Generalizing Parallel Processing Designs for Machine Learning Based Suitability Modeling of Cash Crops in California

Alejandro Hernandez, Nilay Nagar, Dr. John Korah
Department of Computer Science @ California State Polytechnic University, Pomona

Contact Us!

Abstract

Global demands for food production are expected to rise 70% by 2050 as the world population surpasses 9 billion, alongside an ever-growing pressure on farmers to produce more with the same resources.

- In response, statisticians, computer scientists, and Geography Information System (GIS) specialists have united in the practice of Precision Agriculture, a strategy to rely on statistical modelling to optimize crop productivity.
- A subtopic of this area is the problem of identifying ecological niches of a crop within a geographic space. Crop suitability modelling, as it is referred, has the potential to radically improve crop production by mapping over a region, each location's relative suitability.

➢ **Research Goal**

Designing and implementing generic parallel algorithms for machine learning techniques (e.g. MaxEnt) to map the suitability of cash crops over the state of California, USA. We present expected results and performance metrics to illustrate improvement with parallel execution.

Research Challenges

A presence-only dataset is a collection of coordinate points where a species has been observed. In company with maps of recorded climate, one may build a statistical model to map environmental suitability. However, two challenges must be addressed.

1. Sampling bias is inherent to any presence-only dataset because it is virtually impossible to certify any location as truly unsuitable for a species just because it was not observed there (see Fig. 1). Uncertainty is built-in to the records, regardless of sampling effort. Because any model's output will reflect patterns/biases of the data, using presence-only data prevents us from using traditional models. See *Machine Learning for Crop Suitability* for more.

2. Enormous datasets are expected for this work, as suitability is often mapped over extensive regions (each climate map in Figure 2 contain over 1.4 million pixels). As a result, data-processing during model building may take hours on a single computer, which is far from acceptable. See *Parallel Processing Design* for more.

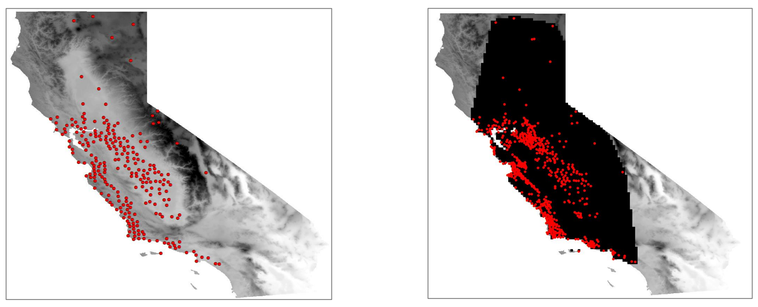

Figure 3. (Left) Mapped presence points of strawberries in California, collected in 2018. (Right) Bias region used to identify trusted pseudo-absences.

Machine Learning for Crop Suitability

Machine learning (ML) is a popular intersect of computer science and statistics used to build powerful models. Predicting crop suitability is ultimately a binary classification problem (suitable/unsuitable), for which many standard ML models are well suited. However, inherent bias of presence-only data prevents us from using these traditional algorithms.

- **How is the data biased? Why can't we trust an "unsuitable" classification?**

If there is enough belief that a presence-only dataset is incomplete, we cannot treat non-presence points as true absences; this renders many traditional binary classifiers as unfit.

- **What model can we use?**

This research studies the *MaxEnt* (Maximum Entropy) algorithm [1]. The idea is that we can generate many models that output a probability of suitability- but we choose the one that has the greatest *informational entropy*; this is synonymous to a probability density function that maximizes the probability of any location being suitable, under certain constraints.

- **What are the model inputs?**

For the model to differentiate environmental conditions under which a species can occur, inputs include climate maps, presence points, and a *bias* area that visualizes the limit of locations from which the model "learns" (Fig. 2 & 3). This selected region illustrates our distrust in the accuracy of recorded absences in areas distant from true presence locations.

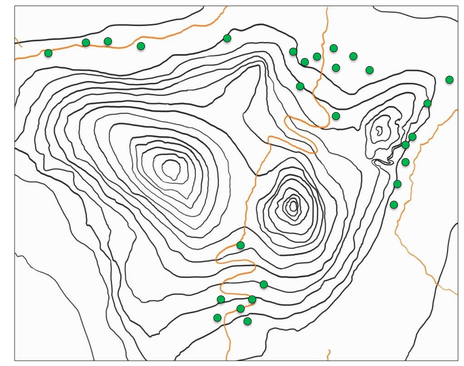

Figure 1. Example elevation map displays presence locations (green) that are all recorded suspiciously close to roads (orange). This illustrates sampling bias, as surveyance efforts in hard-to-access regions were likely weak.

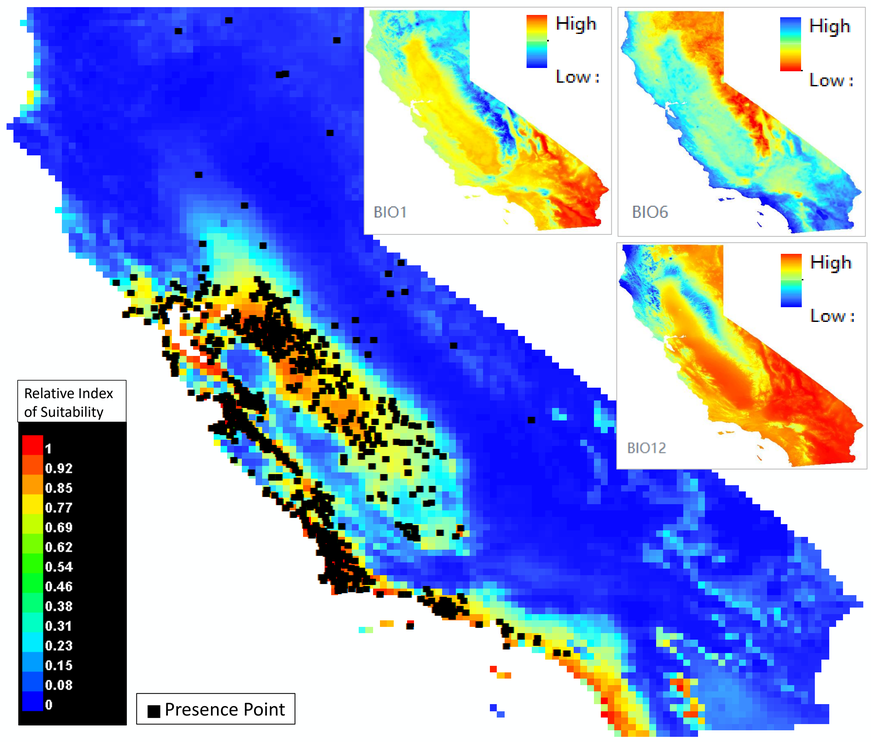

Figure 2. (Top right) Climate maps of Annual Mean Temperature (Bio1), Min. Temperature of Coldest Month (Bio6), Annual Precipitation (Bio12); (Center) Suitability map for 2018 Strawberry (from Figure 3) in California, generated with a MaxEnt program by S.J. Philips, M. Dudík, R.E. Schapire [3].

Expected Results

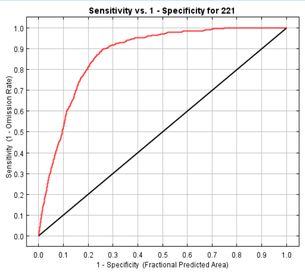

Figure 5. ROC plot from MaxEnt

Model accuracy is captured by plotting *sensitivity* (True Positive Rate) and *specificity* (True Negative Rate) at various classification thresholds. Figure 5 shows Maxent's performance with AUC = 0.872 (red line), compared to a random binary classifier with AUC = 0.5 (black line).

Parallel Processing Design

Parallel processing is the computing method of running multiple processors (CPUs) to handle separate parts of an overall task.

Initial Parallel Processing Design

➢ The initial approach used in this project is called Data Parallelism.

➢ Our aim is to parallelize multiple homogeneous models on partitions of the dataset to reduce computational time complexity.

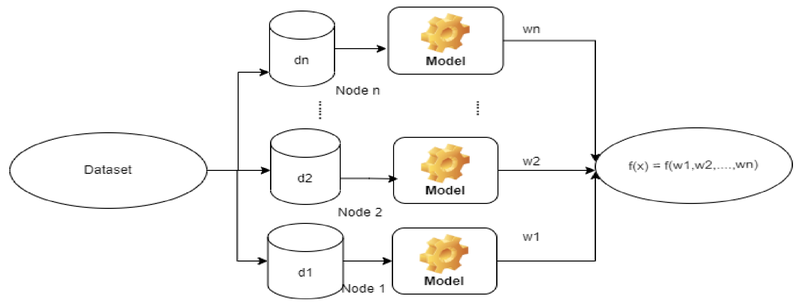

Figure 4. Flow Diagram for proposed Data parallel model approach

➢ Partitioning the dataset into chunks in a random fashion.

➢ Training the partitioned datasets using multiple homogeneous classifiers/models.

➢ Combining the model predictions and reduce them to get the required result.

➢ In the below diagram, w represents the predictions generated from each model.

➢ We plan to use Message Passing Interface (MPI) for the proposed architecture, which is a standardized message-passing developed for distributed and parallel computing. MPI allows to scale over different systems.

Conclusion and Future Directions

➢ We have successfully formulated initial MaxEnt based models for 'Grapes', 'Strawberries', and 'Lettuce' data points in California region using SDM ToolBox and ArcMap.

➢ We have developed python scripts for data pre-processing required before training on MaxEnt.

➢ We plan to develop parallel designs for MaxEnt in python using the data parallel approach to increase speedup.

➢ We plan to generate and use pseudo-absence data in classic machine learning techniques and compare its accuracy with MaxEnt.

References

[1] S. J. Phillips, R. P. Anderson, and R. E. Schapire, "Maximum entropy modeling of species geographic distributions," *Ecological Modelling*, vol. 190, no. 3-4, pp. 231–259, 2006.
[2] M. Dudík, S. J. Phillips, and R. E. Schapire, "Performance guarantees for regularized maximum entropy density estimation," *Learning Theory*, pp. 472–486, 2004.
[3] Steven J. Phillips, Miroslav Dudík, Robert E. Schapire. [Internet] Maxent software for modeling species niches and distributions (Version 3.4.1). Available from url: http://biodiversityinformatics.amnh.org/open_source/maxent/.

Acknowledgements

All maps were generated in ArcMap (ESRI). We acknowledge the research collaboration with Dr. Gabriel Granco (Dept. of Anthropology & Geography, CPP) and Dr. Subodh Bhandari (Dept. of Aerospace Eng., CPP).