Investigating the Impact of Unsupervised Feature-Extraction from Multi-Wavelength Image Data for Photometric Classification of Stars, Galaxies and QSOs

Domain: Astronomy, classification

Data source: Sloan Digital Sky Survey (SDSS)

Challenge: Feature generation

Approach: Deep Learning, Computer Vision, Unsupervised Training

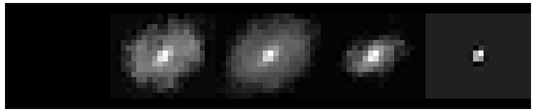
Background

Class labels

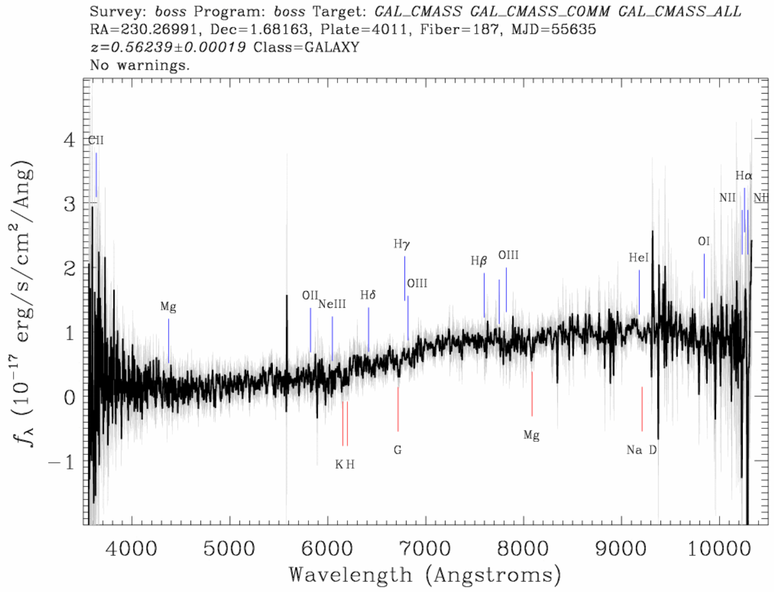
- STAR / GALAXY / QSO

Photometric vs spectroscopic

- spatial vs function of wavelength
- abundant vs precise



Multi-wavelength image data of SDSS J152104.77+014053.8



Optical spectra of SDSS J152104.77+014053.8

Photometric Classification Gaps
(STAR / GALAXY / QSO)

- **Feature generation** – handcrafted rules lead to higher risk of bias
- **Image data** – potentially useful, but not well-explored
- **Interpretability of the model** – important for astronomers, but often lacking in complex classifiers

Deep Belief Networks (DBN) – a solution?

- **Feature generation** – unsupervised method lessens the risk of bias
- **Image data** – well-explored as a Computer Vision algorithm
- **Interpretability of the model** – generative model with layer-wise training

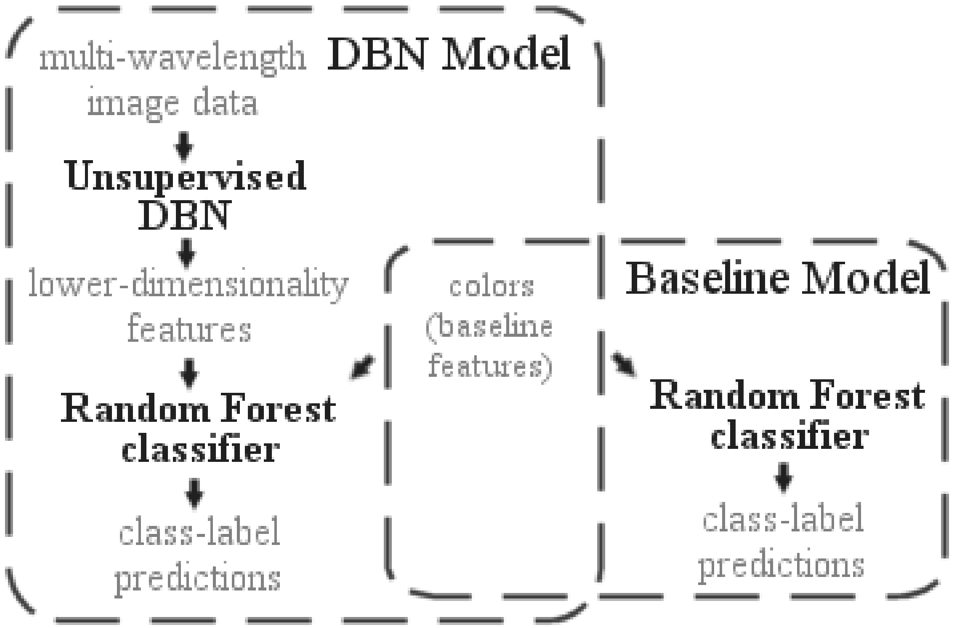
Research Question

"Can unsupervised feature-extraction from multi-wavelength astronomical image data provide a significant advantage in the photometric classification of stars, galaxies and QSOs?"

Experiment setup

Baseline Model
- Traditional parameters

DBN Model
- Traditional parameters
- DBN-features



Model overview.

Sampling

Full population (3 537 411 objects)

- All objects with a spectroscopic classification

Clean population (3 140 92 objects)

- No duplicates
- Valid measurements
- Minimum quality criteria
  (ONLY based on SDSS parameters)

Sample (10 000 objects)

- Random sampling
- Stratified on class labels

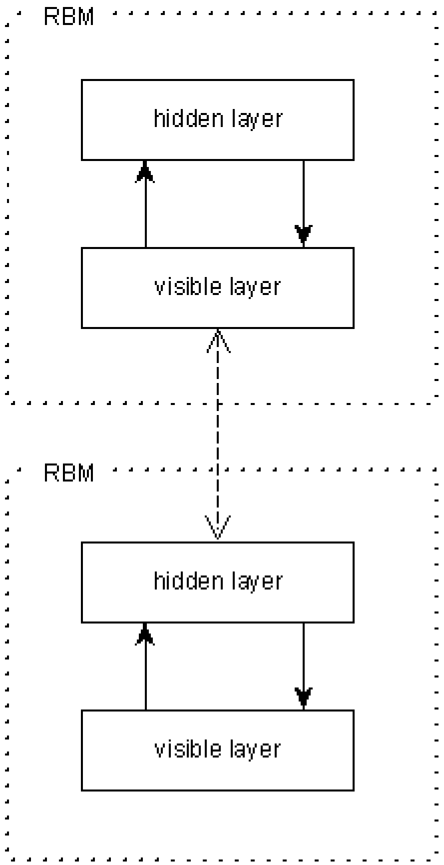
Class label	Sample n = 10,000	Clean population n = 3,140,923	Full population n = 3,537,411
STAR	23.97%	23.97 %	22.83 %
GALAXY	62.10%	62.10 %	64.29 %
QSO	13.93%	13.93 %	12.88 %

Class label proportions in the sample and populations.

Deep Belief Network

DBN

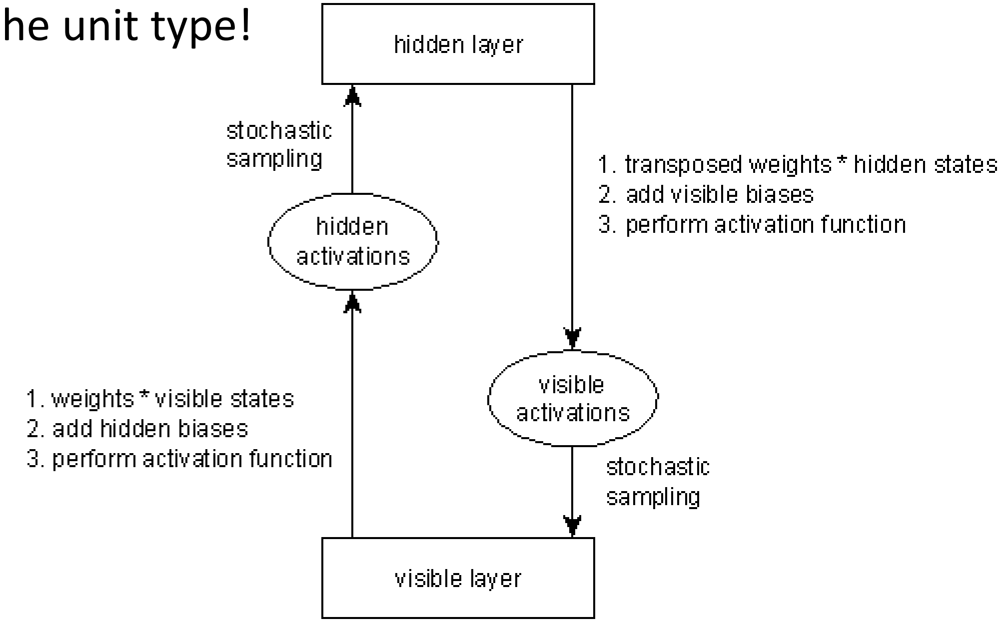
- Stack of RBMs
- Layer-wise training



2 RBMs forming a DBN.

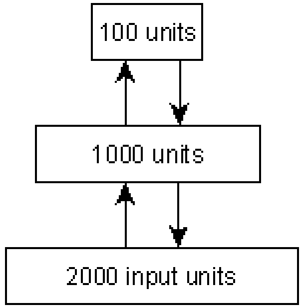
Restricted Boltzmann Machine

- Symmetrical weights
- Encoding / generative
- Activation and sampling depends on the unit type!



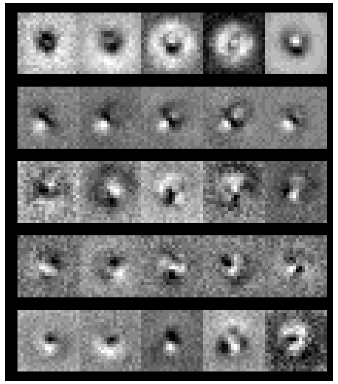
A single RBM.

What did the model learn?

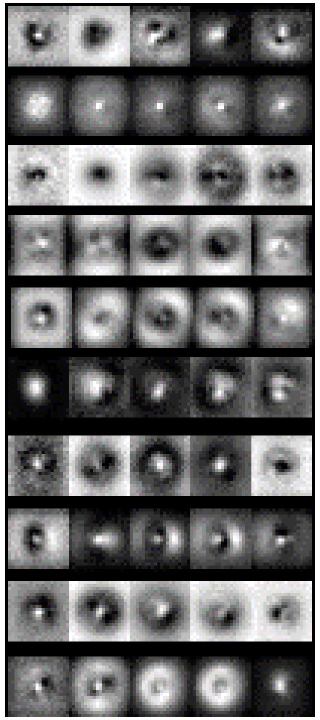


Thesis model.

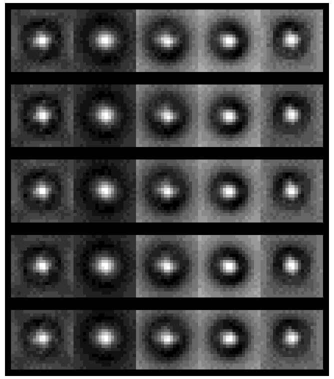
First-layer features.



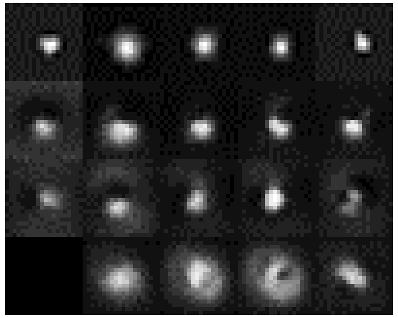
Top 10 second-layer features.



Redundant features.



Sampling from the model.



Visualizations of what the model has learnt.

Results and comparison

	Baseline model	DBN model	Absolute performance increase	Relative performance increase		2 x 5 magnitudes (Brescia et al., 2015)
Macro-averaged F_1 score	0.8321	**0.8682**	0.0361	4.35%		0.9135
F_1 score, STAR	0.8037	**0.8664**	0.0627	7.81%		0,8996
F_1 score, GALAXY	0.9376	**0.9582**	0.0206	2.20%		0,9522
F_1 score, QSO	0.7549	**0.7801**	0.0252	3.32%		0,8866

A comparison of results.

Contributions

- A ready-to-use method.
  Presented a method for astronomical classification tasks that is both interpretable and competitive in results.

- A basis for further research.
  Provided statistically verified evidence in support of the approach of using unsupervised feature-extraction from image data for astronomical classification tasks.

- Cross-domain source code for the algorithms.
  A scalable technical implementation of Restricted Boltzmann Machines and unsupervised Deep Belief Networks.

Future Work and Recommendations

- Training for sparsity.
  - Improved interpretability.
  - Decreasing the complexity of the task for the DBN.
- Multi-wavelength vs single wavelength images.
  - Is one wavelength band enough?
  - Better discrimination by training on each band separately?
- Transfer Learning.
  - Decreased training time for other tasks? (e.g. galaxy morphology, redshift, etc.)
  - Build a catalog of lower-dimensionality intermediate features.

Thank you for listening!

Image from Galaxy Zoo (https://www.galaxyzoo.org/).