EcoSurvey Analysis Progress Report

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, Scientific Modeling, Collaborative Modeling, Teacher Differences, Classification

ACM Reference format:
Conor McNamara and Shawn O'Grady. 2017. EcoSurvey Analysis Progress Report. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

Our research focuses on issues related to student scientific modeling in a student-centered, inquiry-based curriculum. Modeling practices are essential to student learning because they reveal new learning analytic methods to better characterize student scientific modeling practices and to examine classroom level differences. We will be use computational techniques to study students' modeling practices at a detailed level, at unprecedented scale, and in new ways. Our computational techniques and interpretation of their results will be guided by theories and prior research on effective modeling practices arising from learning sciences research.

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. Figure 4 shows an example of an organism card for inputing data in EcoSurvey version 1. Figure 5 shows an example of of an organism card for version 2, which has normalized certain data fields, such as role of organism (either predator, prey mutually beneficial, etc.) relationship, as opposed to being an open field. 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.

Scientific models are tools for explanation and prediction. A complete scientific explanation should ?explain observed relationships between variables and describe the mechanisms that support cause and effect inferences about them? [23]. Thus, to support student explanations, a scientific model of a phenomenon should include important components (?vari- ables?), their interactions (?relationships?), and define the mechanisms involved. When modeling an ecosystem, these correspond to the organisms in the ecosystem (animals, plants, insects, fungi, etc), how these organisms interact with each other and the environment (predator, prey, producer, de- composer, etc), and the involved processes (abiotic, biotic, etc). Professional biologists use this information to measure the biodiversity of an ecosystem in terms of species richness, evenness, and divergence.

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.

We will also be comparing and contrasting the two different deployments of EcoSurvey to understand how the differences between the two versions influenced the students learning. In order to do this, We will be using data mining and machine learning techniques. By understanding and analyzing student models, we can develop theories in regards to the nature of these students learning and use this to help develop student-centered curriculum.

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

Conference'17, Washington, DC, USA
© 2017 ACM. 978-x-xxxx-xxxx-x/YY/MM...$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

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.

We are also drawing a large part of our research and background from the work previously done by David Quigley, Jonathan Ostwald and Tamara Sumner called "Scientific Modeling: Using learning analytics to examine student practices and classroom variation."

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

The phases of data mining and machine learning consist of exploring the data, data fitting/selecting features and analyzing methods. Once again, it is important to note that these phases are not linear. Exploring the data consists of exploring the data visually. For this research, we use Microsoft Excel to visualize the data and mysql to gather the data in the appropriate manner. The next step is data fitting, where we use Python to write code that selects the correct and relevant data. Finally, we will use Weka to analyze that data and determine the accuracy of the fit. In order to do this, we will use supervised methods for data analysis, namely classification and clustering in Weka. we will then record our findings and proceed as the exact same process for the data of the first version of EcoSurvey. We will then compare our findings for the first version?s data to the findings from the second version?s data and determine the significant differences. We will record these differences and use them to determine our ultimate analysis. This analysis will help improve both the EcoSurvey tool and assist in the development of a foundation for a more student centered curriculum.

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
	Perform error analysis on data set from the first version
	Draw conclusions based on results from error analysis of the the first version
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
	Perform feature selection analysis to determine most predictive features of the data set
4/11/17-5/5/17	Compare and contrast findings and use them to develop conclusions

7.1 What We have Achieved So Far

We have successfully developed a series of queries that have taken all of the data given to us, and have organized it into a CSV file. This has also allowed us to determine what data points were recorded properly and which ones were not. For example, the ?group id? value was not recorded properly in the mysql relational database. We have recorded discoveries like this so that we can use this information on our Weka preprocess phase. Our CSV file has the following data fields for the purposes of our research: subject_time_stamp, subject_name, subject_description, subject_url, subject_lastmod, subject_gid, subject_owner_id, subject_image, subject_lat, subject_longitude, subject_cTmp, subject_roleName, subject_tagName, obj_time_stamp, obj_name, obj_description, obj_url, obj_lastmod, obj_gid, obj_owner_id, obj_image, obj_lat, obj_longitude, obj_cTmp, obj_roleName, obj_tagName, relationTableID, obj_id, relationship_id, subject_id, tmp, relationship_name, relationship_description, relationship_type, relationship_inverse, relationship_inverse_id, subject_group_number, subject_owner_role, subject_owner_user_id. In addition, we used the EcoSurvey application graph functionality to visualize our data and confirm that our CSV was correct.

In addition, we have performed an in-depth error analysis on the EcoSurvey Version One data set. This error analysis will be utilized in our compare and contrast phase as well as determining the relevant machine learning algorithms that we will run on the version two data set with which we are working. The error analysis as well as research on the various types of classification algorithms revealed that the Random Forrest Classification Algorithm not only yielded the best result, but was our choice as the algorithm we wished to use. Our confusion matrix below (Figure 1) displays the results.

Therefore, we will be utilizing the Random Forrest Classification Algorithm when we do our version two analysis.

7.2 What Remains to be done

After confirming the accuracy of our CSV file and visualizing the data, we are currently in the process of developing our python program in order to translate our data file into an ARFF file so that it can be utilized in Weka, where we will preform our preprocess and analysis phase. We are currently in the process of finishing this program and hope to do so by March 15th, 2017. Afterwards, we will begin our preprocessing phase and run our machine learning algorithms on the data set.

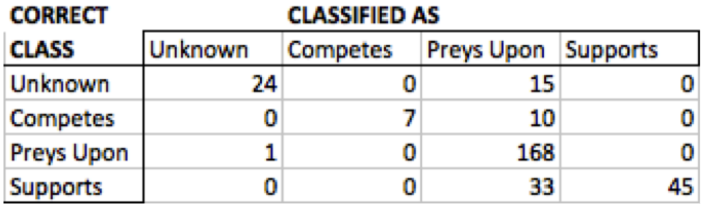
In addition, we will been running analysis to determine which features are the most predictive of relationships given the object and the subject information. We will compare these predictive features to those features that were most predictive in the first version of the data set. We will use Weka's 'Feature Selection' functionality in order to perform this analysis. We have determined that we will be utilizing this functionality buy using three different methods of feature selection and then analyzing the results. The methods that we will be using are Correlation Based Feature Selection, Information Gain Based Feature Selection and Learner Based Feature Selection. After running these algorithms we will determine our most predictive features for the second versions data.

8 RESULTS SO FAR

Our results so far primarily are from the work with our error analysis of the version one data set and with our mysql data base queries.

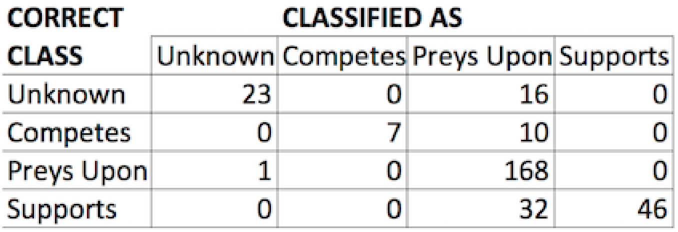
Our error analysis of the version one data set has revealed to us that the algorithm which we would like to use for the purposes of our research is the Random Forrest Classification Algorithm. It has also yielded a confusion matrix revealing 80% accuracy for predicting the relationship between a subject and an object in the version one data set by using the Random Forrest Classification Algorithm.

We have also developed and created a detailed text document outlining the various mysql quires which we developed and implemented in order to convert the data to a CSV file. Through this process, we have learned which data points were not well recorded and can use those results as a part of our preprocess phase.



CORRECT	CLASSIFIED AS			
CLASS	Unknown	Competes	Preys Upon	Supports
Unknown	24	0	15	0
Competes	0	7	10	0
Preys Upon	1	0	168	0
Supports	0	0	33	45

Figure 1: Confusion Matrix with Selected Attributes



CORRECT	CLASSIFIED AS			
CLASS	Unknown	Competes	Preys Upon	Supports
Unknown	23	0	16	0
Competes	0	7	10	0
Preys Upon	1	0	168	0
Supports	0	0	32	46

Figure 2: Confusion Matrix with all Attributes

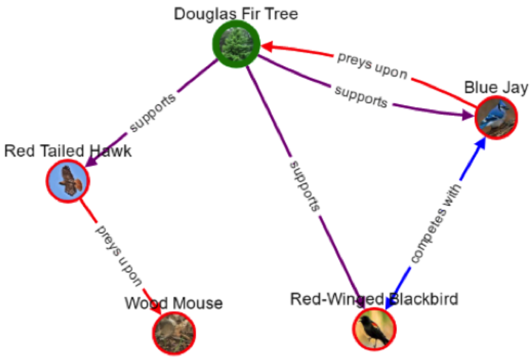
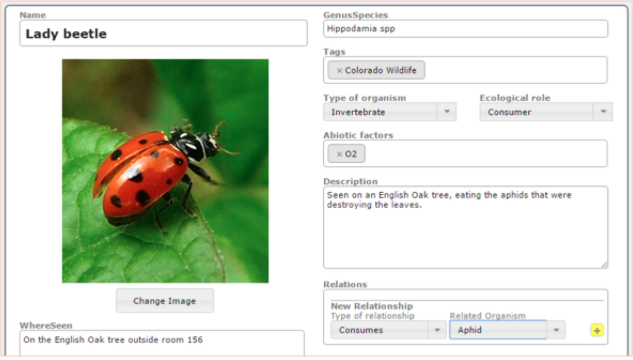


Figure 3: Visualization of Data



Name
Lady beetle

GenusSpecies
Hippodamia spp

Tags
× Colorado Wildlife

Type of organism
Invertebrate

Ecological role
Consumer

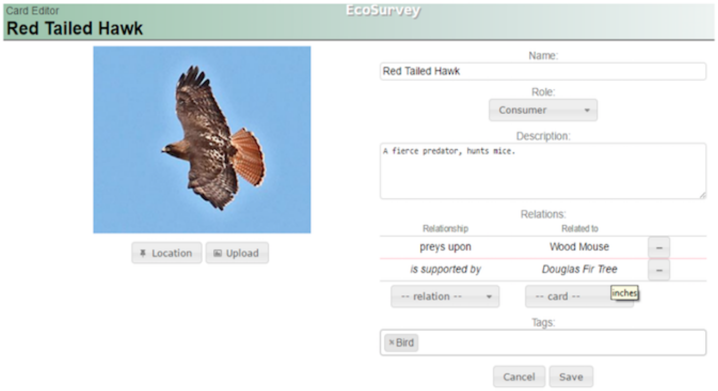
Abiotic factors
× O2

Description
Seen on an English Oak tree, eating the aphids that were destroying the leaves.

Relations
New Relationship
Type of relationship
Consumes
Related Organism
Aphid

WhereSeen
On the English Oak tree outside room 156

Figure 4: Organism Card Example



Card Editor
Red Tailed Hawk

Name:
Red Tailed Hawk

Role:
Consumer

Description:
A fierce predator, hunts mice.

Relations:

Relationship	Related to
preys upon	Wood Mouse
is supported by	Douglas Fir Tree

-- relation -- -- card --

Tags:
× Bird

Figure 5: Organism Card for Version 2 of EcoSurvey

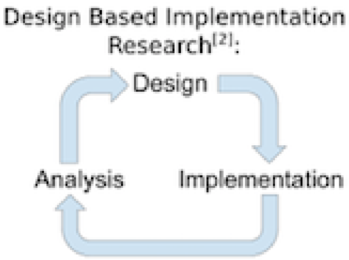


Figure 6: Design Based Implementation Research