Graph Theory for Dimensionality Reduction: A Case Study to Prognosticate Parkinson's

CSE-0410 Summer 2021

Taskir Rahman Tasin and Yeasmin Akter
Department of Computer Science and Engineering
State University of Bangladesh (SUB)
Dhaka, Bangladesh
taskir.rahman72@gmail.com and yasminepsha@gmail.com

I. Summery

The study made its inference on the optimum number of features which might help to get the actual variables influence on the dependent variable like Parkinson's Disease (PD). It's always been tough to consider multiple regressors as principle and their level of significant correlation to the predicted variable due to the huge dimensionality of the small observations. The study explain its significance by ensuring some methods like Disjoint Set Union (DSU) and dimensionality reduction to show the accurate and precise number of features which are highly explained the predictive variables. For so, the study undertake computational graph theory which tries to define the construction of a formal mathematical model of real system by the proposition of pruning technique with others. Along with the research highly focused on dimensionality reduction for reducing the number of random variables under consideration, by reducing high dimensional space to the low dimensional space. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Generalized Discriminant Analysis (GDA) are the participative methods of Dimensionality reduction. But here the research narrow down its focus on PCA and LDA for approaching the variables to the linear correlation. The PCA and LDA helps to minimize the variance and maximize the distance between control and predicted variables which are mentioned in the study. Ultimately, the study proposes a means to construct network graphs from features using strong linear correlations, promoting the use of Spearman's r in a current world exalting Pearson's r. Notwithstanding, it's also observe that, dimensionality reduction with PCA does not reveal what features it projected onto another dimension, but DSU can reveal what features have been unified under a root till a timestamp. The elucidation has shown to improve performance through statistical inference on completely randomized, evenly sized k-fold (k = 10) datasets, for each decrement of r, but, multiple iterations are needed to tune r optimally for the most representative set.

II. Proposed Methodology

Parkinson's Disease (PD) is a neurodegenerative old-age complication that affects the motor muscular system. It may cause a subject's speech to become monotonous, mumbled, hoarse and conversations often drift away from the topic—endorsing speech as a potential detector of the disease. The progressive nature of the disease merits a longitudinal study, sampling an individual after certain intervals. 2015 witnessed 6.2 million people falling victim to PD, resulting in 117,400 deaths worldwide. The dataset we analyze for the proposed identification of Parkinson's is enriched with features defining 10 perspectives they are baseline , MFCC, intensity , wavelet , formant frequencies , TQWT, bandwidth , longitudinal study ID , vocal , gender. The study has a longitudinal nature and hence there exists an ID for the ease of tracking, along with different medical metrics.

A. Preparation of Adjacency Matrix

This paper bundles up multiple features applying DSU on a network graph produced on the basis of Spearman's rank correlation coefficient (r Spearman).We know,

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

Fig. 1. r Spearman

The concept of graph theory has edges linking nodes. Similarly, in this work, we create a network graph that we subsequently represent using a 2D adjacency matrix.
A legit question may arise, as to why the usage of Spearman's (in place of Pearson's) r for defining the edges. It is due to certain advantages of the approach over Pearson's: robustness to extreme observations, ease of calculation and that the two variables can be ranked separately. We know r Person is:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Fig. 2. r Person

However, the concept of correlation is bidirectional in nature and outputs a value within -1 to +1 (inclusive).
Firstly demolish boundaries 10 features. Then set a tolerable threshold for spearman's r and calculate Spearman's r for each features against all other features. After calculate consider absolute values of r Spearman. Then prepare a unidirectional network graph basis Pearson's r. Prepare an upper triangular matrix of Spearman's r. The output count of connections.

B. Performing DSU on Features

We assume a connection to be there between any two features if they are highly correlated. The threshold of the said high correlation is alterable and its tuning leads to different counts of features in the final feature-sets. From Preparation of Adjacency Matrix the connections induct features and initialize features to their own roots. Then input the edges from the adjacency matrix. From disjoint sets until the list of edges is exhausted. After exhausted extract out the final roots. And lastly finalize the set of roots as a constructed feature set.

C. Fitting Models, Validating Results

Representative features (roots) from all disjoint sets are found by running DSU until the list of r-based connections (edges) is exhausted. With this output, we manually transfer the feature-sets to Tensor flow for predictive analysis. The higher the tolerance for r Spearman, the more the features. We first set a benchmark accuracy and incrementally improve on this by tuning the r-threshold 5 times. For results free of bias, specifications are kept the same for all cross validation. The features have been scaled via normalization for a smooth convergence of Gradient Descent (GD).

III. Advantage

1. The paper precisely done on linear data correlation and firmly defines the principle variables influences on Parkinson disease.
2. Helps to remove data redundancy and compress the dataset precisely.
3. In this paper, using so many graph and table so, the reader read easily and understood properly.

IV. Disadvantage

1. The research didn't consider Generalized Discriminant Analysis (GDA) which certainly reject some random variables might influence predicted variable; Parkinson Disease.
2. The PCA LDA solely focused on linear correlation among the control and predicted variables mentioned in the paper are quite undesirable to presume the optimum number of control variables.
3. Higher possibility of data loss, mean and covariance might not enough to define dataset precisely.
4. PCA shows a 'black box' tendency by not clarifying which features have contributed the most/least to the final projections.

V. Every terminology of the paper used

Disjoint set union, dimensionality reduction, Spearman's r, Pearson's r, Parkinson's disease, statistical inference.

VI. Why this paper is unique

This paper has introduced a novel feature-exclusion method based on the application of DSU, an algorithm belonging to computational graph theory. The research proposes a means to construct network graphs from features using strong linear correlations, promoting the use of Spearman's r in a current world exalting Pearson's r. In order to optimize computational costs regarding r calculation and DSU, the study seeks to make the network graphs unidirectional and reduces the corresponding matrix to an upper triangular form. The solution has shown to improve performance through statistical inference on completely randomized, evenly sized k-fold (k = 10) datasets, for each decrement r. Dimensionality reduction with PCA does not reveal what features it projected onto another dimension, but DSU can reveal what features have been unified under a root till a timestamp.

VII. Experimental Result section explanation

The documented research attempts to justify the effectiveness of a novel application of graph theory for a reduction in dimensionality. The course of this discussion starts out by demonstrating success in optimally training, eventually making its way to show the significant positive impact on incrementally increasing accuracies—using statistical inference. The discourse is concluded by comparative studies involving other dimensionality-reducing technologies and related literature. Before putting forth the argument bolstering the supremacy of our method, it is imperative to set a fair field for experimentations with all alterations. For all alterations, we cross-validate the results k-fold where k = 10, a widely-accepted numeric for validating medical diagnosis. Qualitatively, PCA shows a 'black box' tendency by not clarifying which features have contributed the most/least to the final projections; whereas DSU has the capability to reveal features that have flocked up.

VIII. Future work of the paper

The future research may find out some non-linear correlations and undertake GDA methods to show the random relationships among the considering number of predicted and control variables.

Acknowledgment

I would like to thank my honourable**Khan Md. Hasib Sir** for his time, generosity and critical insights into this project.

References

[1] Maitra, S., Hossain, T., Hasib, K. M., Shishir, F. S. (2020, November). Graph Theory for Dimensionality Reduction: A Case Study to Prognosticate Parkinson's. In 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0134-0140). IEEE.

[2] Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.