INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 1)
Available online at www.ijariit.com

Automatic Feature Selection from EHR & DNN Modeling

Shreyal Gajare
Maharashtra Institute of Technology, Pune, Maharashtra
shreyalgajare@gmail.com

Shilpa Sonawani
Maharashtra Institute of Technology, Pune, Maharashtra
shilpasonwani@mit.edu.in

Abstract: Recently there are a lot of advancements in healthcare technology. Amongst, Electronic Health Record (EHR) is an upcoming trend which stores patients' demographics, lab tests & results, medical history, habits etc. collaborated in electronic form. EHR is huge data, which is difficult to maintain and retrieve. So the idea of health risk prediction is formulated in this work. To get the relevant data from EHR, feature selection technique is used. Feature selection is responsible to collect only important and needed data from the dataset. For feature selection regression method is used in which loss function is proposed due to which accuracy and performance of the model are increased. Further risk prediction is done using neural network model. Deep Neural Network (DNN) is best suited for pattern learning and prediction purpose. It consists of various layers which have their specific function. DNN uses transfer learning to avoid repeated training for the whole system. Dataset considered here is of hypertension. EHR data is also synthetically created for analysis.

Keywords: Electronic Health Record (EHR), Feature Selection, Regression technique, Deep Neural Network (DNN).

INTRODUCTION

In today's world, healthcare is on critical stage to achieve better, cheaper and safer health that offers improved results with speedy & efficient treatment. In the growing population, people have the desire to live longer and healthier life to avoid preventable death. Digitization is growing enormously with the advent of machines. Thus Electronic Health Record (EHR) came into existence. EHR is a digital version of patient's record chart. They are real time records available instantly and securely for authorized users. It includes patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results. It can also allow access to evidence-based tools that providers can use to make decisions about a patient's care.

EHR contains a huge amount of data so it is necessary to store it in the proper format for further processing. To extract the needed data from such large dataset is a challenge. A lot of research is going on in this area. Here feature selection is the method used to select appropriate features from the dataset. There exist three main methods for feature selection like filter methods, wrapper methods, and embedded method. The motivation behind using feature selection is general data reduction, feature set reduction, performance improvement and data understanding.

Due to the advent of technology, machine learning is widely adopted in the healthcare industry. Deep Learning applies a set of machine learning algorithms at multiple levels for learning and prediction purpose. Deep Neural Networks (DNN) is used as a risk prediction model which can be extended with many hidden layers being an iterative training process. The network weights can be adjusted by minimizing the difference between the network outputs and the desired outputs. Thus the accuracy of the model can be increased by adopting this method for predictive analytics. Other approaches using statistical techniques do not support incremental learning as the neural networks do.

RELATED WORK

Deep Learning has been very famous as a lot of research is going on over it. The reason to implement Feature selection before applying it to the neural network model is to increase the performance and response time of the system.

The deep model that simulated the thinking procedure of people and combine feature representation and learning in a unified model was given. A modified version of convolutional deep belief networks is used as an effective training method for large-scale data sets. The restricted Boltzmann machine was used for learning from electronic medical records [3]. But it could not optimize the parameter for large datasets. In [8] diabetic detection method using ANN and a feature set formed by adopting singular value decomposition (SVD) and Principle Component Analysis (PCA) has been proposed. But the drawback was the learning process was very slow. G. Canino & Q. Suo designed a workflow based model for analyzing biological values. The system is able to relate biological data to diagnosis codes and with additional information integrated and correlated to EMRs data. They used Recursive Feature Elimination along with sparse logistic regression which could not focus on loss function as well as regularization parameter. David Barber, Ken Martin & Farhana Zulkernine have suggested Artificial Neural Network (ANN) approach to capture the patient's data and predict the risk faced by them. But it was restricted to only a few parameters and was inefficient due to limited number of layers in neural network.

METHODOLOGY

In this work, an approach is made to enhance the performance, speed, and accuracy of the model. Thus feature selection methodology is applied on EHR dataset. Automatically the irrelevant and noisy data will be removed before passing it to the prediction model.

FEATURE SELECTION

Feature selection enables selection of relevant parameters to find the dependency among the variables. It can reduce the complexity of the problem to a large extent. In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, [3] is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Feature selection techniques are used for four reasons:
- Simplication of models to make them easier to interpret by researchers/users
- Shorter training times
- To avoid the curse of dimensionality
- Enhanced generalization by reducing over fitting

After feature selection is performed, the selected features are stored in vector form. The attributes selected possess the score based on their value. Hence to store the data in the proper format before providing it to the neural network is stored in feature vector form known as representation learning.

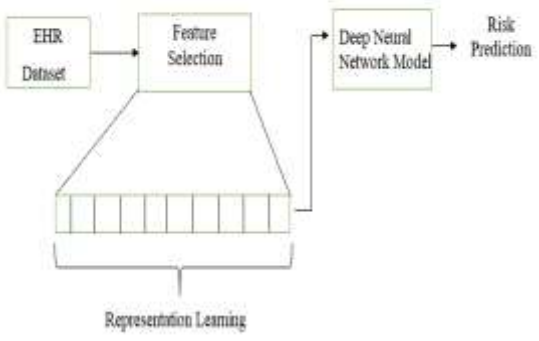
DEEP NEURAL NETWORK

Nowadays deep learning majorly makes use of neural networks in its perspective. Neural networks serve a huge purpose in the area of machine learning. Deep-learning methods are part of distributed representation-learning algorithms that attempt to extract and organize discriminative information from the data by discovering features that compose multi-level distributed representations. Deep neural networks are typically trained, by updating and adjusting neurons weights and biases, utilizing the supervised learning back propagation algorithm in conjunction with optimization technique such as stochastic gradient descent.

Transfer learning is used in many areas of machine learning without retraining the whole system. For eg. One can get pre-trained network and add an extra classifier on top and train only that classifier on new training samples while keeping the weights fixed. DNN can be used for transfer learning to keep their training session as simple as possible without the need to train the whole system.

PROPOSED SYSTEM ARCHITECTURE

In healthcare industry, it is necessary to provide timely treatment to the patient. Similarly, it is also possible to decide the risk of patient's health condition depending on his/her medical history, current state, and other health parameters. For that purpose, a model which can predict the risk of patient's illness is developed with the help of neural networks.



Feature Selection is a technique used for selecting the subset of relevant features used for model construction. Amongst the available methods of

feature selection regression is well suited for hypertension problem because the requirement is to evaluate the risky attribute of the patient. When there are multiple co-related features involved, the model becomes unstable due to minor changes in the data and can cause large changes in the model. Thus regression method is used for selecting the subset of features for hypertension parameters. Wrapper methods are expensive to use for large feature space because of high computational cost and each feature set must be evaluated with the trained classifier that ultimately makes feature selection process slow. Hence embedded methods are a recently developed approach which utilizes the advantages of both the methods using the independent test as well as performance evaluation function tests.

REGRESSION

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. Various types of regression are linear, polynomial, stepwise, logistic regression etc. Thus embedded methods consist of two types of regression techniques LASSO (L1) and RIDGE (L2) regression. Recursive Feature Elimination (RFE) is also used along with L1 regression but in terms of Area under Curve (AUC) parameter L1 is more stable and accurate.

ALGORITHM

Algorithm - Sparse Logistic Regression
Input: EHR Dataset D
Output: feature ranking
for $\lambda_i \in \lambda$ (i=1.....n) do
train sparse logistic regression on D

$$P(y|x) = \frac{1}{1+e(-y(w^T)x+b))}$$

Minimize loss function,

$LL(w, b) = \frac{1}{n}\sum_{j=1}^{n} \log(1 + \exp(-y_i(w^T x_i + b)$

Replace loss function with logistic loss

Minimize $\sum_{xy} log(1+\exp(-w^T x.y)) + \lambda w^T w$

Evaluate model performance

Calculate weight W_i^k for each feature k

end for Calculate feature score $S_k = \frac{1}{n}\sum_i^n W_i^k$

Determine appropriate number of remaining features

Hence using sparse logistic regression (L1) method, features are selected appropriately from EHR record, related to hypertension. Generally, in other methods of logistic regression, the only estimator is trained and feature score is evaluated. Rather in this work, after training the dataset loss function is replaced with logistic loss function. Thus it helps to improve the accuracy and performance of the model.

In DNN architecture, it comprises of 11 input layers for the attributes selected from EHR for hypertension, 4 hidden layers, and 3 output layers.

The input for DNN will be the feature score i.e feature vector formed from attribute selection. As the training progresses, network weights are adjusted in multiple iterations to align the outputs. Until the error is reduced to an acceptable level or a threshold number of iterations is reached. So, in each iteration of the input vector, the error (*e*) is calculated from the difference between the desired (d_p) and obtained (y_p) output values.

$$e = (d_p - y_p)$$

The back propagation algorithm tries to minimize the error by following the downward slope of sum squared error value, which is called the gradient descent approach

$$\sum_p (dp - y_p)^2$$

The sigmoid function is used to calculate the final output in the form of probability,

$$S = \frac{1}{1+e^{(-m*net)}}$$

Where net is the total input signal of node, m is the slope i.e m=1

Thus sigmoid function evaluates whether the patient is normal, high risk or critical condition of hypertension in the form of probability.

CONCLUSION

Although different feature selection methods are available for selecting the subset of specific attributes, it needs to be checked thoroughly with various measures. The objective of feature selection is to reduce feature space. Thus L1 logistic regression is well suited for feature selection with sparse features selected. The selected features are stored in vector format with their ranks. So that DNN model can take appropriate parameters for risk prediction of hypertension. In DNN, due to an improved technique of Transfer learning pattern recognition is done. Hence DNN can predict the risky parameters with more accuracy.

FUTURE SCOPE

In future, the aim is to provide more accuracy and efficiency to the model by collating L1 and L2 regression methods. As L1 is good for regularization and L2 for loss function they can be combined for the betterment of the system. The linear combination of L1 and L2 is elastic net regression which can be used in further work. Present work gives the overview for risk prediction of hypertension, similarly, by using multi class prediction, it can be extended for all 6 to 7 types of diseases.

REFERENCES

1. Daniele R, C. Wong, M. Berthelot, "Deep Learning for Health Informatics", Journal of Biomedical and Health Informatics, 10.1109/JBHI.,IEEE 2016.
2. B. Shickel, Patrick Tighe, A Bihorac, Parisa Rashidi, "Deep EHR: A Survey of Recent Advances on Deep Learning Techniques for Electronic Health Record (EHR) Analysis, Journal of Bioinformatics,10.1109/JBHI.2767063, IEEE, 2017
3. G. Canino, Q. Suo, P.Guzzi,"Feature Selection Model for Diagnosis, Electronic Medical Records and Geographical Data Correlation", DOI.10.1145, 2975167, ACM, 2014
4. Fei Wang, P. Zang, X. Wang, "Clinical Risk Prediction with Multi linear Sparse Logistic Regression ", KDD DOI. 10.1145/263330, ACM, 2016
5. T. Tran, Richard Kennedy, Ann Larkins, "A framework for feature extraction from hospital medical data with applications in risk prediction", BMC Bioinformatics, DOI.10.1186/s. 2015.
6. Elyne Scheurwegs, Boris Cule, Kim Luyckx,"Selecting Relevant features from the electronic health record for clinical code prediction", Journal of Biomedical Informatics, Elsevier, 2017
7. Alexios Koutsoukas, Keith J. Monaghan, Xiaoli Li and Jun Huan,"Deep learning: investigating deep neural networks hyper parameters and comparison of performance to shallow methods for modeling bioactivity data", Journal of Cheminformatics, DOI 10.1186/s13321-017-0226-y, Research Gate, 2017
8. Trang Pham, Truyen Tran, D. Phung, "Deep Care: A Deep Dynamic Memory Model for Predictive Medicine, 10.1007/978-3-319-31750-23 Springer, 2016.
9. David Barbar, Ken Martin, "Using Machine Learning to Predict Hypertension from Clinical Dataset", 7840087/ 7849361/07849886, IEEE, 2017.
10. Riccardo Miatto, Brian Kidd, Joel Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from Electronic Health Record, Scientific Reports, DOI 10.1038, IEEE, 2016.
11. Sarfarez Ahmed, "Medical Diagnosis using Neural Networks", Conference Proceeding (NCETSE), ISSN: 2321-9939, IJRED 2014.
12. Jing Zhao, Lars Asker, Henrik Bostrom,"Learning from Heterogeneous temporal data in electronic health records, Journal of Biomedical Informatics , DOI. 10.1016, Elsevier, 2016.
13 Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe,"Deepr: A Convolutional Net for Medical Records", IEEE Journal of Biomedical and Health Informatics, DOI 10.1109/JBHI.2633963,2016

