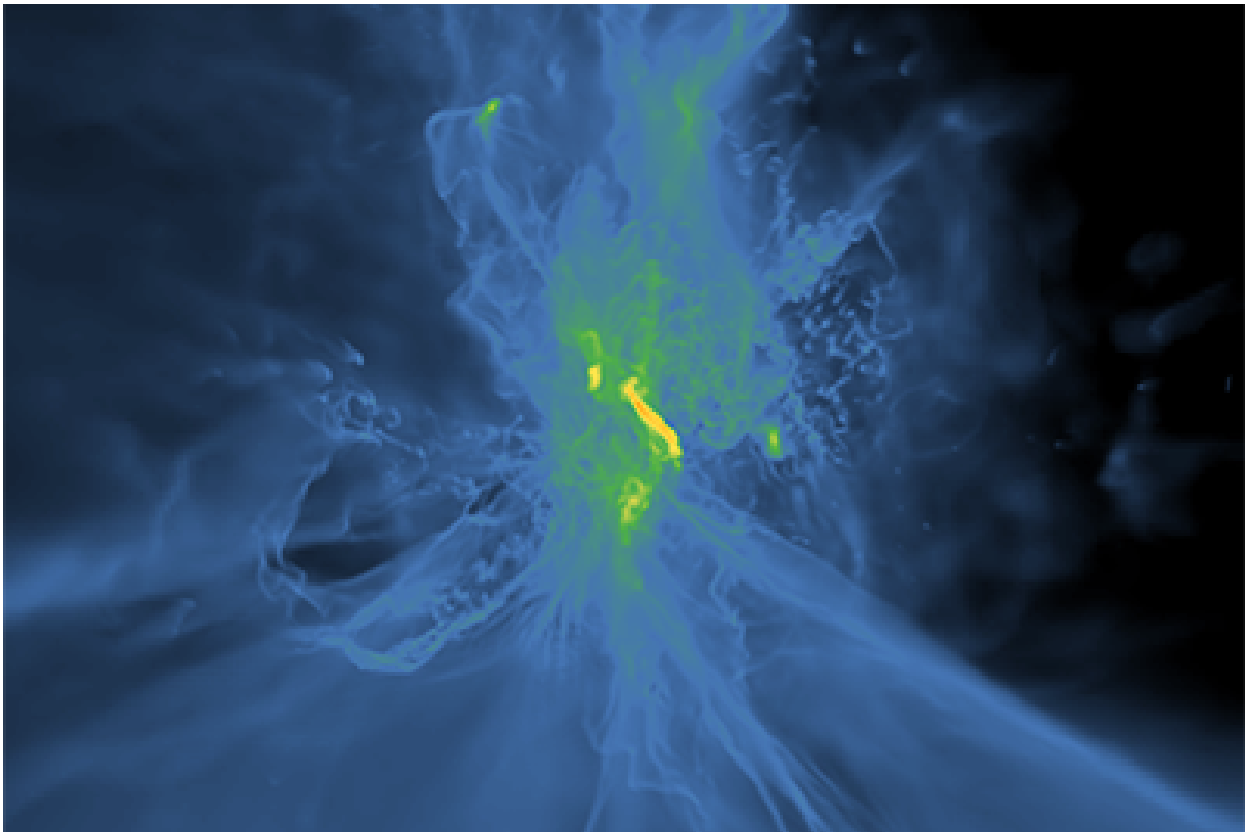

Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

2021

M. Sweety Reddy
Undergraduate Student/ Data Science Enthusiast (UPES Dehradun)

Overview

A very crucial component of galaxies, the circumgalactic medium (CGM) has assumed an important role in our understanding of galaxy evolution owing to rapid advances in observational access to this diffuse, nearly invisible material. The CGM is a source for a galaxy's star-forming fuel, the venue for galactic feedback and recycling, and perhaps the key regulator of the galactic gas supply. Observations from all redshifts and from across the electromagnetic spectrum indicate that CGM gas has a key role in galaxy evolution.

Goals

1. Review our evolving knowledge of the CGM with emphasis on its mass, dynamical state, and coevolution with galaxies.
2. Reduce the dimensional results so that we can maintain a high level of accuracy compared to the original quasar absorption spectra datasets.

Specifications

There are several techniques for dimensionality reduction such as :

- Missing Values Ratio, Low Variance Filter, High Correlation Filter.

- Random Forests/Ensemble Trees. Principal Component Analysis (PCA). Principal component analysis (PCA), Backward Feature Elimination, Forward Feature Construction.

There are other recent advanced Techniques for Data Dimensionality Reduction. Three newly added techniques being: linear discriminant analysis (LDA), neural autoencoder, and t-distributed stochastic neighbor embedding (t-SNE). So basically, there are many dimensionality reduction algorithms to choose from and no single best algorithm for all cases.

In terms of overall accuracy and reduction rate the random, forest-based technique might be the most effective in removing uninteresting columns and retaining most of the information for the classification task at hand.

Even a High Correlation Filter is an optimal choice as data columns with very similar trends are also likely to carry very similar information, and only one of them will suffice for classification. Here we can calculate the Pearson product-moment correlation coefficient between numeric columns and the Pearson's chi-square value between nominal columns.

Milestones

I. Review the dataset and analyze the best dimensionality reduction technique suitable for it.

After reviewing the dataset, we would be able to choose one among the researched dimensionality reduction techniques on the quasar absorption spectra dataset.

II. Compare and choose the best model according to the accuracy achieved.

After applying the suitable techniques to our dataset, our main goal would be to achieve the highest accuracy rate by modeling it accordingly.