Adventures of Two Student Research Computing Facilitators

Extended Abstract

Noah Howard
School of Computing and Information Science
University of Maine
noah.howard@maine.edu

Nicholas Colella
Harvard University
Department of Chemistry and Chemical Biology
colella@g.harvard.edu

Kasey Legaard
Research Scientist
University of Maine
kasey.legaard@maine.edu

Saritha Nellutla
Bridgewater State University
Department of Chemical Sciences
snellutla@bridgew.edu

Evan McCoy
Department of Electrical and Computer Engineering
University of Maine
evan.mccoy@maine.edu

Larry Whitsel
Advanced Computing Group
University of Maine System
larry.whitsel@maine.edu

Chris Wilson
Advanced Computing Group
University of Maine System
chris.wilson@maine.edu

Bruce Segee
Advanced Computing Group
University of Maine System
segee@maine.edu

ABSTRACT

The NSF-sponsored Northeast Cyberteam program (https://necyberteam.org) is matching student research computing facilitators with research projects at small and medium sized institutions that need help making use of high performance computing resources. Students are selected based on relevant domain knowledge and level of interest in exploring the Research Computing Facilitator role as a career path. Each student is paired with an experienced mentor, and each project lasts 3-5 months.

The poster presents results from two of the ~10 Northeast Cyberteam projects that are either completed or in progress at the time of the conference. Students will be prepared to discuss the research projects that they supported, how their efforts advanced each project, reflections on what they learned about the Research Computing Facilitator role, and recommendations on how other sites might make best use of students as Research Computing Facilitators.

In the first project, conducted at the University of Maine, improvements to application performance and parallelization have enabled production of forest attribute data at significantly higher volume, changing the philosophy of forest mapping from creating a single map to creating and assessing thousands of maps, making it possible to assess and (to some degree) trade off errors between maps.

The second project expanded efforts to introduce computational chemistry into the undergraduate chemistry curriculum at Bridgewater State University. Having access to a high performance computer cluster allows for studying "real world" systems and also provides students with an opportunity to experience how research computing is done in academia and/or chemical/pharmaceutical industries.

CCS CONCEPTS

• **Social and professional topics** → *Professional topics*; *Computing education*; *Informal education*;

KEYWORDS

Genetic Algorithms, Support Vector Machines, Forest Mapping, Remote Sensing,Physical chemistry, Inquiry based laboratory experiments, Gaussian09, Infrared spectroscopy, Diels-Alder reaction, Fluorescence spectroscopy

ACM Reference Format:
Noah Howard, Nicholas Colella, Kasey Legaard, Saritha Nellutla, Evan McCoy, Larry Whitsel, Chris Wilson, and Bruce Segee. 2018. Adventures of Two Student Research Computing Facilitators: Extended Abstract. In *PEARC '18: Practice and Experience in Advanced Research Computing, July 22–26, 2018, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3219104.3229251

1 INTRODUCTION

The Northeast Cyberteam program aims to make regional and national cyberinfrastructure more readily accessible to researchers at small and mid-sized institutions in Northern New England, enabling researchers to produce faster/more accurate results. Central to the program is the idea of giving student-facilitators hands-on Research Computing Facilitation experience while also advancing research and education projects that need help. This poster shows

PEARC '18, July 22–26, 2018, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-6446-1/18/07.
https://doi.org/10.1145/3219104.3229251

two early results, where the program matched students with researchers and experienced mentors to advance a research project and an education project. In one case, the student facilitator is an undergraduate student, and in the second, a professional educator exploring the Research Computing Facilitation profession.

For the research project, "**Using Genetic Algorithms and Support Vector Machines in Forest Mapping**", improvements to application performance and parallelization have enabled production of forest attribute data at significantly higher volume, changing the philosophy of forest mapping from creating a single map to creating and assessing thousands of maps, making it possible to assess and (to some degree) trade off errors between maps.

The education project, "**Integrating Computational Chemistry into the Undergraduate Physical Chemistry Laboratory with HPC**", expanded efforts to introduce computational chemistry into the undergraduate chemistry curriculum at Bridgewater State University. Having access to a high performance computer cluster not only allows for studying complex systems encountered in âĂIJreal worldâĂİ but also provides students with an opportunity to experience how research computing is done in academia and/or chemical/pharmaceutical industries.

2 PROJECT 1: USING GENETIC ALGORITHMS AND SUPPORT VECTOR MACHINES IN FOREST MAPPING

Student Facilitators: Noah Howard, Evan McCoy, Undergraduate Students University of Maine
Researcher: Kasey Legaard, Doctoral Candidate University of Maine
Mentors: Larry Whitsel, Cyberinfrastructure Engineer, **Chris Wilson**, Data Architect, **Bruce Segee**, Director, Advanced Computing Group, University of Maine System

2.1 Project Abstract

Satellite-derived maps of forest conditions play diverse roles in research and resource management. Maps provide a basis for planning and executing field studies, developing and calibrating models, quantifying ecosystem processes or services, and evaluating environmental change. Natural resource managers use maps to characterize resource conditions, project changes, and direct management actions. However, inferences and decisions must be made within the context of map error, and methods used to produce maps generally result in patterns of error that are potentially detrimental. The researcher has developed machine learning techniques that effectively reduce undesirable systematic error when mapping forest attributes from satellite imagery and geospatial data. The approach is based on support vector machines (SVMs) using a multi-objective genetic algorithm (GA) designed to simultaneously minimize both total and systematic error. This approach has been used to map tree species abundance and forest disturbance across a northern Maine study area, obtaining outcomes that compare well against other approaches previously applied either regionally or nationally. These algorithms are computationally demanding, and large-scale applications will require use of high-performance computing resources.

To get these computationally demanding algorithms to run at large scale, the project developed enhanced parallelization and algorithm improvements coupled with more efficient and more automated methods for input and output data handling. The primary project outcome was software that supports locally adaptive mapping of forest resources and environmental conditions across large spatial scales through innovative algorithms and efficient use of available cyberinfrastructure.

The researcher's original model was developed in MatLab and designed to be run on a single laboratory machine. The model took advantage of the thread pool parallelization provided by MatLab, but was limited by memory and performance concerns. In addition, the researcher hoped to be able to checkpoint the modeling process at key locations as to provide a branching point for exploring various sub-problems. The original MatLab code was ported to a hybrid C/C++ environment to accommodate an improvement to parallelization with minimal disruption to the algorithms being used in the original model. To that end, the libsvm library was adopted to implement the Support Vector Machines while a C++ version of the Non-Sorting Genetic Algorithm (NSGA2) multi-objective algorithm was used. The OpenMP parallelization model was chosen to improve performance on the new target platform, the supercomputer at the University of Maine.

Machine learning is a tool that can be used in multiple ways. Some problems of interest to the researcher involved using the model to classify data into two categories. These classification problems provided the ability to rapidly identify areas of forest that had been disturbed – by insect activity in the case of one problem (monitoring defoliation by eastern spruce budworm) or through forest harvesting in other problem spaces. The same set of algorithms employed in a slightly different way allows the researcher to perform regression tests that allow predictions of the abundance of a particular tree species based on satellite data. In the regression problem use-case physical observations are matched to patterns of satellite data and SVMs are trained to produce a prediction of species abundance given the available data.

To give an idea as to the scope of the problems being solved, we offer the following example of a classification problem. Training data consisted of 750 actual observations spread evenly across the area of interest. The genetic algorithm used 500 individuals – each of which was a particular parameterization of a support vector machine. The individuals were allowed to evolve over 120 generations. For each generation, the total population was randomly divided into 10 partitions of 75 individuals. Nine of those partitions were used to train the model 10 times with the 10th partition being used to calculate the accuracy of the resulting model and area being predicted by the model. The average of each result were used as the fitness objectives for the individual. In each generation the NSGA2 model is used to rank each model with the best models being used to generate the next generation of models. In the example here, the result at the end of computation was a set of approximately a dozen SVM models that represented the best matches selected from the 6 million models constructed and evaluated. The resulting SVM(s) are then used across an area of approximately 7000 sq. mi. to identify/predict the conditions inherent in any particular pixel of data. The results output of the SVM is then used to generate an overlay usable with GIS mapping tools.

The model optionally allows for feature selection and exclusion learning. If you think of the input data as an m-by-n matrix of locations in rows and approximately 100 columns of properties from the satellite-data, feature selection allows the algorithm to ignore columns of data that reduced the accuracy of the model, while exclusion allows the algorithm to ignore rows of data that reduce the accuracy. This ability to allow the model to excise data that is producing poor results increases the overall accuracy of the machine learning.

Improvements to application performance and parallelization have enabled production of forest attribute data at significantly higher volume, changing the philosophy of forest mapping from creating a single map to creating and assessing thousands of maps, making it possible to assess and (to some degree) trade off errors between maps.

One of the key pieces of the software project was to take the output from the SVM and turn it into a map. For this, a custom piece of software was developed that took the model and raster files as input and generated a map. The software is written in Java and makes use of the libsvm and GDAL libraries for interacting with the SVM model and processing the rasters. It is based on software that was originally written by the PI in Matlab, but was ported to Java for better runtime as well as compatibility with the new model generator. Using multithreading, we were able to reduce map generation time from 40 minutes to 4 minutes. The software was written to allow for different amounts of input as well as a variety of problem types with the idea that the software will eventually be used in different regions for classification or regression problems.

This project was mutually beneficial for all parties involved. The student had significant involvement in a real-world software development project including writing, debugging, testing and verification. The researcher was key in explaining the algorithms and interpreting the results. The benefit to the researcher was code that can be run on a High Performance Computing environment, rather than an overtaxed desktop system. The Mentors utilized their technical, organizational, and people skills to lead the project to a successful conclusion. The overall CITeam project gained valuable insight into how a multidisciplinary cyberinfrastructure facilitation team can operate effectively. [1–6]

3 PROJECT 2: INTEGRATING COMPUTATIONAL CHEMISTRY INTO THE UNDERGRADUATE PHYSICAL CHEMISTRY LABORATORY WITH HPC

Student Facilitator: Nicholas Colella Harvard University Department of Chemistry and Chemical Biology
Researcher/Educator: Saritha Nellutla Bridgewater State University Department of Chemical Sciences

3.1 Project Abstract

Six computational chemistry laboratory experiments were developed using Gaussian09 to complement the Physical Chemistry II course at Bridgewater State University (BSU). Each of these "dry labs" can either be implemented as a stand-alone experiment or combined with complimentary wet lab experiment to create a more engaging inquiry-based laboratory experience for students. Furthermore, these experiments highlight the quantum chemistry underpinnings of various spectroscopic techniques such as UV-visible, nuclear magnetic resonance, infrared, and Raman spectroscopy typically covered as part of physical chemistry II course at BSU.

Students will follow the steps listed below to perform the desired Gaussian09 calculation on the high performance cluster, namely C3DDB, at the Massachusetts Green High Performance Computing Center.

(1) Draw the molecular structure (using a molecular editing software such as ChemDraw and/or Avogadro)
(2) Generate a complete input file .gzmat (using Notepad++)
(3) Generate a .job file (using Notepad++)
(4) Upload these files to the server (using FileZilla)
(5) Run the files on the cluster (using PuTTY)
(6) Download the .out file (using FileZilla)
(7) Analyze the results (using Avogadro and/or Notepad++)

Complexity of the newly developed experiments range from simple to complicated tasks as illustrated in a few examples described below.

- Students will learn how to balance between computation time and accuracy of results by testing a variety of basis sets to calculate the optimized geometry of a copper(II) ion surrounded by six water molecules, including experimentally observed features like Jahn-Teller distortion and angled water molecules.
- Students will calculate the infrared spectrum of aspirin, a common analgesic. They will use their results to visualize animations of the vibrational motion of functional groups of aspirin such as carboxylic acid, ester and C-H bonds in the benzene ring.
- Students will evaluate the optimized molecular geometry and molecular orbitals of 7-hydroxycoumarin-3-carboxylate, a fluorescent dye commonly used in protein assays. They will then use the results to predict the characteristic absorption and fluorescence spectra of this compound.

Through these laboratory experiments, the students will gain knowledge of the breadth and depth of computational chemistry modeling available through computational software packages like Gaussian09. Moreover, the use of high performance computing cluster to implement these calculations has added technical and logistical advantages such as: (i) ample processing power and memory, which allows students to perform multiple calculations, each employing multiple processors, at once by exploiting parallel processing capabilities of Gaussian09, (ii) faster computing speeds, which will allow students to apply the quantum mechanical theories they learned in class to larger "real" systems like aspirin and 7-hydroxycoumarin-3-carboxylate instead of model systems like water and carbon dioxide, and (iii) fewer points of failure because of circumvention of common problems such as software/operating system incompatibility or instability on personal computers.

The development of these labs benefited the community at multiple levels. The student facilitator gained more experience using computational chemistry in the classroom, particularly interfacing with the computing cluster. The educator obtained dry labs which are both self-contained and modular; the experiments are

a comprehensive introduction to using Gaussian09 with high performance computing and provide a foundation for expansion (e.g. different molecules, other spectroscopies). Most importantly, since use of HPC is becoming more mainstream for research computing in academia and industry R&D these labs provide undergraduate students the opportunity to utilize HPC as part of their education and hence can serve as a gateway to a potential career pathway in research computing for chemistry majors at BSU.

REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[2] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification. (2003).

[3] Kasey R Legaard, Steven A Sader, and Erin M Simons-Legaard. 2015. Evaluating the impact of abrupt changes in forest policy and management practices on landscape dynamics: analysis of a Landsat image time series in the Atlantic Northern Forest. *PloS one* 10, 6 (2015), e0130428.

[4] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. 2003. Building text classifiers using positive and unlabeled examples. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 179–186.

[5] David Meyer. 2004. Support vector machines: The interface to libsvm in package e1071. (2004).

[6] C OpenMP. 2002. C++ application program interface.