

EcoSurvey Analysis

Conor McNamara
Shawn O'Grady

Project Description

Total Data Points: 61,639,282 (Rows: 142354 Columns: 433)

Problem Space: EcoSurvey is an application that students use to build digital models of their local ecosystems. The data gathered from these models as well as the students interactions with the application allows us to observe student learning habits in order to develop student centered curriculum. An effective way to do this is to use data mining and machine learning techniques to gather information from the data set.



Our Questions

- 1) Using object and subject features, can we predict the relationship between the two organisms?
- 2) Which features from the data set are the most predictive of relationships?



Prior Work

- 2 Versions of EcoSurvey
 - Conor McNamara ran analysis on Version 1 already
- All of our data is from Version 2
 - After Version 1 analysis was completed, it was implemented to improve the tool with Version 2
- Data has been stored in a mysql database and organized into various (40) tables

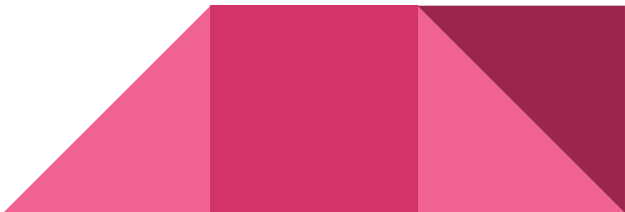


Data Sets

- 1) EcoSurvey Base (mysql)
 - a) Found by Conor McNamara from Research Work
 - b) Requires Citi Training and authorization to access
- 2) Findings for EcoSurvey Version 1
 - a) Written by Conor McNamara
 - b) Previously stored in an Excel document



Proposed Work

1. Data cleaning
 - a. Organizing mysql base to get necessary info
 - b. Converting to .csv
 - c. Converting to .arff
 2. Preprocessing
 - a. Using Weka to Preprocess data
 3. Feature Selection/Analysis
 - a. Using Weka to determine predictive features
 4. Classification
 - a. Using Weka to run classification algorithms of various types
 5. Error Analysis
 - a. Observing Confusion Matrices
- 

Tools

- GitHub
- Mysql
- Microsoft Excel
- Python
- Sublime
- Weka
- Microsoft Word



Evaluation of Results

- Error Analysis
- Compare and contrast with first data set findings
- Draw conclusions
- Determine answers to our questions



THANK YOU!
Questions?

