Exploration of Marine Biology

Identifying marine life forms by using Machine learning on Flow Cytometry data

By

Abhigyan Kaustubh

Elton Dias

Tanmay Modak

Under the guidance of

Dr. Bill Howe (Project Guide/ Sponsor)

Mala Sarat Chandra (Adviser)

Isabel Carrera Zamanillo (TA)

EXECUTIVE SUMMARY

We will be undertaking our capstone project in collaboration with researchers from the UW eScience Institute. The eScience Institute launched in July 2008, dedicated to the support of research computing and to leadership in key areas of eScience expertise, including data mining, machine learning, and sensor networks. Our emphasis is on the emerging need to support data-centric high performance computing. As available data grows exponentially, researchers must apply computational and data management approaches to their research. Access to computer science expertise is key to this endeavor. Our eScientists act as matchmakers, helping domain scientists apply the most appropriate technology to their research.

The eScience team consists of individuals with backgrounds in physics, astronomy, bio-engineering, bioinformatics, data management techniques, and computer science. As part of this initiative we will be studying and working with large data sets related to fields of the physical sciences and applying various computational algorithms to derive new insights. Our objective using this project is to define a novel or news ensemble techniques that could be used by researchers for solving similar research problems which can be deployed through various big data platforms and applicable to various sciences. Our final deliverable will be a set of findings and possible recommendations for solving data science problems in a particular domain.

I. Organizational background

The eScience Institute launched in July 2008, dedicated to the support of research computing and to leadership in key areas of eScience expertise, including data mining, machine learning, and sensor networks. Our emphasis is on the emerging need to support data-centric high performance computing. At the core of the eScience Institute are individuals who have proven track records in developing and applying advanced computational methods and tools to real world problems. Their task is to seek out and engage researchers across disciplines where eScience approaches are likely to have the greatest impact. To ensure that researchers have access to the necessary physical infrastructure, the Institute has undertaken coordinated planning and support for advanced local and remote computational platforms. This includes developing relationships with commercial and non-commercial service providers as well as the development of shared facilities on campus. This support extends to assistance in the preparation of select proposals where we are able to focus resources, improving their chances for success.

II. Information Problem

As an abstract of the problem that we plan to tackle through this capstone project would be to create a better data analysis technique to analyze oceanography data produced through SeaFlow – A Novel Flow Cytometer developed for continuous real-time observations of natural assemblages of small phytoplankton cells, including Prochlorococcus. Current automated flow cytometers primarily target the dynamics of phytoplankton communities in coastal environments and lack the sensitivity to resolve the spatial distribution of Prochlorococcus, the smallest phytoplankton that dominates the open ocean. SeaFlow makes it possible to explore surface phytoplankton dynamics at a spatial scale ranging from a few meters to thousands of kilometers. Our aim would be too create an effective data analysis medium to automatically cluster and count micro-organisms in the scanned ocean areas also with geo-referenced data visualization to prove our findings.

III. Project Description

Through our project entitled "Flow Cytometry based exploration of surface Marine Biology" we intend to accomplish deliverables as outlined below (rough steps) -

1. Talk to the customer (user research) and perform preliminary visual and exploratory analysis of the data to better understand the problem statement.
2. Clean and format the data. Estimated 100 three minute files data window sets.
3. Decide on the necessary Machine Learning algorithms to be implemented (2nd 3rd order derivatives, Experimental performance, etc.), as well as certain pre-requisite procedures like clustering to isolate the biogeographical patterns.
    a. Tools: Python using libraries like scikit-learn, mlpy.
4. Produce visualizations to showcase useful insights (Draw the cruise track in colors by cluster label).
5. Once achieving a suitable performance and results, do a comparison of different Big Data systems available that will allow us to effectively scale the solution.
        1. Reproduce algorithms on other Big Data systems.
        2. Optimize them accordingly to make them faster.
6. Finalize results, produce visualizations. Create Poster.

IV. Related Work.

Current automated flow cytometers primarily target the dynamics of phytoplankton communities in coastal environments and lack the sensitivity to resolve the spatial distribution of Prochlorococcus, the smallest phytoplankton that dominates the open ocean. Here we present SeaFlow, a novel flow cytometer developed for continuous real-time observations of natural assemblages of small phytoplankton cells, including Prochlorococcus. Unlike other flow cytometers, SeaFlow does not use sheath fluid. Instead, a virtual-core is used to determine the position of a particle in the stream of seawater. By eliminating sheath fluid, SeaFlow can continuously sample the seawater stream directly from a ship's intake system. Image analysis is used to automatically align the laser with the optical system and then monitor and correct for drift. SeaFlow performs rapid quantification (up to 24,000 cells per second) of multidimensional characteristics of phytoplankton cells in the pico- to nanophytoplankton size range (0.5-20 μm) to analyze the equivalent of 480 traditional flow cytometry samples per day while on board a research vessel. Data analysis tools have been created to automatically cluster and count phytoplankton populations with geo-referenced data visualization. SeaFlow makes it possible to explore surface phytoplankton dynamics at a spatial scale ranging from a few meters to thousands of kilometers. An example data set is presented from a 450-km long transect near the Hawaiian Islands with the continuous data aggregated in 3 min intervals. As SeaFlow is further refined to become increasingly autonomous, adaptive sampling based on observed changes in community structure will be possible. The ability to sample at positions along a transect where the most significant biological gradients are observed will provide new insights into microbial structure and functions. The tools required to interpret a real-time shift in community structure will need to be developed and tested. Simple and obvious cytometry metrics such as rise or fall in event rate, or the event

rate associated with a particular population have been tested. A more complex analysis based on the cytometric diversity measurement developed by Li (1997) has also been run on the SeaFlow data in real-time. Regardless of the measure, determination of a significant biological signal "in the moment" is challenging primarily because the candidate signaling algorithm must be predictive. At present, useful interpretations of measures based on high-resolution cytometry, alone or correlated to other shipboard data streams, do not exist. Data mining of current and future SeaFlow data sets will help in development of these predictive tools.

V. Intervention, Outcome, and Evaluation Criteria

Our evaluation criteria would be in accordance with our customer/ researchers, and their needs/ things that they are looking to find in the data. They have already been working on this for a while, and would be stating their requirements as we progress with the project. Being a new and unexplored area, the exact requirements are not fully know yet (other than the algorithms that they wish to employ on the data set.) The primary successful evaluation criterion would be that those algorithms run successfully on the given initial data sets (100 three minute files) using python and tools like scikit learn on local servers.

We will then deploy these developed algorithms on each of different systems. As we would be experimenting with different systems, we expect different levels of performance and accuracy. We will then conduct an analysis of each of these systems and come up with a set of inferences and recommendations about how to approach such data science problems while using different systems, and possibly be able to also recommend the best system for such problems.

VI. Citations

We are still working on the literature review and are waiting to get more documents form our sponsor regarding the same. For the time being, we are referring general research work:

http://ec.europa.eu/research/energy/pdf/seaflow_en.pdf/

http://www.aslo.org/lomethods/free/2011/0466.pdf

http://aem.asm.org/content/65/10/4404.full.pdf+html/

http://escience.washington.edu/