Weekly Research Project Status Report

Project Title : Stellar spectral classification using Active learning approach
Report Date : 12th April, 2025

Summary of Progress

- We have successfully implemented the Active Machine learning on our dataset.
- For Active Learning, we divided our total data (59805 spectra) in three datasets :
    1. Training set – 200 spectra
    2. Pool of unlabelled data – 47108 spectra
    3. Test set – 11777 spectra
- Currently we have implemented uncertainty sampling for querying the data from the pool of unlabelled data. In uncertainty sampling, the model randomly selects a new sample from the pool of unlabelled data and then asks oracle for the label corresponding to it.
- We used Random Forest Classifier as our base model and performed Active learning for different number of queries.
- To draw a comparison between traditional ML and Active learning, we compare the accuracy score obtained from only Random Forest model (i.e. traditional ML) and Random Forest + Active Learning for classification of stars in terms of T_{eff}.
- In traditional ML with Random Forest classifier, we allocated 80% data as training set and 20% as test set. With this approach, we gained 82.5 % accuracy score.
- In Active Learning with Random Forest classifier, we allocated 200 samples (out of 59805) as training set and out of rest dataset we allocated 20% as test dataset and rest as unlabelled pool dataset. With this approach, we gained 80 % accuracy score with just 250 queries. Hence, we show that with limited dataset, Active Learning approach gives better result compared to traditional ML.
- We show the change in the accuracy score with number of queries below for illustrating the Active Learning approach quantitatively.

No. of queries	Size of pool of unlabelled data	Accuracy Score (%)
0	47108	76.54
50	47058	78.52
100	47008	78.90
150	46958	79.84
200	46908	79.77
250	46858	80.77