Investigating the Efficacy of Data Mining Techniques for Crowdsourced Biological Datasets

Ilona Regan
Bioinformatics and Computational Biology Program
Worcester Polytechnic Institute
Worcester, MA, USA
imregan@wpi.edu

Dany Alkurdi
Department of Near Eastern Studies
Princeton University
Princeton, NJ, USA
dalkurdi@princeton.edu

Sarun Paisarnsrisomsuk
Department of Computer Science
Worcester Polytechnic Institute
Worcester, MA, USA
spaisarnsrisomsu@wpi.edu

Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute
Worcester, MA, USA
ruiz@wpi.edu

Elizabeth F. Ryder
Department of Biology and Biotechnology
Worcester Polytechnic Institute
Worcester, MA, USA
ryder@wpi.edu

Rob Gegear
Department of Biology
University of Massachusetts Dartmouth
Dartmouth, MA, USA
rgegear@umassd.edu

***Abstract*— The rapid worldwide decline of wild pollinators over recent years poses a significant environmental threat due to the critical keystone role that pollinators play in terrestrial ecosystems. In order to gain insight into the major anthropogenic factors causing these declines, researchers have collected large amounts of ecological data. Yet, they often lack the computational tools needed to analyze the information contained in such datasets. We investigate various data analysis techniques for the Beecology Project, which is a citizen science based effort to rapidly collect data on foraging habits of bumblebee species native to Massachusetts. Different data mining approaches were explored, including association analysis, trend analysis, classification, regression, and clustering. It was found that different techniques were more suitable depending on the biological research question. Future work will focus on making tools utilizing these approaches available online through the Beecology website, where they can be used by the public to determine how best to protect our native pollinators and the diverse ecosystems they support.**

Keywords—biological data mining, environmental data analysis, machine learning

I. INTRODUCTION

In an increasingly data-driven world, more problems are being addressed using large datasets. As database size and complexity increase, it is essential to have efficient techniques to analyze data. Data mining focuses on finding patterns in data through a variety of methods[1]. The purpose of this paper is to pinpoint specific data mining techniques that are well-suited to support biologists, environmental scientists, and the general public in exploring and analyzing biological data and investigating biological hypotheses in a data-driven manner. This investigation is conducted in the context of an environmental citizen science project, the Beecology Project.

The Beecology Project is a project about wild bumblebee pollinators native to Massachusetts [2]. Exploring the Beecology database helps researchers to determine which pattern discovery techniques will be most effective to analyze various biological hypotheses. Furthermore, the Beecology Project is part of a larger effort, the Bio-CS Bridge Project [3], aimed at interconnecting the teaching of biology and computer science in high school curricula. It provides high school teachers and students with a biologically-motivated, computational framework and curriculum. Data analysis techniques identified in this paper as well-suited to investigate biological hypotheses will be made available to the general public via the Beecology and the Bio-CS Bridge websites. It is expected that these pattern discovery techniques will be useful not only for testing hypotheses about bumblebees using the Beecology database, but for testing hypotheses related to other biological datasets as well.

Funding provided by the National Science Foundation

II. BACKGROUND

A. The Beecology and the Bio-CS Bridge Projects

Pollinator decline is a pressing issue. Wild pollinators are keystone species that promote ecosystem biodiversity and resilience. The Beecology Project uses a citizen science approach to collect ecological data on bumblebee pollinators native to Massachusetts [2]. It focuses on data collection and on the visualization and analysis of the pollinator data stored in its database. This Beecology database is described in Section III.A.

The Beecology website hosts a web app for data collection and a data visualization tool, both of which are connected to the Beecology database [4]. The website also contains an agent-based simulation that explores how bee and plant populations are affected by different environmental conditions. There is currently no data analysis tool available on the website.

The Bio-CS Bridge is a related project at WPI that aims at creating a synergistic approach to the teaching of biology and computer science in high school [3]. It utilizes the Beecology Project, to illustrate how the investigation of an environmental problem, that of pollinator decline, can be addressed using computational thinking and approaches. The current work is the first step towards the creation of a data analysis tool for the Beecology Project that will allow students, teachers, researchers, and citizen scientists to handle and analyze the Beecology dataset to determine the most effective strategies to manage and preserve our native ecological systems.

B. Data Analysis Approaches

Data mining aims at extracting patterns in data. There is a variety of data mining approaches and techniques. Four data mining approaches are described below.

Association Analysis asks whether there are attributes in the dataset that are correlated to one another. This technique involves measuring the strength of the relationship by comparing multiple attributes. **Trend Analysis** analyzes the changes in an attribute or combinations of attributes over time. **Prediction** includes multiple techniques that can be used to generate predictions. Classification uses given attributes, or classifiers, to predict the class. Decision trees are a supervised

2019 IEEE MIT Undergraduate Research Technology Conference (URTC) | 978-1-7281-5818-1/19/$31.00 ©2019 IEEE | DOI: 10.1109/URTC49097.2019.9660523

machine learning classification technique. Regression is used to measure and predict the associative relationship of multiple attributes. **Clustering** is used to examine similarities between entries. K-means clustering uses Euclidean distance to compute clusters that minimize the distance within the clusters. Hierarchical clustering minimizes the distance between clusters as well as within clusters..

III. Methodology

A. Beecological Database

The Beecology dataset, containing nominal and numeric attributes, acted as a model in this project to examine the viability of various data analysis techniques for datasets with similar data analysis goals. Each entry in the Beecology database represents an observation of a bumblebee foraging on a flower. Attributes for each entry consist of:

- Bee information including species (*B. affinis, B. bimaculatus, B. borealis, B. fervidus, B. griseocollis, B. impatiens, B. pensylvanicus, B. perplexus, B. ternarius, B. terricola* and *B. vagans*); gender (female, male, worker, queen); and foraging behavior (either collecting nectar or collecting pollen).
- Flower information about the flower where the bee was observed, including species (193 different species), color, bloom time, and tube shape (closed tube, long tube, open tube, short/no tube, and tube with spur).
- Date and location (longitude, latitude, and elevation) of the observation.

There are over 24,000 entries in the database and more than 95% of entries come from the New England area. All entries correspond to museum bee records (starting from the early 1900's), field data from the Gegear laboratory (starting in 2013), or entries submitted through the web app (starting in 2018). The data entries from the web app and field data contain more attributes, such as flower characteristics, than the data entries from the museum records.

B. Methods

The following data mining approaches were used for exploratory analysis of the Beecology database. **Association Analysis**: A contingency table was created between two nominal attributes from which relative frequencies were calculated for each bee species. Due to large differences in sample sizes, relative rather than absolute frequencies of various attributes were calculated for each bee species. **Trend Analysis**: Trend lines were explored through graphing with matplotlib. **Prediction**: Decision trees were created with the scikit-learn library and were used to further investigate relationships that were discovered using association analysis. Regression, both linear and non-linear, was performed through matplotlib as well as numpy libraries to further investigate relationships between attributes. **Clustering**: K-means clustering was explored on a set of coordinate points using the scikit-learn library using Euclidean distance metric and randomized initial centers. Hierarchical clustering was performed on nominal attributes that were encoded into numeric values using a SciPy dendrogram algorithm that calculated Ward distance metric.

IV. Results

A. Association Analysis

Association analyses can be used to investigate hypotheses about relationships between attributes in a dataset. For example, the graphs shown in Fig. 1 support the hypothesis that individual bumblebee species prefer either high or low elevation and that these preferences can change over time. For example, *B. bimaculatus* and *B. borealis* have been observed in larger numbers at high elevation in recent years, and *B. perplexus* and *B. vagans* numbers at higher elevations have also slightly increased. This result may suggest that bees are able to change their preferences for elevation and adapt to new locations as the environment changes.

This type of association analysis can test hypotheses about the relationship of one attribute with another, including how bee species' preferences for certain flowers have changed. Association analyses provide insight into how attributes are related to one another and can reveal patterns in the data.

Fig. 1. Bar graphs comparing relative elevation freuqncies for several bumblebee species. Low elevation <1000 ft above sea level; high elevation >1000 ft above sea level.

B. Trend Analysis

Bumblebee species vary in the length of their tongues; they are categorized in the dataset as short-, medium-, or long-tongued. Tongue length is one factor that determines the species of flowers on which bumblebees forage. To investigate whether short- and long-tongued bumblebee species declined over the past two decades, data entries were combined based on the tongue length category of their species. Relative frequencies of each category were then plotted against time. In Fig. 2, it can be seen that short- and long-tongued bee species have decreased in abundance over time, while medium-tongued bee species have increased in abundance. Medium-tongued bees, which are more versatile in flower selection than their counterparts, have grown to account for over 60% of the entries for both the 2000s and 2010s, whereas they only accounted for 10-30% in prior decades. This suggests a recent change in environment that favors the versatility of this tongue length.

978-1-7281-5818-1/19/$31.00 ©2019

Overall, trend analysis can be used to examine hypotheses about attributes with respect to time and location. It is also possible to filter the data beforehand to compare trends between different groups. It is important to note that the trend lines themselves do not measure the strength of the relationship and only show how it changes over time.

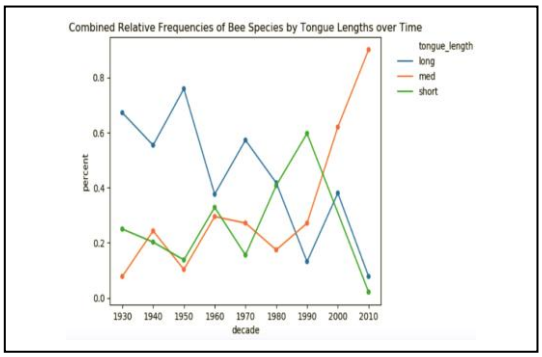

Fig. 2. Plot of relative abundance of bees based on tongue length against time.

C. Prediction

a) Classification: Fig. 3 shows a decision tree that predicts the shape of the flower that a bee would visit based on just the bee's tongue length. For example, if its tongue length is short, the decision tree would predict that the bee will prefer flowers that have a short tube or no tube at all. Classification using decision trees, a supervised machine learning technique, can be used for prediction of essentially any attribute in the dataset based on combinations of other attributes. It works by grouping the data at each level so that the homogeneity of the groups is the greatest, using entropy as the criterion. In order to create these decision trees in Python, any nominal data needed to be one-encoded. This process creates an array that denotes which nominal value is represented without assigning numerical ranking, to account for the inherent lack of ordinality in nominal data.

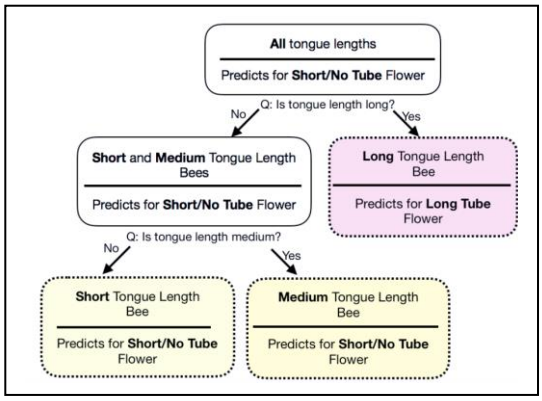

Fig. 3. Decision tree predicting flower shape from tongue length using entropy as the criterion.

Ultimately, decision trees as a technique were most useful for examining relationships in the data and discovering which attributes most significantly contributed to the prediction of an attribute and those which had a lesser effect.

b) Regression: Fig.4 analyzes the abundance of *B. vagans* over months of the year. The equations of the fitted line are different for the 1980s and the 2000s. A second degree polynomial is the best fit for both graphs rather than a linear regression, since bee populations typically follow a curve, with lower abundance at the beginning of the season (April-May) that rises to a peak mid-season (July-Aug.) and then decreases (Sept.-Oct.). Considering only the scatter plot, the month with the most average observations changes from August in the 1980s to July in the 2000s. It is important to note that the number of observations per year increased over time as well due to more concerted data collection efforts. Regression is useful to test hypotheses about changes in bee abundance over time. Comparison of the fitted equations allows us to analyze changes in these curves over decades.

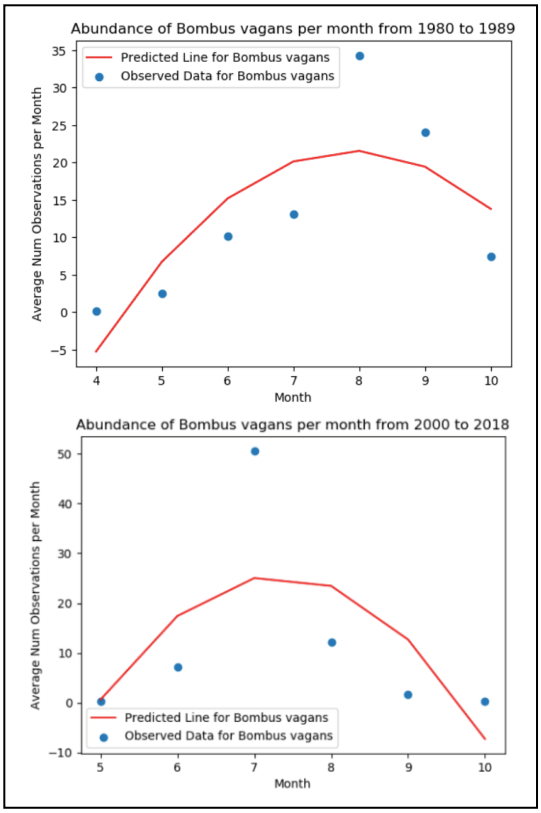

Fig. 4. Non-linear second-degree polynomial regression on abundance of *B. vagans* bumblebees per month in different time frames.

D. Clustering

a) K-Means Clustering: Clustering analysis was used to determine whether sites were surveyed consistently over time. In Fig. 5, each dot is an observation at the given location, each shade of blue represents a cluster, and each salmon-colored dot is the center of a cluster. Although at first glance one may interpret changes in cluster location over time as changes in bee location preferences (i.e., small migrations), given the nature of the Beecology database, these changes in cluster locations are related to changes in locations chosen for data collection over time. Cluster overlapping throughout the years means that the intersecting area was well surveyed in historical records and in current entries from the web app. If there appears to be no overlap over the years, this means the site was not surveyed throughout the years consistently. A lack of data in historical records cannot be recovered. However, if there is a lack of data from 2000 to 2019, then the gaps can be filled in by obtaining more data from that location in upcoming years. This would allow analysis to be conducted in a site-specific manner at locations that are consistently surveyed.

978-1-7281-5818-1/19/$31.00 ©2019

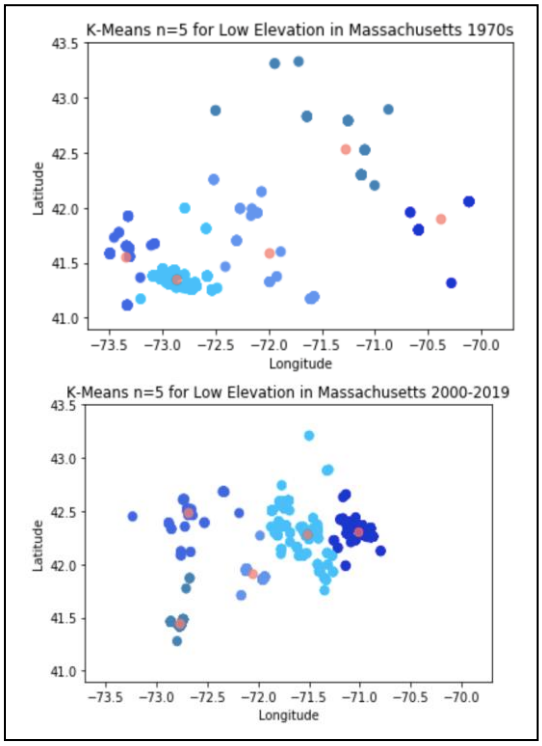

Fig. 5. K- Means clustering based on location from various years to find survey sites in historical and current records.

b) Hierarchical Clustering: Fig. 6 shows an example of hierarchical clustering, visualized as a dendrogram through the use of Scipy. The clusters are represented by different colors. The cophenet value of Fig. 6 is 0.733. This represents the strength of the hierarchical clustering by measuring the distance between clusters.

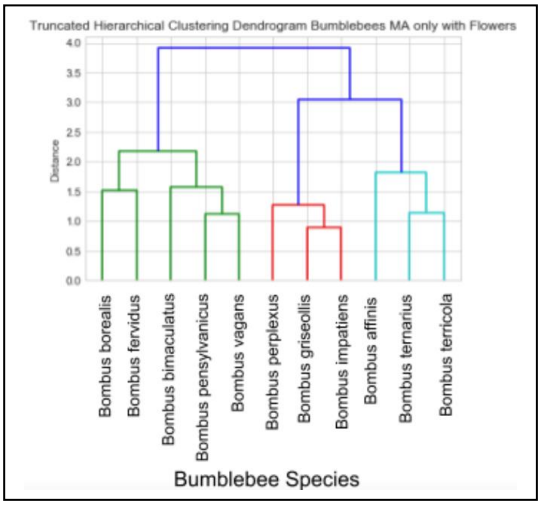

Fig. 6. Hierarchical Clustering for all Bumblebee species in Massachusetts from 2015 to 2019.

In Fig. 6, the clusters are calculated from the following attributes for each species: tongue length, abdomen coloration, relative frequencies at high and low elevation, and relative frequencies at five locations within the Massachusetts region. Since the SciPy hierarchical clustering function requires all values to be numeric, relative frequencies were calculated for elevation and location. Tongue length and abdomen coloration were encoded into numeric values based on a ranked scale. This clustering shows which bumblebee species are most similar based on the given attributes. Consistent with expectations based on domain knowledge, *B. borealis, B. fervidus,* and *B. pensylvanicus* cluster together.

V. Conclusion

From the experiments performed, it is shown that different data mining techniques may be utilized depending on the nature of the dataset, the attributes, or the goal of analysis. Different pre-processing techniques may be used for attributes that are nominal or numeric. If the attribute is nominal, it can be transformed into a numeric value through encoding or calculating relative frequencies. Ultimately, different data analysis techniques can be used depending on the type of hypothesis posed by biologists. The data mining techniques used in this paper can be similarly used on other large biological datasets which are similar in nature to the Beecology dataset.

VI. Future Work

Future work includes implementing these data analysis techniques online as part of the Bio-CS Bridge and Beecology projects. The data analysis tools will be incorporated into the Beecology website, leveraging the website's modular platform. Users will also be enabled to download their own customized Beecology data subsets to examine the data and utilize the analysis tools.

Acknowledgment

We thank the National Science Foundation for funding this project. Authors Regan and Alkurdi are supported under the NSF#1852498 grant "REU WPI SITE: Data Science Research for Healthy Communities in the Digital Age" and authors Paisarnsrisomsuk, Ruiz, Ryder, and Gegear are supported under the NSF#1742446 grant "Building Educational Bridges between Computer Science and Biology through Transdisciplinary Teamwork and Modular Curriculum Design." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank the members of the Beecology and Bio-CS Bridge projects for their input.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. 1996. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.

[2] Beecology Project. (n.d.). Retrieved from https://beecology.wpi.edu

[3] Bio-CS Bridge. (n.d.). Retrieved from https://biocsbridge.wpi.edu

[4] Wang, Xiaojun. 2018. Bringing 'Bee-cological' Data to Life through a Relational Database and an Interactive Visualization Tool. Master's Thesis, Bioinformatics and Computational Biology Program. Worcester Polytechnic Institute. 2018.

978-1-7281-5818-1/19/$31.00 ©2019