Prepend: A Novel, Deterministic CUR Matrix Decomposition

Lane Lewis, Kevin Lin, Alexa Aucoin

May 2022

1 Introduction

A common issue when dealing with large datasets in fields such as Genetics is finding ways to run standard statistical analysis on the entire dataset while being limited computationally by the algorithms for analysis. As an example, if a researcher at a biotech company has a dataset of 10,000 genes from 5000 subjects and they wish to find a subset of genes that predict a variety of diseases, they may wish to run a clustering algorithm on the dataset, compute the correlation between the different genes, and compare the distribution of genes across subjects. Altogether, the time and computing resources needed to perform this analysis may exceed the budget that the company allows. There are several ways the researcher may get around this issue. One solution for the researcher is for them to reduce the size of their dataset by selecting out only a subset of the columns or rows to use when performing their analysis. The researcher now has two choices, to select features in a supervised or unsupervised manner.

To use a supervised feature selection algorithm means that they selects columns and rows based on some measure of performance each column/row gives: such as the amount of correlation each gene has with odds of getting Alzheimers disease or the mutual information it shares with it. This has the advantage of giving a subset that contains information on the target predictor, but this also can be a disadvantage. Choosing rows and columns of the dataset that contain information on one type of question doesn't necessarily mean that those columns/rows contain information on any question that could be asked of the dataset. However, the researcher may want to perform a variety of analyses on the data with different predictor targets such as odds of developing Alzheimers disease, diabetes, or heart failure. In this case, a supervised algorithm for column/row selection may not be the best option.

In an unsupervised feature selection algorithm, the columns and rows are selected not based on any external predictor targets, but based on some internal measure of information each column/row has within the dataset. So, a researcher may want to use this type of algorithm if they want to reduce the total size of the dataset while still preserving the maximal amount of information about the dataset in general. One common algorithm used to perform unsupervised feature selection is Principal Component Analysis (PCA). PCA constructs new columns composed of linear combinations of the original columns of the dataset, in a way such that each new column is uncorrelated with the others. A new column created by this algorithm then would be something like: (PC1 = .2*Gene1 - .01*Gene2 +.13*Gene3+ ...). From a dataset of 10000 genes, the researcher may be able to explain the majority of the data by using only 20 or so 'new columns' called Principal Components.

One of the issues with using PCA is that the Principal Component columns created by the algorithm may be not easy to directly interpret by the researcher. Using the above example, if a Principal Component were found to significantly predict Alzheimers disease, it is unclear what direct action should be taken. Since PC1 = (.2*Gene1 - .01*Gene2 +.13*Gene3+...), should the company invest in developing a drug to de-express Gene1, Gene3 or others? The answer isn't as straightforward as it would be if the PCs were themselves genes.

CUR matrix decompositions are unsupervised feature selection algorithms similar to PCA, but instead of creating a new dataset of Principal Components, they chooses to keep a set of existing columns and rows out of the dataset. The way in which the CUR decomposition algorithm does this, is it finds the columns and rows of a matrix that (in some way) contain most of the information of the Principal Components. It then takes the subset of found columns and places them into a matrix C and finds a subset of rows and places them into the matrix R. A matrix U is also constructed such that the composition of $CUR \approx$ Original Dataset. Through use of the C and R matrices, the genetics researcher would have a way of representing nearly all the information of the dataset, while still having just a subset of the original genes and rows of the genes. In this way, a CUR matrix decomposition accomplishes a very similar representation as PCA but remains much more interpretable. If a researcher finds that a column from the C matrix correlates with risk of developing Alzheimer's then this simply says that the gene corresponding to that column is related to Alzheimer's risk.

CUR matrix decompositions aren't unique, as in there are many algorithms that can generate useful C, U, and R matricies. The current most popular algorithm for performing a CUR decomposition is very fast but non-deterministic (i.e. if run multiple times it will produce different choices of columns in C and rows in R). In some instances where the decomposition needs to be replicable - such as if the genetics researcher needed to make consistent choices of genes to target - this type of algorithm might not be as useful as a deterministic one. In this paper, a deterministic CUR matrix decomposition was developed that will produce the same columns in C and rows in R when run multiple times.

2 Using This Paper

Outline
- Find dataset (genetics probably)
- Show Graphs