Automated classification of scientific publications linked to GES DISC datasets

NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

Rohan Dayal[1, 2], Irina Gerasimov[1, 2], Armin Mehrabian[1, 2], Jennifer Wei[1], Mohammad Khayat[1, 2], Andrey Savtchenko[1, 2]

[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Background and Motivation

- The data collections archived and distributed by the GES DISC NASA data center are widely utilized for various Earth Science studies.
- GES DISC collects these publications and provides their citations for the users, so it is helpful to categorize these papers based on how they relate to the datasets they are associated with.
- GES DISC scientists came up with four major categories of publications with regard to how they related to the Earth Science collections.

Category Name	Paper Category Description
Algorithm	• Written by scientists to describe the mathematical basis for the algorithms used for creating datasets. • Describe algorithms that convert sensor observations (radiances, brightness) into physical variables (water vapor, precipitation). • Can be descriptions of a modeling or assimilation approach. Can help dataset users better understand the mathematics involved in dataset production, dataset provenance, structure, parameters, limitations, etc.
Validation	• Describe validation of the datasets done by comparison between datasets coming from different sources, e.g. platforms/instruments and models, or the ground validation of Earth observational data. • By science teams or the teams producing ground observation data. • Important for dataset creators as well as users to understand the datasets uncertainties, errors, and limitations in applicability.
Application	• Largest category on how dataset was used in real life: analysis of events, phenomena, system modeling, environmental trends, etc. • Created by science teams and researchers who acquire the data • Have largest diversity of publication sources in terms of the location and science impact factor. Can help to evaluate dataset usability for specific studies as well as dataset science impact.
Overview	• General descriptions of missions, sensors, projects, experiments, field campaigns, or assessments, that produced the datasets. • Written by principal investigators • Highly recommended to understand the basics of the project's goals, the observational and mathematical principles employed.

- The process currently requires manual labeling. We seek to create an automated classifier using manually labeled publications and their abstracts as training data for supervised machine learning algorithms to make category predictions.

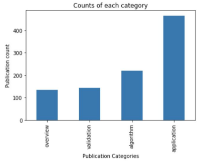
Data Pre-Processing

- Data pre-processing steps are then applied to clean the abstracts
  - Lowercasing all text so that same word uppercase and lowercase counts as one
  - Removing all punctuation marks since they do not provide much knowledge
  - Removing all stop words to make the data less noisy
  - Lemmatizing the text - Reverting all words to their base form

Feature Engineering

- The text is converted to a numerical format through the TF-IDF weighting factor
- TF-IDF is the frequency of the word in the current document multiplied by the inverse of the frequency across documents of the word
- Each individual word in an abstract has its TF-IDF number calculated
  - Combination of all these weights across an abstract creates a vector representation for the abstract
- The TF-IDF vectors will be used to train the machine learning algorithms employed

Data Distribution and Machine Learning



- Publication category distribution:
  - application: 466 publications
  - algorithm: 219 publications
  - validation: 143 publications
  - overview: 135 publications
- Distribution is extremely biased to "application"
  - Population is also extremely biased this way
- Models run with imbalanced distribution may not perform well
  - Two approaches to resolve: under sampling, over sampling
- Two algorithms of interest are Random Forest and Naive Bayes
- Naïve Bayes algorithm - based off of Bayes Theorem
  - Utilize each word's TF-IDF weight to calculate the probability that a word belongs to a category
  - Documents are classified as belonging to the category for which the product of probability for each individual word belonging to that category is the highest
    - Ex: P(App) = P(App|word1)*P(word1) * P(App|word2)*P(word2) * …
    - P(Val) = P(Val|word1) *P(word1) * P(Val|word2)*P(word2) * …
    - If P(Val)>P(App), output P(Val) for the document classification

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events
$P(A|B)$ = probability of A given B is true
$P(B|A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B

- Random Forest algorithm is a tree ensemble algorithm
  - A single decision tree can be used for modeling simple scenarios
  - Random Forest algorithm trains multiple decision trees fit for different segments of the training sample
  - Ensemble algorithm works in that the result which is output from the algorithm as a whole is the most popular output from amongst all of the trees in the "forest"

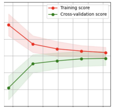
Initial Results

The testing accuracies of the algorithms with different distributions are as follows:

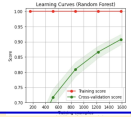
	○ Original Sample	Under Sampling	Over Sampling
○ Naive Bayes:	0.648	0.481	0.636
○ Random Forest:	0.634	0.494	0.914

- Under sampling approach was the poorest performer. Likely as a result of the loss of data under sampling requires by reducing all categories to the minority category size
- Over sampling with Random Forest appears promising, but further analysis demonstrates that the model was overfit to the dataset
- A learning curve should show converging training and cross validation curves. The learning curve for RF lacks this convergence and maintains a gap between CV and training, along with constant training scores, indicating overfitting.

Ideal LC:

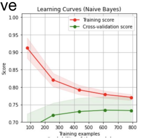


RF LC:



Binary Classifier and Results

- While initial results not ideal, another approach that still could substantially reduce manual efforts is binary classification between "application" and non-"application" publications
  - "Application" categories make up the vast majority of the population of publications which use GES DISC collections
- The sample we have contains 466 applications, and 497 non-applications, so with marginal under sampling, can have balanced data
- Running the same multinomial Naïve Bayes algorithm as a binary classifier between these two categories produced promising results
- Test accuracy score was found to be 0.771, along with similar recall and precision values in the 0.7-0.8 range
- The learning curve for this Naïve Bayes model is much closer to the ideal learning curve model, and shows that overfitting did not occur



Next Steps

- We will apply the algorithm on the newer publications' citations, as well as citations related to other GES DISC data
- We will integrate this classifier directly into the Zotero system so that this process can automatically occur. We will then develop a workflow for the remaining final labeling stages