Use of Statistical Learning Methods to Predict Imbalanced Data

Ty Nunley[1], Anca Constantin[2], & Nusrat Jahan[1]
[1]Department of Mathematics and Statistics, James Madison University,
[2]Department of Physics and Astronomy, James Madison University

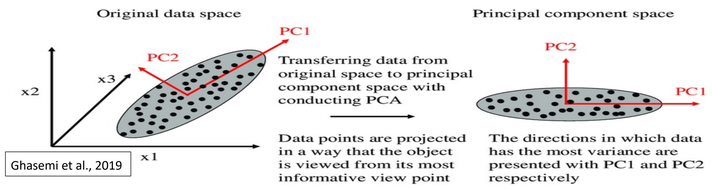
Abstract

We present an investigation of machine learning methods and synthetic minority oversampling techniques (SMOTE) to develop a classification tool to predict classes for imbalanced data. We demonstrate our findings using the data set of Microwave Amplification by Stimulated Emission of Radiation (maser) emissions from galaxy centers. The maser phenomenon is important because it can be used as a unique tool to constrain both the masses of supermassive black holes and the current cosmological models. Unfortunately, maser detections are extremely rare, accounting for ~3 -5% of all surveyed galaxies, leading to a large imbalance in the data. In this work, we aim to classify the emission data into two categories: non-masers and masers based on a priori independent information about their host galaxies.

Tools used for Statistical Learning

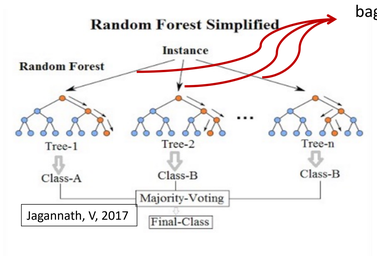
PCA – Unsupervised Analysis

PCA is an unsupervised (exploratory) statistical learning tool used to reduce the dimensionality of a data set. Our goal is to describe the most amount of variability in the data using the least number of dimensions or predictors possible. Through PCA, we are creating linear combinations, or principal components (PCs, or combinations of eigenvectors) of the parameters that characterize a sample of objects (i.e., galaxies). Eigenvalues show how much variability in the data is explained by each PC.



Random Forests – Supervised Analysis

Random Forests (RF) are a supervised learning method used to classify a binary response through a "forest" of decision trees created through bagging (bootstrapping and aggregating) data.



bagging

- Bootstrapping: the process of randomly sampling data multiple times, with replacement.
- Aggregating: tallying the data as it is run down the decision trees created by the random forest algorithm.

Boosting (LogitBoost)

Boosting is a classification method used to help reduce predictive error in classification. It is a method that combines multiple weak learners (i.e., random guesses) to a singular, strong learner. Compared to bagging, which trains weak learners simultaneously, boosting instead trains weak learners sequentially. Because of this, there is a large bias reduction in boosting methods because the tool is improved as more weaker learners are trained. LogitBoost is a method used to minimize the logistic loss of an additive regression. LogitBoost is particularly useful in classifying binary data. It is akin to AdaBoost, another common boosting method, which minimizes the exponential loss of an additive regression. LogitBoost has been found to be slightly more effective than AdaBoost, which is why it is used here (Friedman et al., 2000).

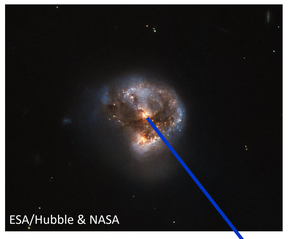
Tools used for Statistical Learning (cont.)

SMOTE (Synthetic Minority Oversampling Technique)

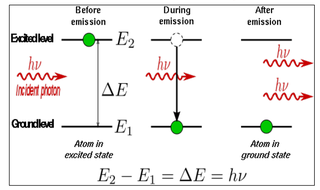
"SMOTEing" synthetically generates new data through finding k nearest minority class neighbors, finds a line segment to connect the data to the kth nearest neighbor, then generates a new data point somewhere on that line segment. It is used to help balance data with unbalanced classes. This method oversamples the minority class and undersamples the majority class to create equal class sizes (Bowyer et al., 2002). Using the new "smoted" data, classification (i.e., random forest analysis) can then be run to classify the data.

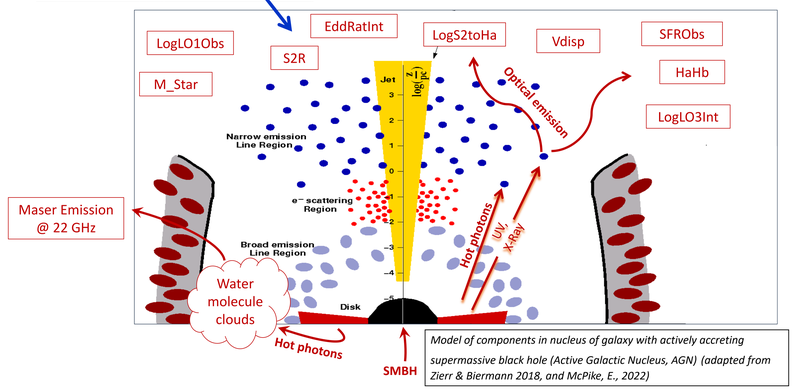
Exploring Maser Galaxies

MASER – ***M**icrowave* **A**mplification by **S**timulated **E**mission of **R**adiation



Model of stimulated atom emitting a photon





Model of components in nucleus of galaxy with actively accreting supermassive black hole (Active Galactic Nucleus, AGN) (adapted from Zierr & Biermann 2018, and McPike, E., 2022)

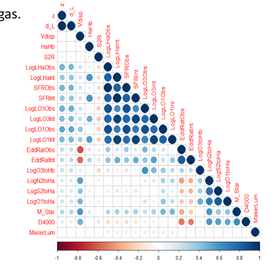
Data and Parameter Information

Parameter Information

The data was sampled from the Maser Cosmology Project, one of the largest catalogs of surveyed galaxies in 22GHz for water maser emission. It was then cross matched with data from the Sloan Digital Sky Survey to create a larger and more coherent catalog of data.

z: Redshift; measure of how fast the galaxy is receding away from us.
d_L: Calculated luminosity distance from the redshift.
Vdisp: Velocity dispersion of stars in the host galaxy; measures the mass of the central black hole.
HaHb: The Balmer Ratio; A measure of the amount of obscuration along the line of sight within the galaxy.
S2R: A line ratio of two emission lines from ionized S; a measure of the density of the emitting gas.
Intrinsic luminosities are obtained from the observed luminosities and are corrected for dust obscuration using the HaHb parameter.
LogLHaObs & LogLHaInt: The H_α observed/intrinsic luminosity.
SFRObs & SFRInt: Star formation rate, obtained from the H_α emission (in luminosity).
LogLO3Obs & LogLO3Int: Observed/intrinsic luminosity of the doubly ionized O.
LogLO1Obs & LogLO1Int: Observed/Intrinsic Luminosity of the neutral O.
EddRatObs & EddRatInt: Eddington ratio; a measure of efficiency of black hole acceleration, based on observed/intrinsic luminosity.
M_Star: Mass of the star in entire host galaxy (in solar masses).
D4000: Measure of the age of the stellar population in the galaxy (based on strength of absorption features in optical spectrum).
Line-flux ratios used as diagnostics for identification of excitation by black hole acceleration:
LogO3toHb: $\frac{O(ionized)}{H}$ **LogN2toHa**: $\frac{N(ionized)}{H}$
LogS2toHa: $\frac{S(ionized)}{H}$ **LogO1toHa**: $\frac{O(neutral)}{H}$

An example correlation matrix of the dataset that reflects the interdependence of parameters.

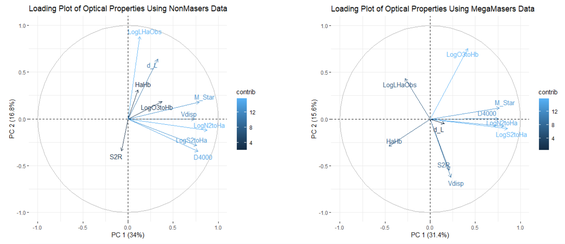


Blue indicates a strong positive correlation;
Red indicates a strong negative correlation;
White indicates weak/no correlation.

Variable Selection

PCA

We run PCA separately on non-maser and mega-maser galaxy samples. We present loading plots that reflect the contributions of individual parameters (optical measurements) to the two main PCs (the x- and y-axis). Brighter colors correspond to higher percentages contributed by the parameters to the PCs. We find that the mega-masers and non-masers are governed by different types of contributions to the main PCs that account for 31-34% and 15-17% variation within the samples, respectively.



Vdisp, HaHb, S2R, LogLHaObs, and LogO3toHb arrows point in significantly different directions within the mega-maser and non-maser samples, suggesting that these parameters carry more significant information about what distinguishes between the two types of galaxies.

Results

Classification

With various splits into training and testing sets, the random forest, SMOTE, and LogitBoost algorithms allow us to create non-maser and mega-maser classifiers based on the optical data, concentrating on the five parameters that PCA revealed as most relevant. While accuracy measures how accurate the tool can predict the data, the Kappa value measures how accurate this data is once random chance is introduced. Therefore, a high value of Kappa is desired when testing the data. Sensitivity refers to the probability that the tool makes a correct non-maser prediction and specificity refers to the probability that the tool makes a correct mega-maser prediction.

Test Set Kappa Values for Various Methods

Split (Train/Test)	Random Forest	RF with SMOTE	LogitBoost	LB with SMOTE
50%/50%	0.1982	0.3157	0.3461	0.3142
60%/40%	0.2814	0.3049	0.2384	0.3008
75%/25%	0.2586	0.3123	0.2122	0.2914

Test Set Sensitivity Values for Various Methods

Split (Train/Test)	Random Forest	RF with SMOTE	LogitBoost	LB with SMOTE
50%/50%	0.991	0.8977	0.9657	0.985
60%/40%	0.9831	0.9041	0.9642	0.9837
75%/25%	0.982	0.8919	0.9636	0.9748

Test Set Specificity Values for Various Methods

Split (Train/Test)	Random Forest	RF with SMOTE	LogitBoost	LB with SMOTE
50%/50%	0.1429	0.6571	0.6154	0.2418
60%/40%	0.2143	0.6071	0.3571	0.2338
75%/25%	0.2222	0.6667	0.3	0.2439

As shown by the tables above, using SMOTE in both the random forest method and the LogitBoost methods significantly improves the Kappa value for all splits except LogitBoost with SMOTE. Sensitivity is above 95% for all the methods due to the very imbalanced nature of the data. For the same reason, specificity is very poor. SMOTE does improve specificity substantially for the RF.

Future Directions

- Investigate variations of SMOTE algorithms to build a stronger classifier of mega-maser and non-maser galaxies.
- Design and develop a web tool with a user-friendly interface that provides mega-maser/non-maser classifications of various probability levels for any input feature set involving galactic parameters tested apriori with (published) observations to correlate with water maser emission of various morphologies.

References

Braatz, J., et al, 2009, ApJ, 695, 287**; Braatz, J,** et al, 2018**; Brinchmann, J.** et al, 2004, **Bowyer** et al., 2002, **Friedman, J.** et al., 2000, **Jagannath, V**, 2017, **Kauffman, G.,** et al, 2003, **Kauffman G.,** et al, 2004; **Landis, J. R.**, & Koch, G. G., 1977, **McPike**, E., 2022; **Ullah, I.**, et al, 2020, **Zierr, C.** & Biermann, P., 2018, A&A, 69, 1