a variant of clique and it relaxes the restriction of cliques, which must present all edges. For get a paraclique, we need to select one maximum clique as a core at first. Then we select one node outside this maximum clique and set a glom factor as the threshold, which means how many edges connected between the outside node and each node in the maximum clique. Using the glom factor, g, as a condition to determine if a node outside the core can be added to the paraclique set. We glom onto a non-clique vertex only if it is adjacent to at least g clique/paraclique members with strong correlations. The process of the paraclique formation as below:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1. Core (maximum clique) | 1. One connected | 1. Three connected |
|  | 1. Paracliqe formation |  |

# Experiments

We applied the method for data reduction and clustering analysis to the two sub-datasets of the exposome: one with 60 variables related to CVD and the other with 567 variables related to infant mortality. Through experiments on these two sub-datasets, we intend to examine the effect of our graph-based data reduction method, demonstrate the outputs with different parameters selection and compare difference of paraclique between high rate and low rate conunty.

## CVD:

* Dataset description

Cardiovascular disease (CVD) is the single largest cause of death in the United States of America (USA), accounting for close to one-third of all deaths. Significant research has been conducted regarding interventions to reduce CVD. There are 60 variables from the exposome were reported be relevant to CVD in the literature.

* Implementation

We calculated all pairwise Pearson correlations. Once we calculate all pairwise Pearson correlations, the result is represented by a weighted graph in which each node is the variable of the CVD dataset and the weight of each edge is the correlation. From the weighted graph, we construct an unweighted graph by retaining only those edges with correlation above some threshold t, in this study, we used 0.6 and 0.5 separately.

* Outputs

By using threshold 0.6, we extracted 2 paracliques: one contains 8 variables and the other contains 3 variables. In contrast, if we changed the threshold to 0.5, then we get one paraclique and ten variables.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1. 1 paracliques by threshold 0.5 | 1. 2 paracliques by threshold 0.6 | |

* Evaluation

Based on the outputs, we decide to use 10 variables from the paraclique as the features and use logistic regression to evaluate the classification effect of those features. Logistic regression, also known as logit model, is a type of statistical model that can be used to predict a categorical response, or outcome. A binary logistic regression is applied when the response variable is dichotomous, which means it has two categories or levels. It can assume either the value 1 or 0. In the logit model the logarithmic of odds of the outcome is molded as a linear combination of the predictor variables. The odds are the ratio of two probabilities, Whenever the logarithm of odds ratio is found to be positive, the probability of success is always more than 50%. When the logarithm of odds ratio is negative, the probability of success is less than 50%.

For the predictor variables, we take a subset of the dataset, including 10 selected independent variables from the paraclique and 1528 observations, which are those high rate and low rate counties by quartile. For the dependent variable, if the ratio of death and population is greater than 0.82%, then the outcome is defined as 1; if the ratio is less than 0.59%, then the outcome is defined as 0. This means 1 represents a high CVD area and 0 tends to be a low CVD place. We split the data into training set and testing set. The training set is used to fit our model which we will be testing over the testing set. By fit the logistic regression model, we get the result below:

|  |
| --- |
|  |
| 1. Figure Coefficient |

|  |
| --- |
|  |
| 1. Figure Deviance |

Now we want to see how the model is doing when predicting the outcome on a new set of data. Our decision boundary will be 0.5. If P(outcome=1|X) > 0.5 then outcome = 1 otherwise outcome=0. Take the test set as a new data, we can obtain the accuracy of the model is “0.918421052631579”, which shows the misclassification rate is only 8.2%.

Another typical performance measurements for a binary classifier are plotting the *ROC curve* and calculate the *AUC* (area under the curve). A good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5. Here is the ROC curve plot and ACU:

|  |
| --- |
|  |
| 1. ROC curve plot and ACU |

## Infant mortality:

* Dataset

This dataset includes 567 independent variables and 1 dependent variable named b\_infant\_mort\_99\_11 (black infant mortality).

* Implementation

We calculated all pairwise Pearson correlations on two groups, which are high infant mortality rate counties and low infant mortality rate counties respectively. The standard divided counties is by quntile Then we take 0.6 as the threshold, since we expect get more nodes with strong relationship and less edges with small weight. We use find maximal clique algorithm identified some maximal cliques, which the maximum cliques size was 20 in high group and 28 in low group. The paraclique algorithm was applied on these two maximum cliques and get two sets of paracliques.

* Comparison of high and low

Through applied the method on different groups of counties, we can see the changes from the original graph construction to the finally paracliques extraction.

|  |  |  |
| --- | --- | --- |
| 1. Table Comparison between high and low | | |
|  | High | Low |
| Nodes | 425 | 428 |
| Edges | 1594 | 2353 |
| Maximal cliques | 368 | 475 |
| Maximum | 20 | 28 |
| Minimum | 3 | 3 |
| #Paraclique | 45 | 41 |
| Paraclique size | 20 ~ 3 | 28 ~ 3 |

# Findings and Results

Utilizing a subset of the Exposome, 60 variables, our method reduced the set of variables by approximately 6-fold. One paraclique was found with 10 variables, thus eliminating 50 variables. In this test, the nodes in the paraclique were proven to have a relationship with CVD by literature, such as when Obesity rate was high, CVD was high; when Inactivity was high, CVD was high, etc. For the analysis on the infant mortality dataset, we can see that paralciques exacted from different groups have different number and size so that getting different variables after data reduction. This method makes sense for data reduction. Given a dataset with large number of attributes, we can extract those variables with strong relationship effectively in a graph form.

# Conclusions

The paraclique algorithm is a promising method for analyzing a large dataset to reduce the number of attributes. Our approach is novel in that no a priori assumptions about the population or data are made. This approach enables the display of the network relationships in the data, thus allowing the data itself to highlight which factors are relevant. This approach can be applied to analysis of different diseases, such as lung cancer etc. We will compare the effect of data reduction between different data samples, to further refine the technique as well as different relationship measurements.

**References**

[1]. Wu X et al (2014). Data mining with big data. IEEE Trans Knowl Data Eng 26(1): 97-107.

[2]. Weinstein M et al (2013). Analyzing big data with dynamic quantum clustering. arXiv preprint arXiv:1310.2700.

[3]. Fu Y, Jiang H, Xiao N (2012). A scalable inline cluster deduplication framework for big data protection. In: Middleware 2012. Springer, pp 354-373.

[4]. Trovati M, Asimakopoulou E, Bessis N (2014). An analytical tool to map big data to networks with reduced topologies. In: 2014 international conference on intelligent networking and collaborative systems (INCoS).

[5]. E. J. Chesler and M. A. Langston. Combinatorial genetic regulatory network analysis tools for high throughput transcriptomic data. In RECOMB Satellite Workshop on Systems Biology and Regulatory Genomics, 2005.

[6]. Jay J, Eblen J, Zhang Y, Benson M, Perkins A, Saxton A, Voy B, Chesler E, Langston M: A systematic comparison of genome-scale clustering algorithms. BMC Bioinformatics 2012, 13(Suppl 10): S7.

[7]. Bron C, Kerbosch J: Algorithm 457: finding all cliques of an undirected graph. Proceedings of the ACM 1973, 16(9):575-577.