Analysis of Projects

Jerome Kovoor

***Abstract*—This document is an analysis of reports and posters on computational projects carried out by various authors**

I. Review of the projects (abstracts and posters)

A. Leveraging Spark and Docker for Scalable, Reproducible Analysis of Railroad Defects

Defects in the tracks are the second leading cause of accidents on railways. This project studied the feasibility of predicting the type of railway defect based on associated data such as geographic region, mean gross tonnage the track has undergone, and the rail type. They used data-driven techniques to predict defects. Two types of datasets were analyzed. One, concerning the defects in the rails themselves and the other one looking at the misalignment of track components. They used Apache Spark, its parallel machine learning library MLLib and Docker to predict defects in railroads. They found that the accuracy of the defect type prediction increased when there was a hierarchical classification of the defects. Also, training classifiers on data from geographically similar regions didn't give a significant improvement in the accuracy but grouping together the regions having similar defect distributions may find useful.

B. Clustering Temporal Gene Expressions of Iron-Oxidizing Zetaproteobacteria

Zetaproteobacteria, an iron oxidising bacteria is commonly found at deep sea hydrothermal vents. They play an importan role in the rusting of ship hulls, metal pilings, pipelines etc. Scientists want to understand the genes of this bacteria that are involved in the oxidization of iron. The authors of this paper used two clustering methods to quickly and effectively find similar gene expression patterns. The resulting clusterings can then be analyzed to find which one will be useful for the scientists which can save them time from otherwise clustering 2000 genes by hand. Their workflow consisted of data collection, normalization, cleaning and clustering. The data was taken from actual bacterial samples collected from Hawaii, which were then preserved at four different times to study the metabolism of iron. The data was normalized using a function called Reads Per Kilobase per Million mapped reads (RPKM). After normalization they used two different clustering methods namely, k-means clustering which favors spherical clusters of equal size and spectral clustering for non-spherical clusters of equal size. They were able to find a cluster with a large number of genes which the scientists have previously reported as interesting.

C. Using Machine Learning to Build a Scalable Tool to support Dietitians to Fight Chronic Diseases

In this project, the authors tried to develop a framework using machine learning to assist dietitians to identify patterns in nutrient intake of people from different age groups, demographics etc. as this can help in fighting chronic diseases. The data set was collected from nationwide surveys conducted by the National Center for Health Statistics and some other health agencies and it consisted of demographics, vital signs and nutrient intakes gathered from several people. The data underwent preprocessing, useful features were selected using MapReduce in Spark, used regularized linear regression, a combination of L1 and L2 regularization called Elastic net and then it was evaluated. They were able to relate some demographics features with the nutrient intake in individuals.

D. Validation of the Short Time-series Expression Miner (STEM) on Iron Cycling in a Shallow Alluvial Aquifer

In this project, the influence of the level of gene expression in an aquatic iron-oxidising bacteria called Zetaprotobacteria was studied. They are responsible for rusting of underwater structure that are made of iron but their metabolic mechanisms are not well understood. An important data which is required to study their mechanism is its gene expression level after initiating iron oxidation once it is exposed to iron, as a function of time. For this, a software developed by scientists at Carnegie Mellon, called the Short Time-series Expression Minder (STEM) was used. The metabolic study data of iron at various time intervals was obtained from a research paper by Jewell et al for different genes. They concluded that the program STEM allocates all known genes to statistically significant clusters.

E. Parameter Tuning of DBSCAN for Medical Data and Diabetes Diagnosis

This project is concerned with preparing a fast and efficient learning algorithm for processing medical data so that machine learning algorithms can analyze patient medical data and provide a diagnosis. The authors collected lab data on diabetes from the National Health and Nutrient Examination Survey. This data was then processed using parallel algorithms built using the MapReduce tool in Apache Spark. The DBSCAN clustering algorithm was then applied to this data inorder to diagnose patients with diabetes. Although, the analysis was not able to differentiate between patients with and without diabetes, their work was able to find optimal parameter setting in the medical data which can help in diagnosis.

II. Commonalities between the projects

All the projects have studied very influential topics, though in various fields. They all have utilized various map reduce techniques or learning algorithms to help scientists do their work efficiently and faster.

III. Abstracts vs. posters

Both the abstracts and the posters cover the topic by giving an introduction, motivation, data analysis procedure and the results. But the way it is represented is what makes a poster and an abstract different. Posters are visual, containing a lot of images and graphs whereas abstracts are more descriptive with words. Abstracts and projects should be considered as complementary. One of these is not enough to completely understand the project.

The pictures in the abstracts are mostly graphs and they are only to support the claims but in the posters, they are mainly for visual aid, to easily understand the project, though graphs are also required in the posters.

Abstracts should have more words because it is like a script, whereas a poster need not have a lot words, Also they ne