Grace Gao

Bass Connections

October 17th, 2020

State of the Question

**Abstract:**

The purpose of this research project is to analyze the effects of the Eighty Year’s War on the migration pattern of artists to and away from Antwerp and Amsterdam, two large cultural centers of Netherlands at that time. This study hopes to answer the following research questions through analyzing the data scraped from the ECARTICO database:

1. How does the background of an artist—nationality, religion, experience, age, gender and socio-economic class—influence their decision of migration?
2. How does the artist migration model differ between Antwerp and Amsterdam? Specifically, 1556-1568 (When Spain conquered Netherlands, but before the Eight Year’s War, 1568-1648 (Eighty Year’s War), and 1648-1675 (Golden Age/Post-War).

To go about conducting the data analysis, I will be implementing the machine learning algorithm—Random Forest Regression—on the extracted, cleaned and tided ECARTICO data. Through this algorithm, I hope to gain a good prediction of the output (resulting migration pattern and destination) from a mixture of continuous and categorical variables (identifying features of the artists).

For the purpose of this literature review, I will be dissecting the methodology, two implementations and the documentation of the Random Forest Regression. Through this literature review, I hope to a gain a better grasp of the algorithm and a refined approach to implementing it for my project.

**Review of Literature:**

*Methodology*

To truly understand the implementation of the algorithm and the application of it to my project, I need to dive into the methodology of the algorithm. The general concept of random forests classification was proposed by Leo Breiman with the specific instance of the RF-CART. Although I won’t dive specifically into the history of the algorithm, I will be discussing the actual logic behind it. As is its name suggests, a regression tree is built by recursively splitting the data into more and more homogenous “nodes” until it reaches the “terminal node” (Grömping 4). Each split is based on the division of a variable, which is choose by a “splitting criterion”. To conduct the predictions, you follow along a branch of the tree until the specific terminal node. With this specific path, you can find the predicted response value, which is just the specifical terminal node’s average response. For the random aspect of the random forests, each tree is built from a random subset of the data, and each split is based randomly on a subset of the different potential variables. Although the randomness causes varying predications in each tree, the averaging of these predictions’ accounts for the variance.

Overall, this algorithm is ideal as it’s able to be implemented in a large set of data (like the ones I will use) and obtain non-linear relationships among the variables. Furthermore, it’s more favorable than statistical models like the linear regression model or forwards/backwards elimination as it’s able to account for categorical variables more accurately than the linear regression model could. Particularly for my data, the only numerical variables are age, the family member count, and the different years. So, it’s crucial I have a good algorithm that can account for the numerous categorical variables and still make an accurate prediction. Finally, the randomness and the averaging behind the algorithm enables a more accurate prediction than other models, especially since many of my variables could be more correlation rather than causation.

*Implementation One: Face Alignment*

In the paper “One Millisecond Face Alignment with an Ensemble of Regression Trees” written by Vahid Kazemi and Josephine Sullivan, they discuss the implementation of regression trees to estimate the face’s landmark positions based on an image. In particular, they present the algorithm’s ability to perform face alignment in milliseconds. With the cascading regression functions, they were able to efficiently estimate the shape of the face from an initial estimate and indexing of the pixel intensities from the initial estimate.

Although this refined version of the basic regression tree algorithm is way above my expertise, I believe there is an essential step of this research that can apply to my research project. In their project, an issue they encountered was the difficulty of accurately estimate the shape because of different factors like facial deformation or lighting. But they standardized the image into a normalized coordinate system based on the current estimate of the shape rather than based on the global coordinate system of the image and then used this normalized coordinate system to apply the regression tree. For my application then, I know the data could be potentially formatted very different across every artist’s datapoints. Thus, I will standardize the format of each datapoint prior to applying my algorithm. By taking this extra step, I will be able to account for a lot of issues that could potentially arise from the lack of uniformity among the data.

*Implementation Two: QSAR study of the Anti-Hepatitis Peptides*

In this study conducted by Gunjan Mishra, Deepak Sehgal and Jayaraman K Valadi, utilized the Random Forest and Extra-trees regressors to predict the anti-hepatitis activity of antimicrobial peptides, which are host defense peptides that could potentially be used to replace broad-spectrum antibiotics.

Similar to my approach, this study also collected data through publically available resources and actually collected 8809 descriptors (identifying features) but cut it to the 157 descriptors that were the most applicable. Although my data would have nowhere near that number of initial descriptors, the approach of only keeping the critical ones is something I have to really consider. Currently, I have a set of variables I hope to analyze. But I also have to consider that many of the variables might not be important at the end or there are variables that I didn’t initially consider in this set. So, the approach I may take is to collect all the possible variables and conduct an initial analysis on them to determine if I will keep them in my final set. Furthermore, their approach of splitting the data set into 80% training set, 10% test set and 10% independent set is something I want to follow. Originally, I was only going to split the dataset in half, one training set and one test set. But I believe this approach with have a greater percentage of the training set will allow for my model to be fitted more accurately.

**Documentation**

Finally, I wanted to become more familiar with the documentation of the RandomForestRegressor from scikit-learn. Scikit-learn is a software machine learning library for Python users to implement. As someone who is more familiar with R rather than Python but this specific algorithm being more developed in Python, it is important that I become very familiar with the different functions and the syntax of the regressor. Even with this specific documentation, it is important to consider that there are many different versions that have been developed from this standard regressor. Therefore, when I implement the regressor, I have to consider the possibilities of utilizing alternative versions best fit for my purpose and my type of data.

**Conclusion**

Through reviewing these scholarly articles, I have developed a deeper understanding about the Random Tree Regressor and a more solidified approach to conducting my own analysis. Taking inspiration from these two studies and becoming familiar with the documentation, I will implement certain aspects of their approaches to help improve my own algorithm.

Works Cited:

“3.2.4.3.2. Sklearn.ensemble.RandomForestRegressor¶.” *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

Grömping, Ulrike. “Variable Importance Assessment in Regression: Linear Regression versus Random Forest.” *The American Statistician*, vol. 63, no. 4, 2009, pp. 308–319., doi:10.1198/tast.2009.08199.

Kazemi, Vahid, and Josephine Sullivan. “One Millisecond Face Alignment with an Ensemble of Regression Trees.” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, doi:10.1109/cvpr.2014.241.

Mishra, Gunjan, et al. “Quantitative Structure Activity Relationship Study of the Anti-Hepatitis Peptides Employing Random Forest and Extra Tree Regressors.” *Bioinformation*, vol. 13, no. 03, 2017, pp. 60–62., doi:10.6026/97320630013060.