Deep Neural Networks for Analysing Cancer Genomics Data

Organizers: Fraunhofer FIT, Germany & RWTH Aachen University, Germany
Md. Rezaul Karim, Oya Beyan, Michael Cochez, Lars Gleim, Nils Lukas and Stefan Decker
Contact person: Md. Rezaul Karim (rezaul.karim@fit.fraunhofer.de)

Motivation

The aim of the hackathon is to explore application of deep neural networks on large scale health data sets. Although classical machine learning techniques allow researchers to identify groups (i.e. clusters) of related variables from data, the accuracy and effectiveness of these methods diminish for large and high-dimensional datasets.

On the other hand, deep neural network architectures (the core of deep learning) can better exploit large-scale datasets to build complex models. Deep learning algorithms need robust feature engineering since this dataset comes with high-dimensional with large samples size. Feature engineering techniques such as correlation matrix, recursive feature elimination, genetic algorithms, and statistical hypothesis tests can help us to deal with these challenges .

The main data source for this hackathon event would be The Cancer Genome Atlas (TCGA) project, which is available on the TCGA Research Network: http://cancergenome.nih.gov/. TCGA is a project that started in 2005 to catalog genetic mutations responsible for cancer using genome sequencing and bioinformatics.

Currently, TCGA has 39 projects, each corresponding to a certain type of cancer. Each cancer type data are high-dimensional and heterogeneous consisting of masked somatic mutation data, copy number segment masked copy number segments, gene expression quantifications, DNA methylations, miRNA expression, and clinical data. Moreover, the TCGA project collects both clinical data and biospecimen(s) from each patient.

Problem Definition

The scope of hackathon will be using robust feature selection and construction, deep neural networks and other machine learning algorithms to identify copy numbers associated with cancers. During the hackathon, the goal will be to solve three prediction tasks:

- **Cancer type detection:** Predict the type of cancers using copy number variation from genomics data and deep learning algorithms.
- **Predicting cancer subtypes:** Based on discrete type of data consisting of estrogen receptor (ER), progesterone receptor (PR), and HER2/neu status, predict cancer subtypes.
- **Predicting survival rate:** Based on continuous survival rate data, determine the future survival rate.

Tasks

Task 1: Cancer type detection

Problem statement: given the copy number variation data of a patient, the task is to identify whether that patient has cancer or not. If yes, then which type of cancer the patient has.

Dataset description: for this event, CNV data having both the normal tissue and cancer tissue samples from 15 different cancer type patients will be used. The input dataset has 569 features and 1 label column (see the data preprocessing document for more detailed information). Since this is al dataset having lots of features, participants should do the feature engineering carefully.

Model evaluation: You have to treat this as a multinomial (i.e. multi class) classification problem. Therefore, the solution should have some performance metrics to judge the model's performance such as accuracy, precision, recall and f1 measure.

Task 2: Predicting breast cancer subtypes

Problem statement: breast cancer patients can be categorized based on the status of three proteins in their body: estrogen receptor (ER), progesterone receptor (PGR), and HER2/neu. Each protein can be determined as "Positive" (if it exists inside patient's body), "Negative" (if it doesn't exist inside patient's body), or "Indeterminate". In this problem, we will make three predictions, each for each protein to determine their status (positive, negative, or indeterminate) based on their genomic data.

Dataset description: there are three types of genomic that can be used: DNA methylation data, gene expression data, and miRNA expression data. Participants can use either one of them or combination of them to make train a multimodal deep belief network (MDBN) to make predictions. This part consist of three different predictions, which are

- Estrogen receptor (ER) status prediction. There are three types of status, which are “Positive”, “Negative”, and “Indeterminate”.
- Progesterone receptor (PGR) status prediction. There are three types of status, which are “Positive”, “Negative”, and “Indeterminate”.
- HER2/neu status prediction. There are four types of status, which are “Positive”, “Negative”, “Equivocal”, and “Indeterminate”.

Model evaluation: You have to treat this as a multinomial (i.e. multi class) classification problem. Therefore, the solution should have the following performance metrics to judge the model's performance: Accuracy, Precision, Recall and F1 measure.

Task 3: Predicting breast cancer survival rate

Problem statement: each patient has a survival rate between 0 and 1 (i.e., continuous), which shows how likely they will survive, with 1 being the likeliest. In this part, we will make a prediction of a patient's survival rate.

Dataset description: there are three types of genomic data that can be used: DNA methylation data, gene expression data, and miRNA expression data. Participants can use either one of them or combination of them to make the predictive model by training a multimodal deep belief network (MDBN).

Model evaluation: You should solve this as a regression task. Therefore, the solution should have the following performance metrics to judge the model's performance: Mean square error (MSE), Root mean square error (RMSE), R2, Mean average error (MAE)

Using dimensionality reduction algorithms (e.g. PCA, SVD, Autoencoders, LDA) is recommended. Additionally, they can use automatic feature selection (e.g. Chi Square Selection, genetic algorithm) and use the most relevant features for making a prediction.

Machine learning/deep learning algorithms

Selection of the deep learning architecture will be based on the type of the data. Some of the most popular network architectures are Feedforward Neural Networks (FFNN, e.g., Multilayer Perceptron, Deep Belief Networks), Convolutional Neural Network (CNN), Long Short-Term Memory network (LSTM) and Autoencoders. In the literature, CNN is well known as an effective network to act as 'feature extractors' for imaging and computer vision, whereas LSTM is a type of powerful Recurrent Neural Networks (RNN) for modelling orderly sequence learning problems.

In our case, since the available input data are not in either imaging or sequence from (e.g., raw DNA sequences or protein sequences), it has been observed that using CNN or LSTM wouldn't be a viable solution since lots of additional steps are necessary to make them fit for these datasets.

Therefore, using FFNN such as Multiplayer Perceptron (MLP) or Deep Belief Network (DBN) with more hidden layers targeted for classification is encouraged. One other positive aspect of using DBN is that we can have a shared representation of features (i.e. multimodal deep belief network - MDBN) and do classification or regression analysis. Nevertheless, participants are free to use other classical machine learning models such as random forest, Support vector machines (SVM), decision trees, gradient boosted trees, logistic regression, etc.

Technologies

The following technologies can be used for the given tasks (we will provide some support with them) :

1. **Apache Spark:** see more at https://spark.apache.org/
2. **TensorFlow**: see more at https://www.tensorflow.org/
3. **H2O and Sparkling Water**: see more at https://www.h2o.ai/
4. **KNIME**: see more at https://www.knime.com/

Access to supplementary materials

Some supplementary resources including the datasets, their description, a sample solution (not hypertuned) will be made available during the hackathon.