Towards a Cognitive Vision System for Astronomy

Aim and Background

South Africa will soon be one of the leading contributors to the field of radio astronomy. The **Square Kilometre Array (SKA)** is anticipated to produce exabytes of data in a single year, at 2 TB/s during operation. To assist in automated and large-scale data processing and analysis, we propose the development of a cognitive vision system for optical galaxy classification. **A Cognitive Vision System (CVS)** or intelligent system, is any system that aims to gather expert knowledge and combine it with predictive capabilities to model and explain future events. Three components will be used in our CVS for astronomy (see Figure 1): Feature extraction, automated machine learning for combined algorithm selection and hyperparameter configuration (CASH), and knowledge representation and reasoning via Bayesian networks. The dataset used is from the The Galaxy Zoo project (see Figure 1)

Feature Selection

Various methods for feature extraction exist for galaxy classification. The hybrid feature selection algorithm developed in this project attempts to solve this problem of choosing the best features for classification by leveraging on the various types of feature selection techniques available. The algorithm makes use of the random forest feature selection technique, recursive feature selection and univariate feature selection to perform feature selection. The algorithm used for feature extraction was the WND-CHARM algorithm, together with hand-engineered features. To evaluate the hybrid feature selection algorithm, comparison were made to the traditional feature selection techniques on a feedforward neural net.

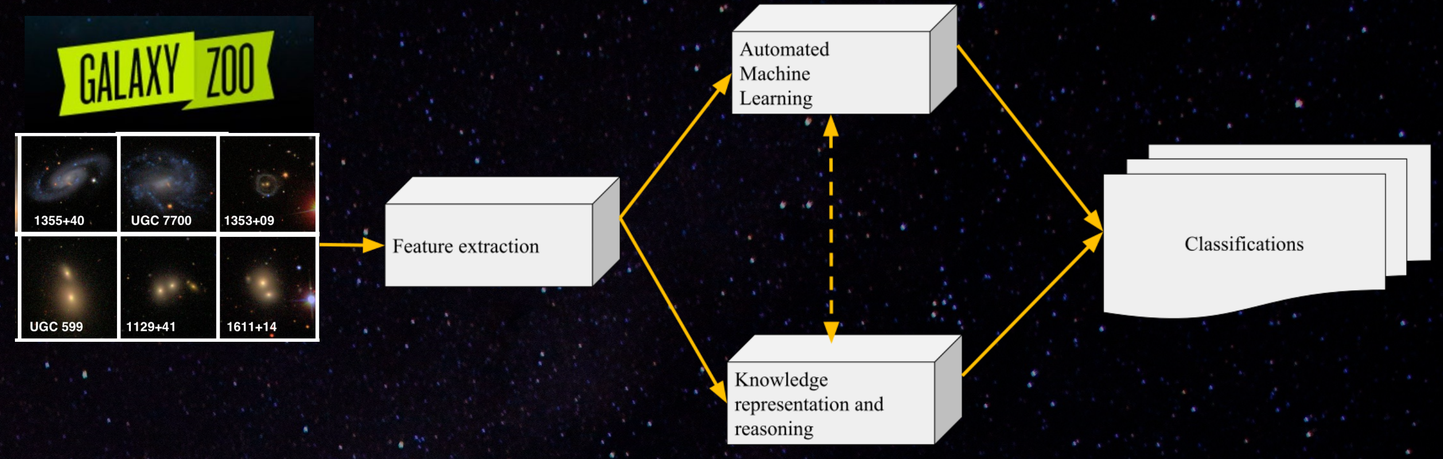


Figure 1) Structure of the proposed CVS for galaxy morphology classification using the Galaxy Zoo dataset (see below). **The Galaxy Zoo,** an openly available, crowd-sourced labelled dataset using optical galaxy images from the Sloan Digital Sky Survey. It contains millions of classified galaxies.

Automated Machine Learning

Part of the problem of using machine learning (ML) in astronomy is that astronomers often lack expert knowledge of state-of-the-art ML models and of how to tune model hyperparameters. The **automated machine learning** component aims to automate and optimize both the choice of ML algorithm and hyperparameters for ML models via CASH through techniques such as Bayesian optimization. Methods used in this project are random search, TPE, SMAC, Hyperband, and Fabolas. The component was evaluated first by running on toy datasets and other astronomy domain datasets to determine the efficacy of automated ML for CASH versus human experts. Then, the methods above were tested for their ability to cope with large astronomy datasets such as the Galaxy Zoo.

Knowledge Representation and Reasoning

The Bayesian network incorporates expert knowledge in the design of the network topology in order to use shape descriptors extracted from an image to identify the type of galaxy present within that image. The parameters of the network were learned using Variational Bayesian approximation from the Kaggle Galaxy Zoo dataset.(see Figure 2 c)) The identification produced by the network is compared with the classifications made by the automated ML module to check agreement, and to fine-tune classifications on unseen data.

Results and Discussion

Results for feature selection showed that only 10 features were necessary to achieve good RMSE scores (see Figure 2 (a)), with the neural net nearly performing as well as a Pearson's correlation feature selection method. Results for the automated ML indicate better-than-human performance on tuning ML algorithms on both toy and astronomy domain problems such as supernovae classification, as well as for galaxy classification for the Galaxy Zoo project. Model configurations that were found achieved better RMSE than previously known configurations on the Kaggle GZ competition, the lowest score being 0.069 versus a previous best of 0.074. Additionally, it was shown that data-efficient methods such as Fabolas can improve speed on model configuration and selection (see Figure 2 (b) below).

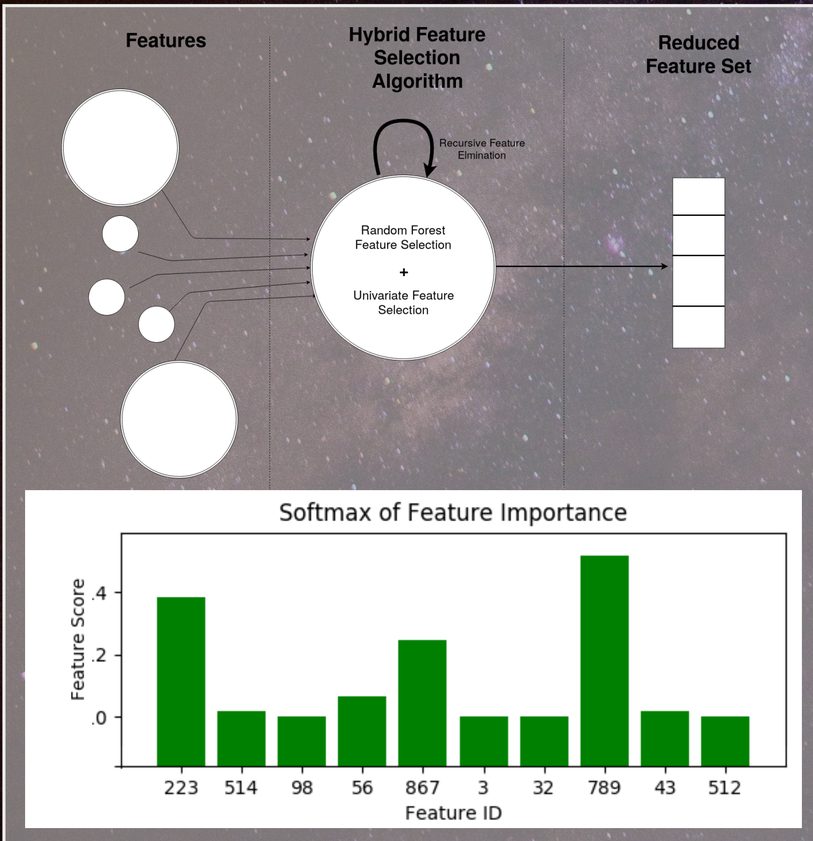
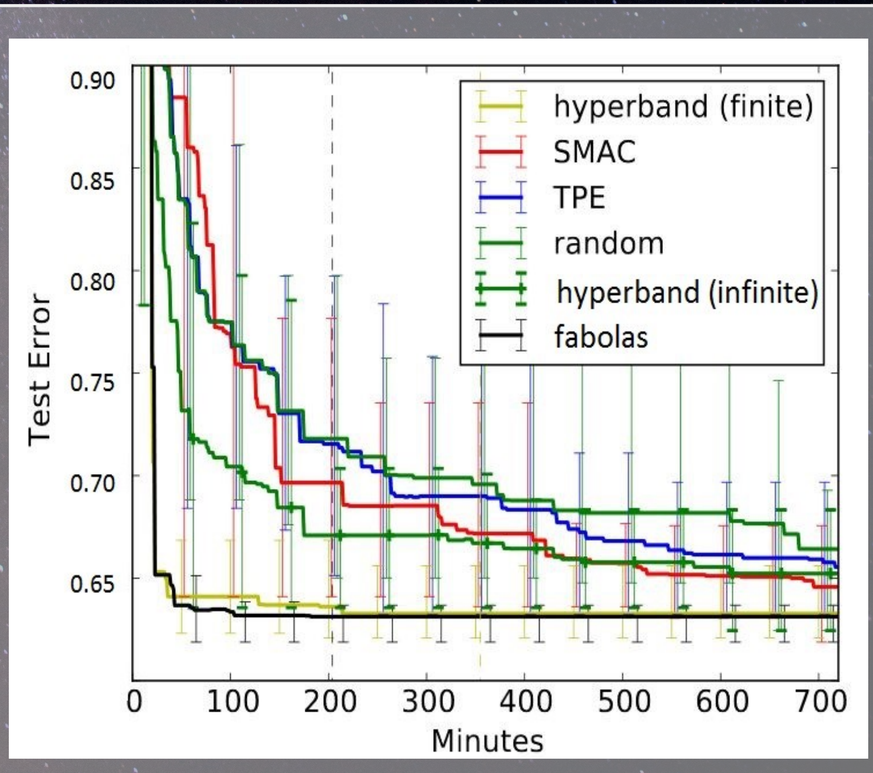
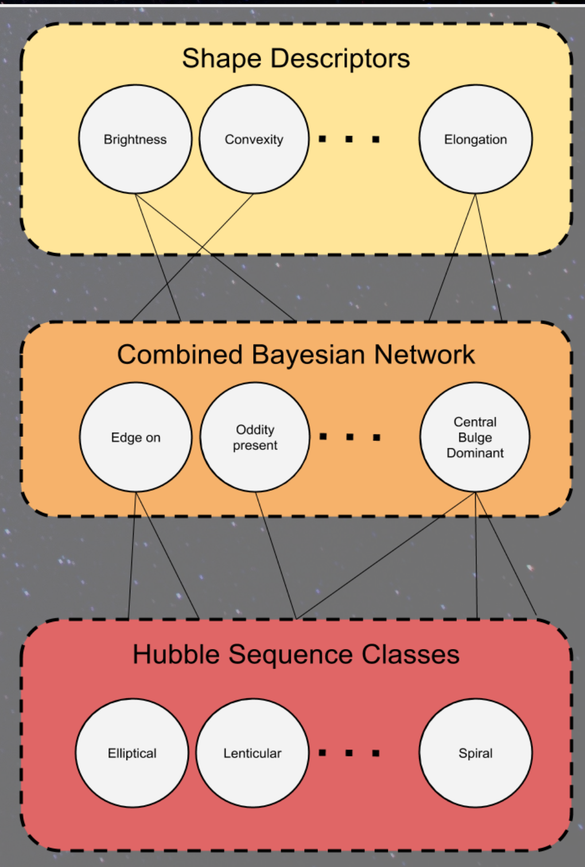


Figure 2) (a) Feature extraction and selection module procedure (above): Extract features, and pass through RFE, or Univariate feature selection. (below) Feature importance of 10 selected features from the WND-CHARM feature set



(b) Performance (RMSE) of various algorithms for **combined algorithm and hyperparameter optimization**. Experiments were run on a convolutional neural net on GZ images. Data subsampling techniques (fabolas) and successive halving (hyperband) find good configurations faster.



(c) Example of a **Bayesian network configuration** using shape descriptors as observations and output labels as evidence.

Conclusions

In conclusion, feature extraction is an important tool for performing both machine learning, and knowledge reasoning. Feature selection can help improve these methods by leveraging the best features. Hyperparameter optimization and model selection are valuable tools that can serve to alleviate much of the experimentation needed to configure algorithms and to assist in automating expert performance. Knowledge reasoning using Bayesian networks is able to provide additional information relating to the identifications made unlike black-box ML approaches such as Neural Nets. While Bayesian networks may not be perfectly suited for pure regression problems, the ability for expert knowledge to be incorporated into the network design as well as the ability to learn using the expert knowledge are key advantages to solving problems that require reasoning.

University of Cape Town
Rondebosch,
Cape Town, 7700,
South Africa

Team Members:
Victor Gueorguiev,
Roy Hana Eyono,
Julius Stopforth

Supervisor:
Prof. Deshen Moodley

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRI

CENTRE FOR ARTIFICIAL INTELLIGENCE RESEARCH