HAL

Humans and Autonomy Lab

HAL2016-03: The Development of Machine Learning Algorithms to Track the African Forest Elephant (*Loxodonta Cyclotis*)

December 30, 2016

Daniel McKee and Dr. Alexander Stimpson

Humans and Autonomy Laboratory
130 North Building
Duke University
Durham, NC 27708

Project Description

The goal of this project is to develop models of African elephant behavior using machine learning and data mining techniques. We are working with a data set which includes geographic coordinate readings documenting the movement of 12 elephants in the Wonga Wongue Reserve located in Gabon, Africa. The geographic data points come from radio collar readings. For each of the timed geographic data points, the data set also includes features describing the specific elephant and location. Our work has included looking for patterns in their daily movements, travel between long term habitat centers, and periods spent in forested areas. We also examined elephant movements towards and away from water. From this information, we are working to create models which predict elephant drinking habits.

Background

Our data set is unique in the number of elephants that are recorded within the reserve location. The availability of this data represents an opportunity to analyze a species which has been reduced in population to critical levels.

This work also presents novel applications of the quickly growing field of machine learning. The work can assist biologists in better understanding the species and help in conservation efforts. We hope that these methods may also be applied to other species.

The radio collar readings which log the coordinate positions of the elephants have been taken at approximately hourly intervals since November 2015. For every data point recorded for an elephant, our data set also provides the distance to water, distance to road, elevation, slope of the location, and whether the location was forested.

Selected Modeling Approach

Clustering approaches were utilized to reveal habitat centers that were significant to the elephants. We used K-means clustering to find the cluster centers and AIC/BIC to measure the optimal number of clusters in the models.

Examining the Euclidean distance between each pair of hourly geographic point readings, we found a lower bound for the distance traveled by the elephants between each of the data points. This allowed us to both analyze the distributions of hourly travel distances and create a new feature to use in our later classification methods.

Our analysis also included examining graphical models of the movement of the elephants with respect to their distance to water and travel into forested or non-forested regions. Locations were documented using a binary forested/non-forested indicator variable for each of the data points. We analyzed the long-term time spent in forested regions along with the fraction of time for each hour of the day spent in forested regions for each of the elephants.

To attempt to model drinking events for the elephants, we used multiple approaches. These included classification of data points as drinking/non-drinking based on a set of selected features and modeling an elephant's distance to water as a time series.

In analyzing the distance from water to understand drinking events, we found that in some periods of time the elephants would spend over a week more than a kilometer away from the water sources documented in our data, a period too long for elephants to survive without water. We took this as a sign that these elephants had found alternative water sources during these periods. By examining graphs of distance to water, we identified elephants with time periods during which regular visits to water implied plausible drinking activity. We used only this data in our modeling efforts for drinking events.

To model the drinking events as a classification problem, it was necessary to label the data points as drinking or not drinking. To do this, we decided to find a cutoff for which to classify all visits within the cutoff as drinking events. We examined the distance to water graphs and found that a cutoff of around 400 meters would result in regular labeled drinking event and avoid long periods without labeled drinking events. After some classification trials, we found that this label resulted in series of points classified as drinking events since elephants often stayed within the cutoff for longer than 1 hour. As a result, we changed the label so that only the first data point in a continuous series would be classified as a drinking event.

For classification, we experimented with a set of features including recent distance traveled, time spent in forest, distance traveled and time in forest since last drinking event, sex of elephants, and measures of recent precipitation. We explored the use of various classification algorithms, including logistic regression, SVM, and tree ensembles.

After finding a particularly high variable importance on the features that gave the time since the last drinking event, we began modeling distance to water as a time series. Our initial efforts in this area included examining the potential of autoregressive models. We analyzed autocorrelations for the periods of time with plausible drinking activity to test the viability of these autoregressive models.

Results

Our clustering analysis revealed habitat locations that were of importance to the elephants. Habitat centers were not clear across all elephants. However, the analysis did show some clear migration between 2 or 3 habitat centers in certain elephants.

By looking at the distribution of distances traveled between hourly readings based on the Euclidean distance metric, we found that most of the travel distances lay within the 100 to 500-meter range. We also found that the elephants were rarely immobile since over 75% of the values lay above 100 meters for most elephants.

In analyzing time spent in forests, we found a range across the elephants. Many elephants regularly alternated between forest and non-forest, while others stayed almost exclusively in forested areas, stayed exclusively outside of forested areas, or alternated regularly between the two. We observed that the graphs of fraction of time spent in forest during each hour of the day for many of the elephants visually showed a rise during the daytime hours, typically between hours 7 and 16.

The increased time spent in the forest during daytime was exhibited in the graphs for all but three elephants. The three elephants that did not clearly display the trend were consistent in that they all spent relatively lower fractions of time in the forest (<0.35) and that none went longer than a couple days before returning to a non-forested area.

We aggregated the data for average binary forest indicator value in each hour of the day across all elephants. Since there was a wide span in the fractions of time elephants spent in the forest, we normalized each of the values for the elephants to a baseline fraction for that elephant. We then tested for statistical significance in the deviation of this result from a uniform distribution across the hours, finding a strongly significant result with chi square test yielding: $\chi^2(23) \approx 180.57$, $p \approx 1.998 \times 10^{-26}$.

In general, the classification rates obtained using the classifier training methods described with the initial cutoff drinking indicator labels were not high enough to be useful in providing predictions of whether a drinking event might occur. Our most effective classification came through the use of tree ensemble algorithms with a random forest classifier yielding an ROC AUC of 0.68. Though we were able to achieve accuracy of over 0.9, this was misleading due to the test data being skewed towards negatives (over 95% negative samples). In fact, this high accuracy came with very low recall of less than 0.05 and precision of approximately 0.05. When classifiers were trained that obtained a higher recall of around 0.7, precision remained less than 0.1, while accuracy dropped to less than 0.6. As mentioned in our motivation for time series analysis, the most effective classifiers revealed strong variable importance on the time since last labeled drinking event and distance traveled since last labeled drinking event.

Upon analyzing the autocorrelations of distance to water readings for the series of plausible drinking, we found high autocorrelations centered around lag 24 hours. This finding could be significant in that it implies that the times of elephant drinking events could be strongly dependent on a 24-hour period. These findings make work using time series models promising.

Future Work

As more data is collected we can make longer term models of seasonal change in the elephants. More data could also allow for building models of transitions between elephant habitat centers identified by clustering analysis. These changes in habitat centers could possibly be viewed as state changes for modeling purposes.

Regarding modeling of drinking behavior, we will continue work with modeling distance to water as a time series and continue work with ARIMA modeling techniques to create predictive models.

Motivated by similar work that has been conducted for caterpillar eating patterns, we also plan to explore the use of Hidden Markov Models and Hidden semi-Markov for behavioral modeling.

Acknowledgments

The Humans and Autonomy Laboratory would like to thank the Duke Tropical Conservation Initiative (DTCI), the Information Initiative at Duke (IID), the Pratt School of Engineering, and the African Elephant Fund (AEF) for their support of this project. We also thank the Agence Nationale des Parcs Nationaux (ANPN) and the Centre Nationale de la Recherche Scientifique et Technique for permission to conduct the research and for their administrative and logistical support. We also thank David Fine and the staff of the Rèserve Prèsidentielle de Wonga Wonguè for their collaboration and support while in the reserve.