Thesis Outline: Machine Learning Experiments on Multi-Modal Astronomical Data

Farukh Rustamov

March 5, 2025

Contents

1 Introduction

This thesis investigates the application of machine learning methods to predict the star formation rate (SFR) in astronomical objects based on photometric and spectroscopic data from the Sloan Digital Sky Survey (SDSS). The study is motivated by the need to process and analyze large volumes of data and to develop optimal methods for the automated interpretation of observational data, a critical area in modern astronomy.

1.1 General Description and Relevance of the Study

Originally, the entire SDSS dataset weighs hundreds of terabytes. Thanks to the HISS-CUBE system installed on the RCI cluster, a subset of the data has been set up, containing more than 150,000 objects. However, for proper statistical analysis and to minimize systematic errors, a filtering procedure is applied using the `FLAG` keyword. When `FLAG` is set to 0, the SFR estimation is considered reliable, and further calculations (e.g., SFR/M*) can be carried out without introducing subtle biases. As recommended on the dataset description page `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/raw_data.html`, we use only objects with `FLAG` = 0, reducing the dataset to about 11,000 objects (approximately 10,000 galaxies and 1,000 quasars).

1.2 SDSS Data Description

The SDSS dataset provides a unique opportunity to study the properties of astronomical objects using comprehensive observations. Each object in the sample is characterized by:

- **Five-Band Photometry.** For each object, five images are available corresponding to different spectral bands (denoted as u, g, r, i, and z). Each image captures a specific portion of the spectrum, enabling a detailed analysis of the structural and physical properties of the objects.
- **Spectroscopic Data.** In addition to the photometric images, each object is provided with a spectrum that offers information on its chemical composition, temperature, and dynamics.

Special attention is given to the quality of the data. The images are available in four resolutions: 64x64, 32x32, 16x16, and 8x8 pixels. Similarly, the spectra are provided in four quality levels, allowing for an experimental selection of the optimal trade-off between data detail and processing speed. The processing and conversion of the raw data into a format suitable for machine learning were performed using HISS-CUBE. More details on this method are available in the publication at `https://www.sciencedirect.com/science/article/pii/S2213133721000172`.

1.3 Prediction of Star Formation Rate (SFR)

One of the primary objectives of this research is to predict the star formation rate (SFR) using the available SDSS data. In our dataset, the SFR is represented by the column `AVR` (mean value of the SFR distribution). Detailed information on the SFR parameter is available at `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html`. By applying machine learning techniques, we aim to evaluate the feasibility of accurately predicting SFR using various types of input data.

- Predicting SFR using only photometric images.
- Predicting SFR using only spectroscopic data.
- Employing a multimodal approach that combines both images and spectra.

A comparative analysis of these approaches will help determine the method that yields the highest accuracy and whether the multimodal approach offers significant improvements over single-modality models.

1.4 Research Challenges

Working with the SDSS data presents several challenges:

1. **Data Filtering.** The original dataset comprises over 150,000 objects, but to ensure the reliability of the SFR estimates, only objects with `FLAG` = 0 are used. This filtering minimizes potential distortions caused by alternative estimation methods.
2. **Quality of Images and Spectra.** The availability of multiple quality levels for both images and spectra allows for optimization of the data processing pipeline. However, determining the optimal resolution that balances processing speed and prediction accuracy is non-trivial.
3. **Multiple Objects in One Image.** There is a possibility that a single image may contain multiple objects. This issue must be addressed since overlapping signals can degrade the performance of machine learning algorithms. Effective methods for automatic object detection and isolation need to be developed.

Addressing these challenges is crucial for improving the reliability and accuracy of the predictive models.

1.5 Comparative Analysis: The Scene Dataset Example

To preliminarily test the advantages of the multimodal approach, experiments were conducted on the Scene dataset, available at `https://www.kaggle.com/datasets/birdy654/scene-classificatio`
This dataset includes two types of data:

- **Images.** Frames extracted from videos that represent eight different types of environments.
- **Audio Data.** For each image, a set of MFCC (Mel-Frequency Cepstral Coefficients) attributes is provided, representing the audio characteristics corresponding to that frame.

Key aspects of the Scene dataset include:

- Images are extracted at regular intervals (e.g., one frame per second) and are accompanied by corresponding audio features.
- The dataset supports two classification tasks: binary (Indoors/Outdoors) and multi-class (8 types of environments).

Preliminary machine learning results indicate that the multimodal approach significantly improves accuracy. For instance, decision tree experiments yielded:

- **Audio:** CLASS1 accuracy of 0.81 and CLASS2 accuracy of 0.66.
- **Combined:** CLASS1 accuracy of 0.97 and CLASS2 accuracy of 0.92.

Additionally, neural network experiments showed:

- **Audio:** CLASS2 test accuracy of 0.94.
- **Images:** CLASS2 test accuracy of 0.99.
- **Combined:** CLASS2 test accuracy of 0.99.

These results support the hypothesis that integrating modalities can enhance prediction accuracy.

1.6 Objectives and Tasks

The primary objective of this thesis is to develop an optimal methodology for predicting the star formation rate (SFR) using SDSS data. To achieve this, the following tasks will be addressed:

1. Perform a detailed analysis of the raw data, assess its quality, and apply filtering using the `FLAG` parameter.
2. Develop algorithms for the automatic detection and isolation of objects within images to address the issue of multiple objects per image.
3. Investigate the impact of different quality levels of images and spectra on prediction accuracy and determine the optimal resolutions for each modality.
4. Compare the effectiveness of models using single modalities (either images or spectra) with multimodal approaches that combine both data types.

5. Conduct a comparative study using the Scene dataset as a preliminary experiment and then adapt the findings to the analysis of SDSS data.

Achieving these tasks will not only improve the accuracy of SFR predictions but also optimize computational efficiency, which is crucial when dealing with large-scale astronomical datasets.

1.7 Terminology and Illustrations

Understanding key terminology and methods is essential for interpreting the data and the results obtained from the machine learning experiments.

1.7.1 Spectra and Spectral Analysis

Definition of a Spectrum: A spectrum in astronomy represents the dependence of an object's emitted intensity on wavelength (or frequency). When light passes through a prism or a diffraction grating, it is dispersed into its constituent wavelengths, ranging from the ultraviolet to the infrared. Specialized spectrographs attached to telescopes are used to record these spectra.

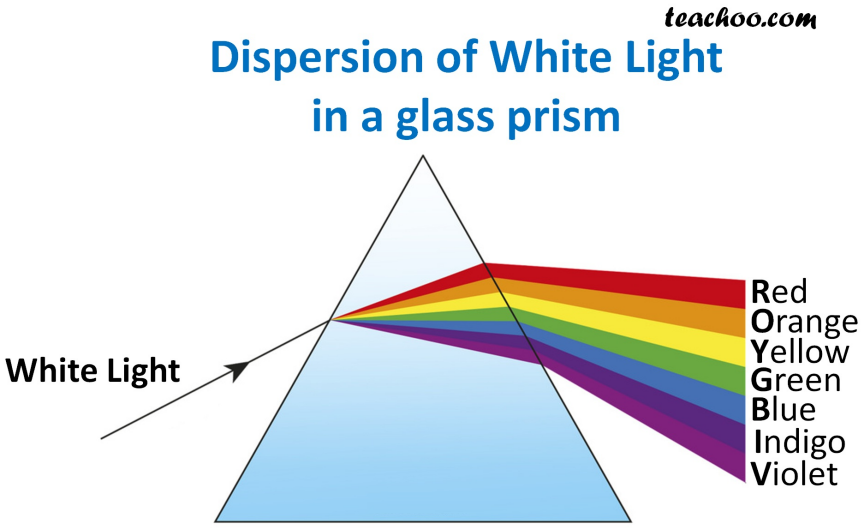

Figure 1: White-light decomposition through a prism.

Why Spectral Analysis Is Needed: Spectral analysis is a powerful tool in astronomy because:

- **Chemical Composition:** Each element exhibits characteristic spectral lines. By analyzing these lines, astronomers can determine the chemical makeup of celestial objects.
- **Velocity Measurements:** Shifts in the wavelengths of spectral lines provide insights into the motion of celestial objects.
- **Physical Conditions:** The intensity of emission or absorption lines can indicate conditions such as ionization levels.

1.7.2 The SDSS *u, g, r, i, z* Filters

One of the most influential surveys in modern astronomy is the Sloan Digital Sky Survey (SDSS). Its photometric system is built around five broadband filters:

- **u**: Near-ultraviolet (centered around 354 nm),
- **g**: Blue-green (centered around 477 nm),
- **r**: Red (centered around 623 nm),
- **i**: Near-infrared (centered around 762 nm),
- **z**: Infrared (centered around 913 nm).

These filters allow for the construction of a broadband Spectral Energy Distribution (SED) for each object.

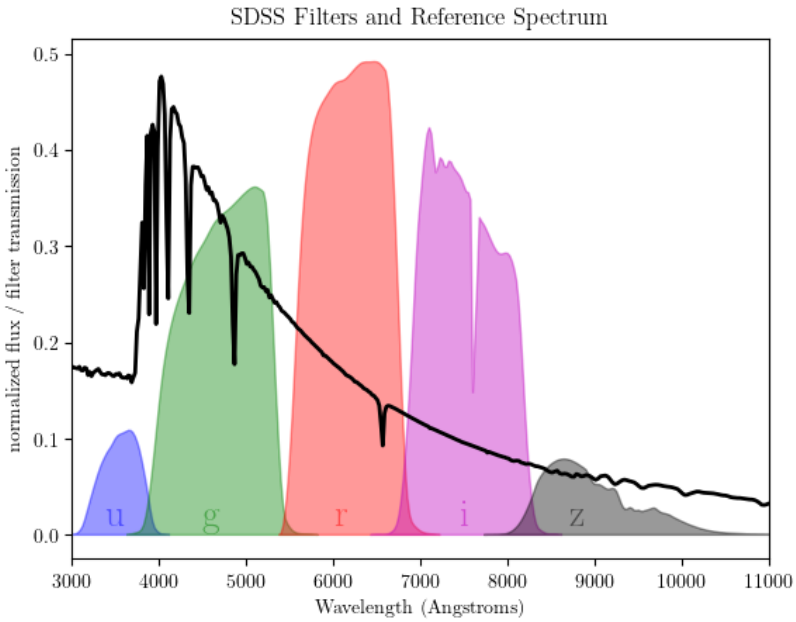

Figure 2: Transmission curves of the SDSS *u, g, r, i, z* filters.

1.7.3 Star Formation Rate (SFR)

What Is SFR: The Star Formation Rate (SFR) quantifies the rate at which a galaxy forms new stars, typically expressed in solar masses per year ($M_\odot\,\mathrm{yr}^{-1}$).

How SFR Is Determined: A common method to estimate SFR is by measuring the intensity of specific emission lines (e.g., Hα) in a galaxy's spectrum:

$$\mathrm{SFR} \propto L(\mathrm{H}\alpha), \tag{1}$$

with a calibration given by:

$$\mathrm{SFR}(M_\odot\,\mathrm{yr}^{-1}) \approx 7.9 \times 10^{-42}\, L(\mathrm{H}\alpha)\,(\mathrm{erg\,s}^{-1}).$$

Other methods may use ultraviolet or infrared measurements, depending on data availability and dust obscuration.

Why SFR Matters:

- **Galactic Evolution:** SFR is a key parameter in understanding galaxy evolution.
- **Energy Output:** High SFRs indicate intense star formation, affecting the interstellar medium.

1.8 Conclusion of the Introduction

In summary, this thesis explores the prediction of the star formation rate (SFR) in astronomical objects using SDSS data. By leveraging both photometric and spectroscopic data, and ensuring data quality through filtering with `FLAG` $= 0$, the study lays the groundwork for advanced machine learning experiments. Preliminary analyses on the Scene dataset suggest that multimodal approaches offer significant improvements in prediction accuracy.

2 Data Acquisition and Preprocessing

2.1 SDSS Dataset Overview

The SDSS dataset offers comprehensive photometric and spectroscopic observations of astronomical objects. The data includes five-band photometry (u, g, r, i, z) and detailed spectral information that captures chemical composition, temperature, and dynamics.

2.2 HiSS-Cube Implementation

Data processing and conversion were performed using the Hierarchical Scientific Storage Cube (HiSS-Cube) based on HDF5. This structure enables the seamless association of five-band images, full spectroscopic data, and computed SFR values in a hierarchical format.

2.3 Data Quality and Resolution Considerations

The images are provided in multiple resolutions (64x64, 32x32, 16x16, 8x8 pixels) and the spectra in different quality levels. Balancing data resolution with processing speed is critical for optimizing predictive performance.

2.4 Handling Multiple Objects per Image

A challenge in the dataset is the potential presence of multiple objects within a single image. Proposed solutions include:

- Object removal techniques to filter out extraneous objects.
- Cropping images to isolate the central object.

3 Exploratory Data Analysis

3.1 Visualization and Inspection of Imaging Data

Initial exploration involves visualizing sample images from the dataset to assess image quality and content distribution.

3.2 Spectral Data Exploration

Examination of key spectral features—such as emission lines and continuum—is performed. Preprocessing steps include normalization and noise reduction.

3.3 Analysis of SFR Distribution

Statistical analysis of the SFR values (represented by the `AVR` column) is conducted using histograms and boxplots to understand its distribution and range, informing subsequent regression tasks.

4 Machine Learning Models: Single-Modality Baselines

4.1 Decision Tree Models on Spectroscopic Data

Baseline models using decision trees are implemented on the spectral data to predict SFR. The model extracts relevant features from emission lines and continuum patterns, and performance is evaluated using regression metrics.

4.2 Decision Tree Models on Photometric Data

Similarly, decision tree models are applied to the five-band imaging data. Comparisons between spectroscopic and photometric baselines provide insights into the strengths and limitations of single-modality approaches.

4.3 Baseline Performance Analysis

The performance of single-modality models is compared using metrics such as RMSE, MAE, and R^2. This analysis serves as a foundation for exploring more complex, multimodal fusion strategies.

5 Multi-Modal Machine Learning Approaches

5.1 Fusion Neural Network Architectures

Advanced architectures combine convolutional neural networks for five-band imaging data with specialized networks for spectral feature extraction. Various fusion strategies are explored:

- **Early Fusion:** Integrating data at the input level.
- **Intermediate Fusion:** Merging feature representations extracted from each modality.
- **Late Fusion:** Combining decisions or predictions from independent models.

5.2 Contrastive Learning and Joint Embedding

Contrastive learning is applied to create a joint embedding space that captures complementary information from both imaging and spectroscopic data. In this approach, similar pairs (e.g., corresponding image and spectrum of the same object) are drawn closer in the embedding space, while dissimilar pairs are pushed apart, thereby enhancing the overall predictive performance.

5.3 Implementation Details and Hyperparameter Tuning

Details regarding network architecture, training strategies, data augmentation, and optimization techniques are discussed to fine-tune the multimodal models.

6 Experimental Design and Evaluation

6.1 Experimental Setup

The experimental setup includes a clear division of the dataset into training, validation, and testing subsets. Evaluation metrics such as RMSE, MAE, and R^2 are used to assess model performance.

6.2 Comparison: Single-Modality vs. Multi-Modal Models

Detailed experiments compare the performance of decision trees and fusion neural networks. Statistical analyses and validation against theoretical models are used to determine the optimal approach.

6.3 Fusion Strategy Analysis

A comparative discussion of early, intermediate, and late fusion strategies is presented. Additionally, the role of cross-attention mechanisms in enhancing feature integration between modalities is examined.

7 Discussion and Conclusion

7.1 Summary of Findings

Key results and contributions are recapitulated, emphasizing the improved accuracy achieved by multimodal fusion techniques in predicting the star formation rate.

7.2 Limitations and Challenges

Current challenges—such as handling multiple objects per image and addressing data heterogeneity—are discussed. Limitations of the current models and potential improvements are also addressed.

7.3 Future Work

Future research directions include:

- Extending the methodologies to predict additional astronomical parameters beyond SFR.
- Applying the developed techniques to upcoming large-scale surveys.
- Exploring more advanced fusion techniques and learning paradigms.

References