spectRoscopic Analysis of oBjects for type 1B supernova classIficaTion (RABBIT)

Abby Stokes, Lexi Leali, Jacynda Alatoma

April 2024

1 Abstract

This study investigates the classification of supernovae based on spectral features, inspired by the the SN1a score model in identifying Type Ia supernovae using binary classification techniques. We attempted to duplicate the SN1a score results and extended the classification technique to include Type Ib supernovae by utilizing Support Vector Machines (SVM), Random Forest, and Simple Logistic Regression/Multinomial Logistic Regression. Our model is aimed at differentiating Type Ib supernovae from other supernova types, based on their characteristic non-ionized Helium emission line at 587.6 nm. This research adds to the automated detection and categorization of supernovae, particularly Type Ia and Ib, by utilizing spectrum analysis and machine learning methods. **add some info about the results we are starting to see **

2 Introduction and Background

SN1a score (cite) has successfully shown that a binary classifier can be trained to identify type Ia supernovae from their spectra. Type Ia supernovae are characterized by their emission line of ionized silicon (Si II) at 657 nm. (add what methods SN Ia score used, briefly). Inspired by the success of SN1a score, we replicated their results to identify type Ia supernovae, and then further extended the classification scheme.

Type II supernovae show emission lines of hydrogen in their spectra, while type Ia does not. Further classification of type II gets more complicated, with the overall shape of the spectra characterizing the sub-type, thus making the features harder to identify by eye. For this reason our model will only extend to identify type Ib supernovae, which are characterized by a distinct emission line of non-ionised Helium in their spectra, at 587.6 nm.

3 Methodology

3.1 Dataset

brief description of data set (Note: could add more supernovae training data from sources Michael mentioned)

The data used comes from two different dataset, both from the Zwicky Transient Facility (ZTF), which were combined based on ZTF ID, which is a unique identifier for Supernovae events. The first dataset comes from the ZTF Bright Transient Survey (BTS), and includes observations of 14155 SN events, with columns for RA, Dec, peak time, peak magnitude, and more, as well as a column for the class label. This dataset includes 3082 SN Ia, and 124 SN Ib events. The second dataset comes from includes detailed information on the spectra (flux per wavelength) for each Supernova event. The wavelength data contains 214 flux observations between the wavelengths of 1000 to 10,000 angstroms. This dataset will be used to extract features based on wavelengths of interest for later classification of SN Ib.

3.2 Data Preprocessing

Oversampling

The first step in the data pre-processing, after merging both datasets on ZTF ID was to perform oversampling of the minority class. In the original dataset, there were 3082 SN Ia, and only 124 SN Ib. Imbalanced datasets such as this can lead to problems with using performance metrics such as accuracy, especially when the minority class is the class of interest. The most common technique for handling imbalanced data is to use Synthetic Minority Over-sampling Technique (SMOTE) (reference 1). SMOTE works essentially by randomly over sampling the minority class to balance the dataset. In our case, we applied SMOTE using the RandomOverSampler class in imblearn library. Our final dataset contained 3067 SN Ia and 3067 SN Ib.

Feature Extraction

The next step was to incorporate the spectra information as a feature that could be used for classification. We decided to extract this feature by pulling the maximum flux value for each SN event between wavelengths of 550 and 600 nm. This should ideally capture the helium emission line at 587.6 and hence improve model performance when distinguishing between type Ia and Ib supernovae.

Splitting Data

Finally, we split our data into train, test, and validation sets using a split of 80/10/10. The training dataset will be used to train each model on finding the ideal parameters, and the validation set will be used to fine tune any hyperparameters of the model using 10-fold cross validation. Finally, we will evaluate model performance and generalizability using the test dataset.

3.3 Random Forest Method

Lexi

3.4 Support Vector Machine Method

Abby

The next method that we used was Support Vector Machine (SVM). SVM works for binary classification by finding the hyperplane that best separates the two classes of points in a high-dimensional space, and then maximizes the margin between the closest two points of the different classes. SVM can be easily extended to multi-class problems using one-vs-rest, where multiple binary classifiers are combined to distinguish between each class and the remaining ones. Additionally, while SVM at its basic form has a linear decision boundary, kernel methods (such as RBF or polynomial) can be used to apply SVM in contexts where the separation between classes is non-linear.

We used binary and multi-class SVM on our dataset, and found that a linear kernel performed the best.

Results (precision, recall, F1 score, Confustion matrix, ROC curve)

3.5 Simple Logistic Regression and Multinomial Logistic Regression Methods

Jacynda

4 Results

figures, quantitative measures of model performance, comparison of all three models

5 Discussion

One limitation of this model comes from the use of oversampling. A potential downside of SMOTE is that it may introduce synthetic samples that are unrealistic or noisy, leading to over-fitting or reduced generalization performance (reference 2). Further, oversampling assumes that the minority data that is present is representative of the true distribution, which may not be the case especially when we have very few samples.

Another limitaion of this model is ...

6 Conclusion

- Which model performed the best?
  - Model limitations and assumptions
  - Implications and future work

References

1. SMOTE: https://arxiv.org/pdf/1106.1813.pdf 2. SMOTE: https://www.ncbi.nlm.nih.gov/pmc/articles/
3. SVM: https://link.springer.com/chapter/10.1007/978-1-4899-7641-3$_9$