EcoSurvey Analysis Proposal Paper

Conor McNamara
University of Colorado Boulder
College of Engineering and Applied Science
Inquiry Hub
Conor.R.Mcnamara@colorado.edu

Shawn O'Grady
University of Colorado Boulder
College of Engineering and Applied Science
Shawn.Ogrady@colorado.edu

ABSTRACT

This paper will outline our groups proposed work for the remainder of the semester to analyze EcoSurvey models. In this outline we cover what knowledge can be mined from the EcoSurvey data, how data mining has been applied to this field previously, how we plan on working with our data set, and our expected steps required for completing this.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; **Model verification and validation**; *Supervised learning by classification*; *Classification and regression trees*; *Bayesian network models*; Ranking;

KEYWORDS

EcoSurvey, prediction algorithm, WEKA

1 MOTIVATION

For this project, we plan on applying data mining techniques to an EcoSurvey model. EcoSurvey is an application that students use to build digital models of their local ecosystems. The data set that we plan on analyzing contains the data about the local ecosystem from the model, as well as information about how the students interacted with the application.

From our analysis, we will firstly try to create an algorithm which takes the object and subject features of two organisms, and predicts how the two are related. Second, we will work from this algorithm to determine which features of the data set are the best predictors of relationships between organisms.

This analysis will have many potential uses, primarily to help develop future models for a more student-centered learning environment.

2 LITERATURE SURVEY

As our work is examines student centered learning, we have examined many similar topics. The most significant prior work that we looked at was self-determination theory, which examines peoples motivation behind their decisions without external influence. This is relevant to our research as we are determining how users interact with our application and exploring what motivates them to make their decisions when creating the model.

In addition to the self-determination theory, our work revolves around digital modeling, which has been extensively explored by the Georgia College of Tech Computing. Digitals models assist us in visualizing relationships between data as well as organizing the data in a way that makes the data easier to both mine and interpret. With the use of digital models and the theories behind student centered learning, we hope to create an interpretation of our data that will help generated a more student centered curriculum.

3 PROPOSED WORK

Thankfully, the entirety of our data set has been generated from the EcoSurvey application, so no additional data collection is required. We plan on using the open source data analysis program WEKA for the majority of our proposed knowledge discovery process, and as such we will need to convert the data we wish to analyze in to a .arff file. The data will have to retrieved from a MySQL database, converted to a .csv, then a .arff file using python scripts.

Preprocessing of the data will be handled using WEKA, as stated before. While we do not have any missing data, we may find during the course of our analysis that outliers or inconsistencies greatly interfere with our process of finding meaningful patterns, in which case we will need to employ data cleaning methods. Additionally, since our data is currently separated in to various tables within the MySQL database, we will need to ensure that proper data integration conventions are followed when we use our full testing dataset.

After the data has been properly preprocessed, we will continue using WEKA to discover the most predictive features of our data set, which will be derived from a training data set. The Attribute-Selection feature of WEKA will allow us to use our training data to determine which attributes we will use to perform our classification, and how to rank them. As one of our project goals is to determine which features of the data set are the most predictive of the relationship between two organisms, this WEKA feature will be very helpful.

In order to get a full prediction algorithm from our data set, we will need to figure out which classification method to use. We will once again primarily use WEKA for this portion of the project, since WEKA has a large number of classifiers built in. For this project,

we foresee primarily using Naive Bayes, Random Forest,Random Tree, and Logistic Regression classifiers.

Finally, we will perform error analysis on the prediction algorithm that we developed by observing the confusion (error) matrix as outputted by WEKA. We will need to keep documentation on how various algorithms performed, and reiterate the attribute selection and classifier selection process as necessary to achieve a level of accuracy that would be deemed acceptable.

Our proposed work differs from the prior work done by others described above in that combines digital modeling and self-motivation theory.

4 DATA SET

As previously stated, our data set comes from the EcoSurvey application, and contains data on both the organisms in the local ecosystem and the students interactions with the application. The data set is stored in a MySQL database, and has been organized in to forty different tables. All tables combined, we have 142354 rows and 433 columns, for a total of 61,639,282 data points. This should provide more than enough information to generate a prediction algorithm.

This is actually the second time analysis has been applied to EcoSurvey data, the first was performed by author Conor McNamara as a part of separate research. The data set that we are planning on analyzing for this project represents a second version, which was constructed from the results of the analysis on the first.

In addition, the data set requires authorization and training from the Collaborative Institutional Training Initiative (CITI) Program to access.

5 EVALUATION METHODS

As stated above, due to this being the second iteration of analysis on a EcoSurvey model, we will be able to compare and contrast our results from this semester with those author Conor McNamara found on version one of the data. Additionally, we will use the confusion matrices generated by WEKA to help in verification of our predictive algorithms.

6 TOOLS

In order to analyze the data from the EcoSurvey model, we will utilize the following programs:

(1) GitHub: We will be using GitHub to manage our code for the project, as well as any required documentation and write ups.
(2) MySQL: All of our data is currently located in a MySQL database, which we will need to continually access throughout the course of this project
(3) Microsoft Excel: In order to better visualize our data, we will use Microsoft Excel's various plotting features
(4) Python: We will be using Python scripts to take the data from the MySQL database, convert it to a .csv file, then convert it to .arff file
(5) Sublime: Clearly, to create Python scripts, we will need to use a text editor such as Sublime
(6) WEKA: The majority of the actual data analysis portion of our project will be performed using WEKA. WEKA allows us to use various attribute selection and classification schemes in order to develop the optimal prediction algorithm for our data set. WEKA will be an extremely valuable tool for this project, particularly for comparing several prediction algorithms.
(7) Microsoft Word and LaTeX : To formally present our findings and describe our methods, we will need to make use of a word processor or typesetting system.

7 MILESTONES

Going forward this semester, in order to fully realize a valid prediction algorithm, we will need to make sure to keep up with the following tasks:

Projected dates	Description
1/10/17-2/1/17	Visualize second data set as well as determine necessary features to be selected
2/2/17- 3/20/17	Write python code to selected the necessary features from the version two data set as well as format properly
3/21/17-4/10/17	Perform machine learning analysis of data from version two
4/11/17-5/5/17	Compare and contrast findings and use them to develop conclusions

8 SUMMARY OF PEER REVIEW SESSION

While there were not many comments from our peers after presenting our proposed project, we were advised to survey more previous projects in the same field we are interested in, which we included in the Literature Survey section above.