Development of a multiclass Naïve Bayes classifier for the retrieval of butterfly trait information from published field guides

Vaughn Shirey
Department of Biology, Georgetown University; Washington, DC, USA

Abstract

Biodiversity literature is a rich resource of knowledge regarding organisms and habitats. Unfortunately, few natural language processing (NLP) tools have been developed to work with many diverse corpora. Field guides represent one domain of literature from which information about species associations, habitat preferences, geographic locations, life cycles, and other valuable attribute data can be extracted. These traits can then be operated upon in a statistical framework to inform biological research communities. Here, I present the development of a multiclass Naïve Bayes (NB) model to discern at the sentence level, which class, if any, of trait data are being presented by the text. Using a variety of features to train the model, the highest performance score obtained was 71%, highlighting the need for further research on the topic.

1 Introduction

The retrieval of critical biodiversity information from literature resources remains a constant challenge for the biology research community, due in large part to a lack of resident expertise in NLP, transcriptions, and translations of primary literature resources (Thessen *et al.* 2012, 2014). Despite these challenges, large online repositories of literature data are already publicly available, and new literature resources are being aggregated daily (Gwinn and Rinaldo 2009).

Field guides on popular organisms (such as birds and butterflies) represent one source of potential knowledge for biological traits (physical, ecological, and behavioral attributes of organisms) that can be operated upon by researchers to produce meaningful models of the natural world.

ButterflyNet (https://www.butterflynet.org/) is an internationally collaborative project that aims to create a complete evolutionary history, range maps, and species trait database for the world's butterfly species (ca. 18,000). Manual extraction of butterfly traits by people is time-consuming and thus, any level of automation through NLP tools will increase the ability to obtain this critical information.

The objective of this work is to develop a multiclass, Naïve Bayes (NB) classifier to classify sentences extracted via OCR into several key trait categories: morphology (discussing some physical attribute of a species), distribution (discussing some spatial attribute of a species), life-history (discussing a particular life stage such as butterfly or caterpillar and behavior), hostplant (describing what a caterpillar eats), and non-target (miscellaneous sentences that do not reflect target trait data).

2 Related Work

Information retrieval (IR) has been a subject of interest within the biodiversity sciences for some time. Related work has included the development of online IR tools that utilize dictionaries and n-gram models to extract information from uploaded PDFS (Muñoz *et al.* 2019). In tandem with this, the development of corpora for named-entity recognition (NER) tasks related to IR has also been of interest (Nguyen *et al.* 2019).

3 Methodology

The dataset consisted of 698 sentences that were labeled by two individuals using labels specified in the introduction for trait data. The sentences came from one family in one text, "The Butterflies of Cascadia: A field guide to all the species of Washington, Oregon, and surrounding territories" by Robert Pyle.

Critical to the development of a high-performing classifier is the selection of features to use in training and validating a model. I employed a scheme of three selected features for the initial exploration of utilizing an NB classifier for labeling sentences into trait categories. NB is commonly used for these tasks as it is a relatively simple and rapidly executing algorithm based on Baye's Theorum and illustrated in Figure 1.

$$pMax(class|feature) = \frac{P(feature|class)P(class)}{P(feature)}$$

Figure 1. An adaptation of Baye's Theorum as applied to the Naïve Bayes classification scheme. Where the probability of a certain class is maximized by the joint probability of features of that class conditioned on the prior probability of the class occurring within the dataset.

First, a model using common single word tokens, and secondly a model trained using common 2-gram features. Two matrices were created for each feature set, one with the raw count of features, and a second, binary feature matrix indicating the presence or absence of the given feature.

Using Python and the NLTK and SciKitLearn libraries, I employed a GaussianNB and BernoulliNB model on the normal and binary feature matrices respectively for each feature set. I compared the results of these models to a zero-rule baseline which took the most commonly occurring label and applied it to all sentences within the validation dataset in order to generate the baseline score. 90% of the data was used in training while the remaining 10% was used for validation in all scenarios. The training/validation selection was kept consistent across all models as well.

In order to assess how the training/validation dataset size impacted scores, I ran the models 6 times, increasing the number of records utilized in training and validation over a sequence starting from 100, to 200, 300, 400, 500 and finally 600 sentences.

4 Results

The results of running all models using both the normal and binary feature matrices are presented in Table 1. From these results, it is clear that the binary feature matrix using single function word features performed best for classifying these data.

	Gaussian	Bernoulli
Single Word	0.47	0.71
2-gram	0.41	0.42
Baseline	0.42	

Our model paradigm encountered an error in the SciKitLearn library when utilizing a 2-gram feature matrix. This was caused by a low (essential no) difference between classifications and their corresponding feature matrices, effectively making certain classes indistinguishable via Naïve Bayes.

Figure 2 represents the model scores for all models and feature sets with an increasing corpus size. Overall it was difficult to detect any strong directional signal with corpus size and model performance, which may indicate that the selection of functions words is not appropriate or expansive enough to encompass all classes within a restricted dataset.

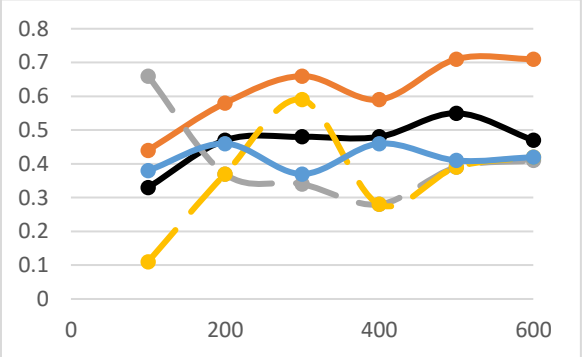

Figure 2. Model performance with increase corpus size. (Orange = BernoulliNB Single, Blue = GaussianNB Single, Yellow = BernoulliNB 2-gram, Grey = GaussianNB 2-gram, Black = Baseline).

5 Discussion

Our best performing model was the BernoulliNB with single word features. This was not surprising as single words are often highly associated with specific classes. For example, habitat descriptors such as “forest” or “open” are largely never applied to morphology, hostplants, or life-stages, making them highly associated with the distributions class. This same logic likely follows for other single word features including colors, metrics, and behaviors (“eats”, “feeds”, etc.).

Although this model was the top performer, missing or misclassifying nearly 30% of the data is not acceptable for the purposes of IR. This is because the labelled sentences to not represent the final desired product for this IR task and will need to be atomized or conformed to a controlled vocabulary in order to be operable by statistical tests. Mislabeling will increase the human workload on the backend.

Several pathways exist to remedy this lower performing score. For example, adopting a feature selection paradigm passed on named-entity recognition algorithms already developed for biodiversity data may lead to a smarter selection of function words based on entities that are already characterized as having some importance to biodiversity.

Another potential avenue to explore would be the use of punctuation and other non-word markings that may denote structure. These would be particularly useful in the case of the hostplant class since these data are often presented in lists delimited by some character such as a comma or semicolon.

Regardless of the chosen avenues for increasing the model performance scores, it seems like for this dataset, utilizing a binary feature matrix with BernoulliNB shows the greatest promise. Future research is needed to tackle the issue of feature selection for this multiclass problem. Including additional books to test this model will also be of critical importance as regional variation among butterfly trait descriptors may be present. Additionally, alternative spellings in American and British English will need to be accounted for. Despite these challenges, the promise of IR using NLP tasks for butterfly traits remains strong.

Acknowledgments

I would like to acknowledge my undergraduate research assistant Erin Leeds for assistance in labelling the original data.

References

Anne E. Thessen and Cynthia S. Parr. 2014. Knowledge Extraction and Semantic Annotation of Text from the Encylopedia of Life. PLoS One 9 (3): e89550.

Anne E. Thessen, Hong Cui, and Dmitry Mozzherin. 2012. Applications of Natural Language Processing in Biodiversity Science. Advances in Bioinformatics 2012: 391574.

Gabriel Muñoz, W. Daniel Kissling, and E. Emiel van Loon. 2019. Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature. Biodiversity Data Journal (7): e28737.

Maria A. Mora and José E. Araya. 2018. Semi-automatic Extraction of Plants Morphological Characters from Taxonomic Descriptions Written in Spanish. Biodiversity Data Journal (6): e21282.

Nancy E. Gwinn and Constance Rinaldo. 2009. The Biodiversity Heritage Library: sharing biodiversity literature with the world. IFLA Journal 35 (1).

Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. 2019. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodiversity Data Journal (7): e29626.