7a.016.DU - A Continual Learning Framework for Domain Adaptation and Provenance Tracking

Gail Rosen[1] (PI), Zhengqiao Zhao[1] (RA), Jay Vandervoort[2] (IAB)
College of Engineering, Drexel University[1], Becton, Dickinson and Company[2]

Project Start: July 2018	End Date: June 2019	Project Budget: $40K	Spent: ~$10k

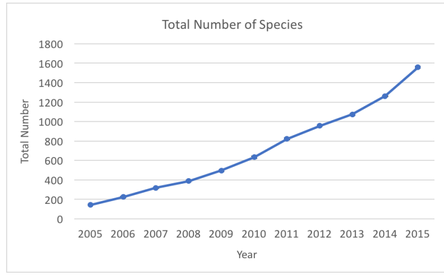
Project Summary:

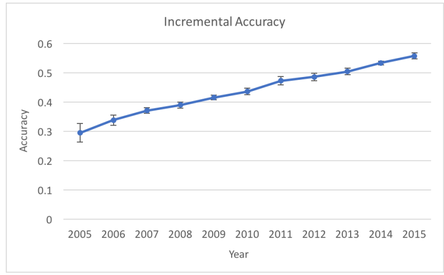
- Implement a continual learning framework that can incrementally update classifiers and efficiently relabel old datasets without the cost of reprocessing the existing database. *Significance*: In previous solution, the classifiers are entirely retrained when the database gets updated by new information – this is redundant and wastes time and computation. Such a framework can greatly reduce the computational redundancies and make instant decisions using latest information.
- Enable Semi-Supervised Learning and novel classes detection. *Significance*: One would like to compare performance of new data to older data and making these comparisons is impossible since the data was classified with older knowledge. By semi-supervised learning, we can make a unified system that will leverage labeled and unlabeled data to incrementally update the classifier and previous datasets. This will enable one to classify on a large volume data, which occurs in healthcare, social networks.

Details of Progress/Achievements:

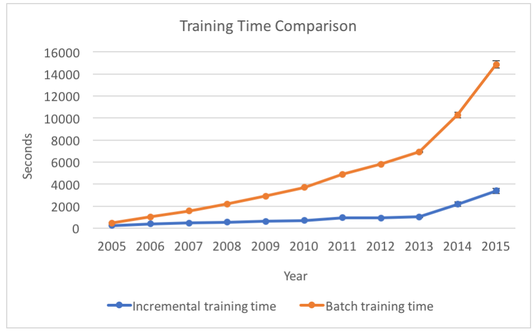
- we have downloaded all the complete bacterial reference genomes from National Center for Biotechnology Information (NCBI): ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/assembly_summary_refseq.txt.
- we organized the genome according to their submission date to simulate the process of development of this database and we created a data pipeline to parse the genome sequence data and simulated the testing reads without sequencing error.
- we built a base learn for taxonomic classification using Naive Bayes classifier and tested the performance. We also incrementalized the base learner and show the accuracy.
- We have completed some experiments and evaluated the performance of our model and got some preliminary results for further analysis

- Simulation (left figure) shows that total number of species in NCBI dataset grows over time. Incremental learning accuracy (right figure) shows that our model can improve its accuracy over time effectively.





- Training time comparison between incremental learning approach and traditional approach (batch update). Incremental learning reduces the total yearly updating time by 73 %.



PROJECT DELIVERABLES

Deliverable	Achievements	Remaining To Do
A continual learning prototype software that ingests, and classifies digital informational objects like DNA sequences.	We have developed a C++ implementation of a continual learning prototype software.	develop the semi-supervise learning and novel classes detection feature.
Publications/patent/documentations that discusses the implementation and validation of the framework.	We have completed some experiments and evaluated the performance of our model.	Complete all experiments and write up the document.
Project report on Potential findings/discoveries in the database using the proposed algorithm.	We have gotten some preliminary results for further analysis	Complete all experiments and write up the document.