Jeffrey Bailey

Craun Intro to Research

5/15/2018

**Background Research on Machine Learning Applied to Heart Attack Prediction**

**Background Knowledge**

The term cardiovascular disease (commonly abbreviated to CVD) refers to a family of diseases that affect the body’s cardiovascular system. A noteworthy example of cardiovascular disease is myocardial infarction, also known as a heart attack, which is an event in which heart muscle dies due to a lack of oxygen, and blood is not successfully pumped through the veins throughout the body. This is caused by coronary artery disease, a form of atherosclerosis that occurs in the coronary arteries. Coronary arteries carry oxygen-poor blood into the heart from the rest of the body. Atherosclerosis is a condition in which plaques, usually made of fats and cholesterols, build up on the walls of arteries and clog the flow of blood through them. High cholesterol, high blood pressure, old age, and diabetes are all risk factors of atherosclerosis and heart attack (Heart Attack, 2018).

Symptoms of myocardial infarction develop over several hours after the artery blockage becomes severe enough, and are deadly in some cases. An estimated 790,000 Americans have a heart attack annually, and cardiovascular disease is the number one leading cause of death in America (Heart Disease Facts, 2017). However, early detection and diagnosis of conditions which cause heart attacks can be life-saving. Many cardiovascular diseases, including coronary artery disease, can be detected by use of an electrocardiogram. An electrocardiogram (abbreviated to ECG) is a device which reads and records electrical signals from the heart. The signals are then drawn by the machine onto a line graph to be analyzed by a physician. Because the heart’s electrical signals are what causes the muscles to move and the heart to beat, ECG signals make a visualization of the heart’s beating rhythm. Thus, electrocardiogram tests can be used to detect most irregularities in a heart’s beating, including artery blockage. Electrocardiograms can also detect if a patient has had a heart attack in the past, which can be used in a prediction of future heart attacks (Heart Attack, 2018).

Machine learning is a branch of data science that deals with algorithms that are designed to adapt to data and “learn” in a way mimicking a human. Such algorithms can be applied to data sets where the relationships between variables are not entirely known or are so complex that humans may have trouble analyzing them. Application of such algorithms can make tasks involving these datasets much more efficient. Datasets used in machine learning generally include many items, each of which have many attributes. The goal of the algorithm is to group items based on their qualities. There are many different kinds of machine learning algorithm, but they all fall into two major categories; supervised or unsupervised. Supervised algorithms are ones where the desired categorization or output from the dataset is known and used to optimize the algorithms prediction using various mathematical functions. Unsupervised algorithms do not have a desired output in the dataset, and the algorithm is tasked with categorizing the data based on its attributes and not by its predicted output or existing categorization. Unsupervised algorithms make their own categories, while supervised algorithms optimize themselves to sort items into existing categories. This research will focus on supervised learning as the desired output, if a person is at risk of having a heart attack, is included in the data set and can be used to train the algorithm.

The dataset for a supervised learning algorithm consists of many items (people, for example) and attributes (age, blood pressure, weight are all examples). Each item has a set of values for each attribute, and one or more of the attributes are to be predicted by the algorithm. Before research, datasets are separated into two sets, the training set and the testing set. The training set consists of the majority of the data and is used during the algorithm’s training. The testing set is used after training, and when it is used with the algorithm the output is not given and the algorithm categorizes the items on its own. These predicted categorizations of the testing set are then compared with the actual outputs from the original dataset for statistical analysis of the algorithm’s effectiveness. After training, the algorithm can attempt to categorize new items based on the variable attributes into the desired categories, which in this example is if the person will have a heart attack in the future. A common issue relating to the application of machine learning models is overfitting, which is when the amount of items is too small, or the number of attributes is too large, and the model makes categorizations based on the specific training items and not on the trends in attributes through the whole dataset. This leads the algorithm to become highly accurate in categorization of the training set, however much less effective at categorization of items inputted after training (Hinton, G., Krizheysky, A., Salakhutdinov, R., Srivastava, N., & Sutskever, I., 2014).

**Current Research**

Neural networks and machine learning in general have recently developed a large role in medical research. As clinical data includes many different variables and values to analyze, use of machine learning is very applicable (Havel, J., Lopez-Rodriguez, A., Pena-Mendez, E., & Vanhara, P, 2016). Machine learning models have already been used in many medical fields. One study done in 2017 by developed a neural network algorithm that could efficiently diagnose various types of pneumonia from a chest x-ray. The input x-rays were entirely black and white, and the algorithm used them to produce colored charts showing fluid buildups, making diagnosis much easier. This research demonstrated neural networks’ strong ability to recognize and manipulate images into a much easier to use form (Irvin, J., & Rajpurkar, P., 2017).

Another study done in 2017 attempted to create an algorithm to effectively predict heart attack risk based on clinical data. Researchers constructed many different supervised algorithms using the clinical data and risk factors as the input, and if the patient had a heart attack in the following ten years as the output (in a true/false statement). Input data included a patient’s age, body-mass-index, blood pressure, cholesterol levels, gender, ethnicity, history of diabetes, and drug use. Algorithms constructed included a random forest, logistic regression, gradient boosting algorithm, and a neural network. The purpose of the research was to develop a program using machine learning to assess a person’s heart attack risk that could perform better than a widely used pre-existing model that does not use machine learning made by the American College of Cardiology (ACC). All four of the machine learning models performed better than the ACC model, having more correct predictions and fewer incorrect predictions. The best performing model was the neural network, which correctly predicted 67.5% of heart attack cases, 4.8% more correct predictions than the ACC model (Garibaldi, J. M., Kai, J., Qureshi, N., Reps, J., & Weng, S., 2017).

A second study done in 2017 developed a model to diagnose arrythmia (irregularities in the heart’s rhythm) using a supervised neural network. The input consisted of hundreds of samples of electrocardiograms from patients who had one of fourteen kinds of arrythmia, and the algorithm was tasked with categorizing the electrocardiograms by arrythmia type, effectively creating a model that could diagnose a patient’s arrythmia based on an inputted ECG reading. A test set was then made with confirmed diagnoses, and a small team of professional cardiologists was tasked with diagnosing each of the ECGs in the test set. The same test set was then run through the neural network, and it was found that the neural network model had a higher rate of correct diagnosis on twelve of the fourteen arrythmia types. This study showed that machine learning algorithms can effectively read electrocardiogram signals and make a diagnosis based on them, and that such diagnoses can rival those made by human cardiologists (Bourn, C., Haghpanahi M., Hannun, A. Y., Ng, A. Y., Rajpurkar P., 2017).

**How My Research is Different**

While the second study previously discussed effectively developed models to predict or diagnose cardiovascular disease, the most accurate model (the neural network) developed in the second study only correctly predicted 67.5% of heart attack cases. This model, however, did not include diagnosis of common heart attack causing diseases in the input, many of which can be diagnosed using an electrocardiogram. The arrythmia study supported the conclusion that machine learning models can effectively diagnose diseases detectible with an ECG. Therefore, it is hypothesized that if a machine learning model is made and trained from an input including both clinical data (similar to that in Weng’s research) and cycles from an ECG test, the prediction of if a person will develop a heart attack will be more accurate than both the existing ACC model and those developed by Weng. Also, Weng’s input set had a large number of input features, which potentially could cause overfitting and lower the algorithms accuracy of prediction outside of the training set. Removing some features that were shown to not have much weight in the developed algorithms could increase the algorithm’s accuracy, as there is a lower chance of overfitting during training. It is therefore quite possible that combining ECG signals with a smaller set of clinical data regarding a patient could result in a model that can predict heart attack more accurately than those created by the ACC and Weng’s team.

**References**

Bourn, C., Haghpanahi M., Hannun, A. Y., Ng, A. Y., Rajpurkar P. (2017). Cardiologist- Level Arrhythmia Prediction with Convolution Neural Networks. Retrieved from https://arxiv.org/pdf/1707.01836.pdf

Garibaldi, J. M., Kai, J., Qureshi, N., Reps, J., & Weng, S. (2017). Can Machine Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?. Retrieved from http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0174944&type= printable

Havel, J., Lopez-Rodriguez, A., Pena-Mendez, E., & Vanhara, P. (2016, January 1.). Artificial Neural ….. . Networks in Medical Diagnosis. Retrieved from

https://www.researchgate.net/publication/250310836\_Artificial\_neural\_networks\_in\_medical\_diagnosis?enrichId=rgreq-5345aec0d0963c4404b1dda6fb416a77-XXX&enrichSource=Y292ZXJQYWdlOzI1MDMxMDgzNjtBUzozMTgxMjAyNjk0MTg0OThAMTQ1Mjg1NzE5MDcyOA%3D%3D&el=1\_x\_3&\_esc=publicationCoverPdf

Heart Attack. (2018, March 10). Retrieved from https://www.mayoclinic.org/diseases-conditions/heart- attack/symptoms-causes/syc-20373106

Heart Disease Facts. (2017). Retrieved from https://www.cdc.gov/heartdisease/facts.htm

Hinton, G., Krizheysky, A., Salakhutdinov, R., Srivastava, N., & Sutskever, I. (2014, June 24.). Dropout: A Simple Way to Prevent Neural Networks From Overfitting. Retrieved from http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm\_content=buffer79b43&utm\_medium=social&utm\_source=twitter.com&utm\_campaign=buffer

Irvin, J., & Rajpurkar, P. (2017, December 25). CheXNet: Radiologist-Level Pneumonia Detection on

Chest X-Rays with Deep Learning. Retrieved from https://arxiv.org/pdf/1711.05225.pdf