Galactic explorations with deep learning and clustering

Michael Young, Fangfei Lan
University of Utah, Data Mining (CS 6140)

God I love space!!!

problem + data

- There's far more astronomical data than we can handle with our measly human brains alone: we need computers to help us do things like classify galaxies and stars and other astronomical objects.
- A technique of growing importance in this area is *surprise* deep learning, i.e. using convolutional neural nets (CNNs) to perform classification tasks.
- The inner workings of CNNs are not always super interpretable.
- This project seeks to explore the features learned in a CNN trained on galaxy images through dimensionality reduction and clustering.
- The aforementioned galaxy images are 60,000+ in number and labeled by GalaxyZoo participants.

key ideas

- ML Interpretability
- Dimensionality Reduction
- Clustering and Clustering Comparisons

what we did

1. Trained a CNN

- Used transfer learning with resnet34 trained on ImageNet. 30,000 images were used (80/20 test/val split)

2. Extracted feature vectors + dimensionality reduction to view

- Fed 20,000 galaxies to our trained network and extracted feature vectors for each galaxy (feature vector = output of the second to last layer of the network).
- Used tSNE, PCA, and LDA to reduce the dimensions and visualize.

t-sne -galaxies

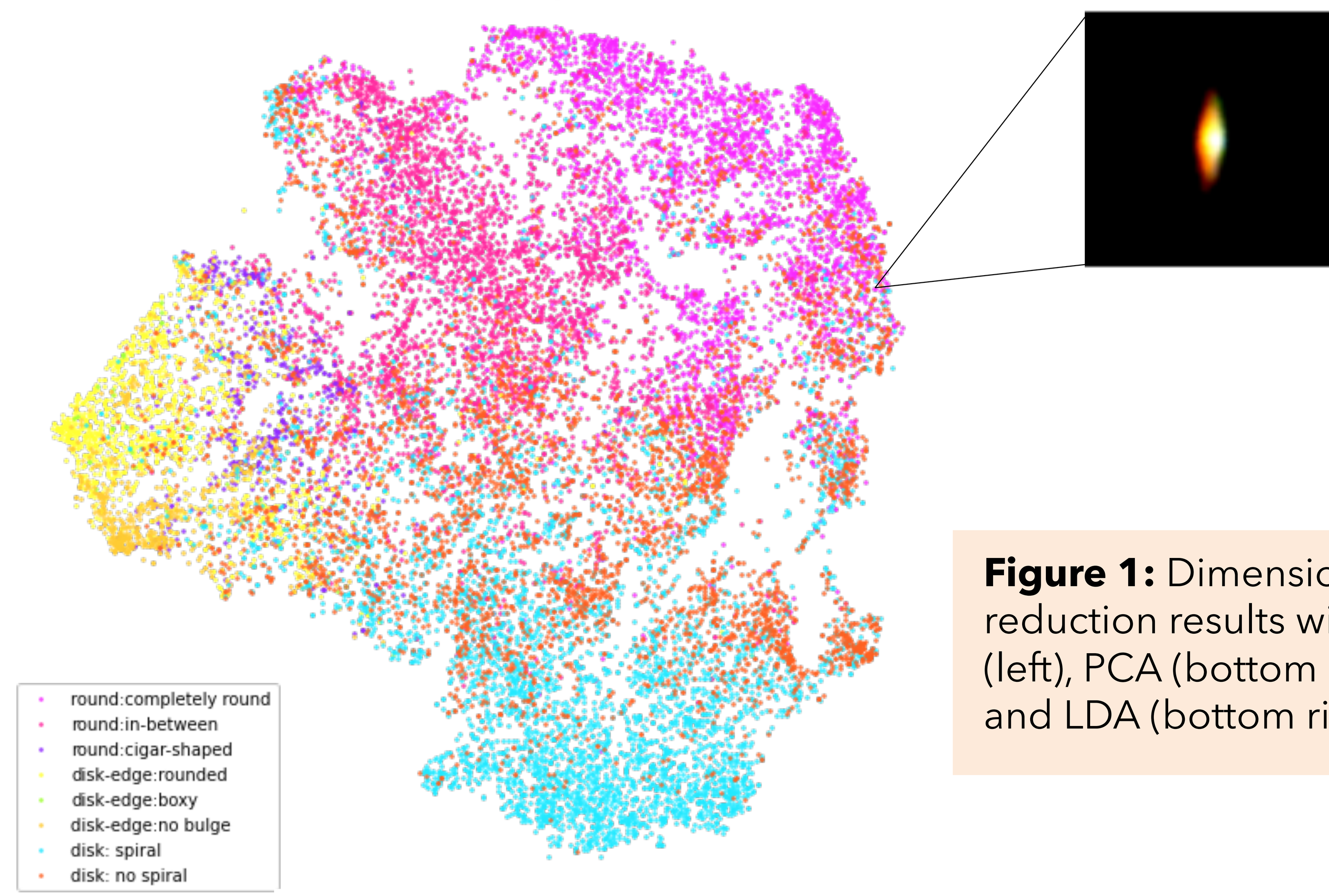
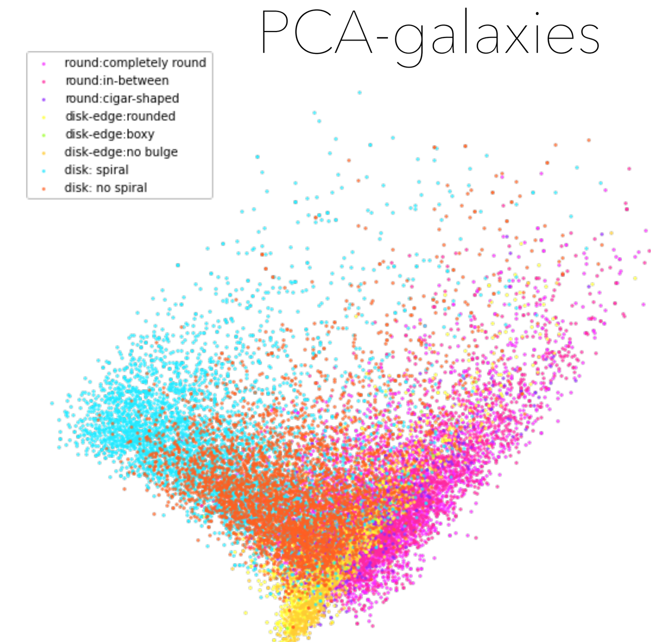


Figure 1: Dimensionality reduction results with tsne (left), PCA (bottom left), and LDA (bottom right).

PCA-galaxies



LDA-galaxies

3. Clustered data and compared clusterings

- Performed K-Means clustering with 3 to 8 clusters and DBSCAN on the data after dimensionality reduction.

DBSCAN w/ t-sne

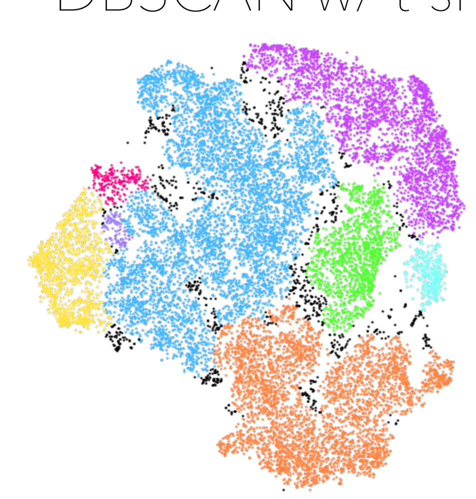
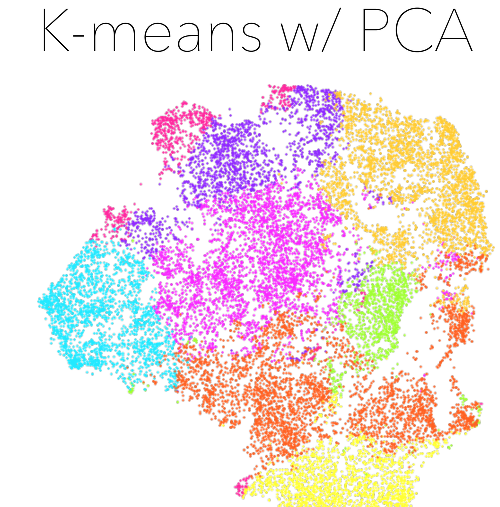


Figure 2: DBSCAN partitioned the data into 9 true clusters and 1 noise cluster. It resulted in the most separable clustering. However, it does not work for higher dimensional data.

Figure 3: K-Means with 8 clusters after PCA generated a clustering similar to that of LDA but different from DBSCAN.

K-means w/ PCA



K-means w/ LDA

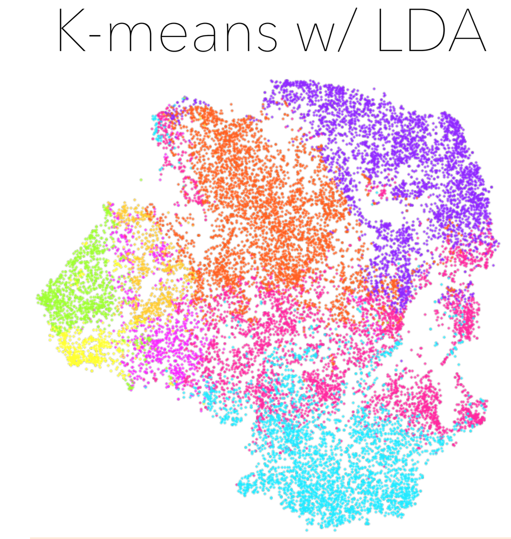
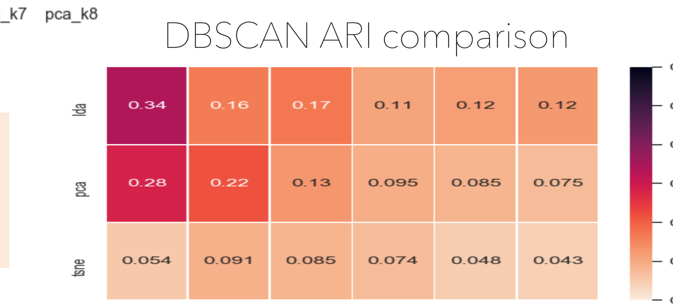


Figure 4: Performing K-Means with 8 clusters after LDA resulted in the most sophisticated clustering of the three. It is also the most similar one to the true distribution shown in Figure 1.

- Compared the clustering assignments with **Adjusted Rand Index (ARI)** and **Adjusted Mutual Information (AMI)** and both measures reached the same conclusion.
- Matrix visualization confirmed that **PCA + K-Means** and **LDA + K-Means** generated the most similar clustering assignments.
- T-SNE + DBSCAN clustering is dissimilar from all other clustering results.

Both ARI and AMI showed that K-Means generate similar clusterings for LDA and PCA with higher numbers of clusters (6, 7, and 8).



T-SNE + DBSCAN is different from all other combinations of techniques.

what we learned

- Features learned from CNN were useful for drawing morphological distinctions between galaxies.
- Disk with no spiral galaxies have the least separability from the rest of the galaxy types.
- DBSCAN does not work well for high dimensional data.
- PCA + K-Means and LDA + K-Means had similar results.
- ARI and AMI computed slightly different individual similarity scores but showed the same overall trend.

references

- D. Harvey et al. Galaxy Zoo – The Galaxy Challenge. Retrieved February 11, 2020 from https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/data

acknowledgements

Acknowledging:
- The Galaxy Challenge from the Kaggle Competition for the carefully curated data
- Professor Jeff Phillips for the guidance and inspiration in research ideas and directions