Deep Learning Project Final Report

Team Members: Ryan Schumm and Theresa Sheets

Date: December 11 2019

Introduction

Cystic Fibrosis (CF) is a progressive genetic disease affecting over 70,000 people worldwide. Though much is known about the progression of the disease, it often progresses rapidly leaving those affected susceptible to opportunistic infection and bacterial colonization. Previous attempts have been made to predict 5 year survivorship of CF patients using traditional statistical measures. We will use a simple Recurrent Neural Net (RNN), a Multi-Layer Long Short-Term Memory RNN (LSTM), and a Fully Connected Multi-Layer Long Short-Term Memory RNN (FC-LSTM). These neural nets will be able to incorporate the temporal information in more complex ways than regression based methods.

Background

CF is caused by an incorrectly functioning protein that causes the mucus in the organs of those affected to become thick and sticky. This mucus builds up in the lungs and traps germs leading to infections and respiratory distress. Many individuals with CF also develop diabetes. While 9.4% of adults in the general population have diabetes, 40-50% of adults with CF develop diabetes.

In CF related diabetes, sticky mucus causes scarring of the pancreas. This prevents the pancreas from producing adequate amounts insulin. Many people with CF related diabetes do not know they have diabetes until they are tested, so part of the CF care guidelines recommend annual diabetes testing of CF patients. CF related diabetes can be well managed with insulin and blood sugar monitoring. An effective way of predicting diabetes diagnosis could help to identify individuals at highest risk and get them treatment more quickly. As CF is a progressive disease, it is important to be able to predict survivorship based on physiological measures. The goal of this project will be to identify the features most likely to predict 5 year survivorship of patients with CF and build a recurrent neural net to predict five year survivorship of diabetes patients.

Previous research has identified forced expiratory volume (FEV) as the most significant predictor of 5 year survivorship of CF patients. While FEV is an incredibly powerful tool to determine those who are most dangerously sick, it is a poor predictor amongst patients with high FEV. We hope using a RNN to analyze the data will help to parse the most powerful predictors amongst groups of individuals with high FEV and low FEV.

Each person in the data set corresponds to a unique observation with a set of physiological features measured for up to 60 time points serving as the information from which we predict the labels of the observations. The network will first be trained to perform a binary classification task corresponding to whether or not the person survived a five year time period. If the network is able to achieve a performance that exceeds that of past work, it will then be trained to predict whether the survivors were diagnosed with diabetes. In addition, variants of the standard recurrent neural network such as the long short-term memory networks will be implemented to see if superior performance can be obtained. The LSTM architecture introduces a cell state which can be modified at each time step by gates and carried through the network. The cell state allows the model to account for long term dependencies among the physiology input vectors better than the standard RNN architecture.

A previous model was developed on this dataset using logistic regression to predict five year survivorship. This project will use the RNN, LSTM, and FC-LSTMs to identify a classifier to predict five year survivorship. Then this model will be compared to the logistic regression based model to see if it is a better method to predict survivorship. Using this as a proof of concept we hope to eventually train new RNN, LSTM, and FC-LSTMs to predict diabetes diagnosis in a similar manner.

Data Set

The data which will be used to conduct this research are a set of 48,000 patients divided into 4 five year cohorts each with 60 time points in which they were assessed by a physician. These patients make up approximately 90% of those individuals diagnosed with diabetes since 1986. The physiological measures for each patient will be used as the input parameters, and the patient's survivorship and diabetic status after five years will be used as the labels for the data set.

For each of those individuals, we have basic physical and demographic information and specifically information on: 'mssa', 'mrsa', 'h_flu', 'pseudo', 'burkho_complex', 'alcalig', 'steno', 'enterobacter', 'serratia_marcescens', 'aspergillus', 'candida', 'scedosporium', 'mabscessus', 'mai', 'sex', 'suff', 'diabet', 'impglu', 'fev1pct_best', 'zscore_best', 'trunc03', 'dflag5'. Most of these parameters are binary classifications regarding the individual's colonization with a variety of pathogens while 'fev1pct_best' measures forced expiatory volume (related to lung capacity) and 'zscore_best' measures relative weight for age. Dflag5 is the binary survivorship label.

The original data set was made of two large data sets with a combined total of 81 parameters and 3,506,568 observations and multiple yearly entries for each individual. One of these data sets contains yearly bacterial cultures and for the patients while the other contains information collected from routine physician visits. We merged these two data sets by patient ID and year so that we were left with 2,483,622 observations of 81 features.

At this point there were no observations without an N/A in at least one of the features. Across the features there were many which were primarily missing data. Given that there are approximately 2.5 million observations, we chose to remove any feature with more than 20% missing data, so we set a cutoff of 500,000 N/As and removed any features with more missing observations than this. This resulted in a data set with 2,483,622 observations of 55 features. A majority of the removed features were associated with various bacterial colonization of patients, but we also removed the third mutation column which was primarily composed of missing data.

Our collaborators informed us that after 2002, the bacterial culture techniques used to diagnose bacterial colonization were changed. This intrigued us so we chose to split the data set into two, one a modern cohort of observations after 2002, and a total cohort including all the data to see if these modern techniques were more predictive than previous methods. There are 1,499,430 observations of 55 features in the modern cohort.

To address the remaining N/As within the data set we chose to create two different versions of each data set, one containing the N/As to be replaced with the average value of that feature, and one with every entry containing an N/A removed. This resulted in a N/A free total cohort of 1,728,227 observations and a N/A free modern cohort of 959,345 observations.

Methods

The data were split into testing and training sets with 80% of the unique patient IDs assigned to training, while 20% were assigned to testing.

The original attempts to train the networks failed because the networks would learn to predict survival for every observation. About 73% of the participants survived at least five years, so the accuracy was high but there were very few true positives. To correct this , the individuals in the training data who did not survive five years were drawn at random with replacement until the resulting training data had 50% of the participants not surviving. The data used to test the network was unaltered.

We built a simple RNN to see if a network could effectively learn on this data set. The network consists of a single vanilla RNN layer with two outputs corresponding to the class scores.

We next trained the LSTM architecture consisting of three RNN layers with LSTM cells. The outputs of the first two layers are sequences of vectors of size 100. The final RNN layer outputs a single vector of size two corresponding to the class scores. The FC-LSTM architecture consists of two RNN layers with

LSTM cells. The first layer outputs a sequence of vectors of size 500. The second layer processes the vector sequence and outputs a single vector also of size 500. This vector is then passed to a fully-connected network with one hidden layer of size 500 with ReLU activations. The output space of the fully-connected network has a size of 2 corresponding to the class scores. All three of these architectures are trained and compared in notebook 1. In notebooks 2 and 3, the FC-LSTM network is trained, optimized, and validated using the different constructions of the original dataset.

Results

1. **RNN** Since the data is unbalanced, both the true positive and true negative rates were used as performance metrics in addition to accuracy. The simple RNN was able to obtain about 80% accuracy, precision, and true negative rates after 1 epoch.

2. **LSTM** The performance metrics improved by about 6% with the multi-layer LSTM architecture and the convergence rate was higher.

3. **FC-LSTM** Like the multi-layer LSTM, the model converged rapidly. It had marginal performance improvements over the LSTM network with no FC layers.

4. **Dataset Manipulations** The addition of 80% complete features did not significantly improve performance nor did restricting the time period to observations made no latter than 2002.

5. **Final Training and Evaluation** The modern data was used to tune the learning rate and for the final training and evaluation. A learning rate of 1e-3 was found to be optimal. The testing data was split into a validation set used at training time and a testing set used for the final evaluation. The model achieved a 85% accuracy, a 90% true positive rate, and a 84% true negative rate on the testing set.

Conclusions

A network with an architecture consisting of two LSTM RNN layers and a fully-connected region with one hidden layer was found to both converge most rapidly and perform the best out of the simpler architectures sampled. The network was capable of predicting the mortality of cystic-fibrosis patients with an accuracy of 85%. In addition, it was found that using only the modern observations of the patients physiological states yielded equal performance to using the entire time period over which the patients were tracked.

References

The following papers will be read to obtain background information for solving the problem:

1. Zachary C. Lipton, John Berkowitz, Charles Elkan: A Critical Review of Recurrent Neural Networks for Sequence Learning

2. Felix A. Gers, Nicol N. Schraudolph, and Jurgen Schmidhuber: Learning Precise Timing with LSTM Recurrent Networks

3. Theodore G. Liou, Frederick R. Adler, Stacey C. FitzSimmons, Barbara C. Cahill, Jonathan R. Hibbs, Bruce C. Marshall, Predictive 5-Year Survivorship Model of Cystic Fibrosis, American Journal of Epidemiology, Volume 153, Issue 4, 15 February 2001, Pages 345–352, https://doi.org/10.1093/aje/153.4.345

4. Adler, F. R., & Liou, T. G. (2016). The Dynamics of Disease Progression in Cystic Fibrosis. Plos One, 11(6). doi: 10.1371/journal.pone.0156752