4033/5033: Final Project Proposal

Naeem Shahabi Sani

1 Problem Statement

This work focus on the commonly used Random Forest algorithm [3], and modify it to properly treat measurement uncertainties. RF is mostly used for classification and regression as a supervised algorithm [4][2][8][7][6][5][9][11], but it can also be used for unsupervised learning [1][10]. In general, this project is a correction on Random Forrest, which applies uncertainty to both input features and labels assigned to samples, and treats with features and labels as a random distribution.

2 Methodology

The PRF method presented here is an RF-based classification algorithm meant to improve the classic RF's prediction capabilities. This is completed by taking into account the input data's uncertainties and using their informative content.

3 Data Preparation

One of these datasets will be used:

1. synthetic dataset with two classes and 10000 samples is used. 5,000 samples are used for the training and 5,000 samples are used for testing. Using scikit-learn, we created synthetic classification data.

2. Titanic Dataset

3. House Prices Dataset

4 Evaluation Plan

Four noise models are examined to test the strength of the PRF in noisy datasets.

1. Noise in the labels

2. Noise in the features (simple case)

3. Noise in the features (complex case)

4. Different noise characteristics in the training and the test sets.

References

[1] Dalya Baron and Dovi Poznanski. The weirdest sdss galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society*, 465(4):4530–4555, 2017.

[2] JS Bloom, JW Richards, PE Nugent, RM Quimby, MM Kasliwal, DL Starr, D Poznanski, EO Ofek, SB Cenko, NR Butler, et al. Automating discovery and classification of transients and variable stars in the synoptic survey era. *Publications of the Astronomical Society of the Pacific*, 124(921):1175, 2012.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Samuel Carliles, Tamás Budavári, Sébastien Heinis, Carey Priebe, and Alexander S Szalay. Random forests for photometric redshifts. *The Astrophysical Journal*, 712(1):511, 2010.

[5] AA Miller, MK Kulkarni, Y Cao, RR Laher, FJ Masci, and JA Surace. Preparing for advanced ligo: A star–galaxy separation catalog for the palomar transient factory. *The Astronomical Journal*, 153(2):73, 2017.

[6] A Möller, V Ruhlmann-Kleider, C Leloup, J Neveu, N Palanque-Delabrouille, J Rich, R Carlberg, C Lidman, and C Pritchet. Photometric classification of type ia supernovae in the supernova legacy survey with supervised learning. *Journal of Cosmology and Astroparticle Physics*, 2016(12):008, 2016.

[7] Karim Pichara and Pavlos Protopapas. Automatic classification of variable stars in catalogs with missing data. *The Astrophysical Journal*, 777(2):83, 2013.

[8] Karim Pichara, Pavlos Protopapas, D-W Kim, J-B Marquette, and Patrick Tisserand. An improved quasar detection method in eros-2 and macho lmc data sets. *Monthly Notices of the Royal Astronomical Society*, 427(2):1284–1297, 2012.

[9] PM Plewa. Random forest classification of stars in the galactic centre. *Monthly Notices of the Royal Astronomical Society*, 476(3):3974–3980, 2018.

[10] Itamar Reis, Dovi Poznanski, and Patrick B Hall. Redshifted broad absorption line quasars found via machine-learned spectral similarity. *Monthly Notices of the Royal Astronomical Society*, 480(3):3889–3897, 2018.

[11] Suk Yee Yong, Anthea L King, Rachel L Webster, Nicholas F Bate, Matthew J O'Dowd, and Kathleen Labrie. Using the properties of broad absorption line quasars to illuminate quasar structure. *Monthly Notices of the Royal Astronomical Society*, 479(3):4153–4171, 2018.