International Conference on Recent Innovations in Science, Engineering & Management
R.R. INSTITUTE OF TECHNOLOGY, Bengaluru (India)
10th - 11th May 2019 www.conferenceworld.in ISBN : 978-93-87793-87-3

Predicting Multiple Diseases Using Machine Learning Techniques

Prof. Shruthi S[1],Baba Bharath G K[2],Vibha Reddy K V[3],Vinutha H[4],Nagaraj T[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, RRIT,(India)*

[2,3,4,5] *UG Students, Department of Computer Science and Engineering, RRIT,(India)*

Abstract

One of the challenges of healthcare outcome is aggregating disparate which represents large amount of asynchronous data source into meaningful indicators of each person. However in this paper we demonstrate that XGBoost classifier is more interpretable than the deep learning models for disease prediction. Additionally, in our work we provide survey of various machine learning algorithm such as Naive Bayes and KNN for disease prediction tasks.

Keywords: Machine learning, Naïve Bayes, KNN and Disease prediction.

I.INTRODUCTION

Nowadays approximately 7.5 millions of people are suffering from health issues such as heart disease, diabetes and chronical kidney diseases etc this has been caused many to consider the possibilities of designing disease detection systems and automated clinical support system. In the past studies, patient laboratory tests, diagnoses and medication were used to predict the diseases. Using this type of model it improves sensitivity, specificity of detecting the disease and to identify potentially unknown risk factors. By using the various methods such as Support Vector, random forests, neural network, time series modelling and logistic regression techniques the diseases have been predicted in recent studies. Deep learning methods (artificial neural network) have been successful in offering insight to both diagnosis and data representation. In our case we are using Naïve Bayes and KNN (K- Nearest neighbour) algorithm for predicting the diseases. The objective of this work is to predict the diagnosis of different type of disease.

II. LITERATURE SURVEY

[1] D. Kartchner, T. Christensen, J. Humpherys, and S. Wade in 2017 IEEE International Conference on Healthcare Informatics (ICHI). "Code2vec: Embedding and clustering medical diagnosis data," Identifying disease comorbidities and grouping medical diagnoses into disease incidents are two important problems in health care delivery and assessment.Using vector space embedding's produced using the Global Vectors (GloVe) algorithm, it is able to find useful vector representations of diagnosis codes that can identify related diagnoses and thus improve identification of related disease incidents.

[2] Y. Choi, C. Y.-I. Chiu, and D. Sontag Learning Low-Dimensional representations of Medical Concepts In year 2016**,** Y. Choi, C. Y.-I. Chiu, and D. Sontag, **"**Learning low dimensional representations of Medical Concepts**",** showed how to learn low-dimensional representations of a wide range of concepts in medicine, including diseases, medications, procedures, and laboratory tests. It expects that these embedding's will be useful across medical informatics for tasks such as cohort selection and patient summarization. These embedding's are learned using a technique called neural language modelling from the natural language processing community. However, rather than learning the embedding's solely from text, we show how to learn the embedding's from claims data, which is widely available both to providers and to payers.

[3] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor AI: Predicting clinical events via recurrent neural networks In year 2016, E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, **"**Doctor AI: Predicting clinical events via recurrent neural networks**",** Leveraging large historical data in electronic health record (EHR), developed Doctor AI, a generic predictive model that covers observed medical conditions and medication uses. Doctor AI is a temporal model using recurrent neural networks (RNN) and was developed and applied to longitudinal time stamped EHR data from 260K patients and 2,128 physicians over 8 years.

[4]N. Razavian and D. Sontag Temporal convolutional neural networks for diagnosis from lab tests. In 2015, N. Razavian and D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests", Early diagnosis of treatable diseases is essential for improving healthcare and many diseases onsets are predictable from annual lab tests and their temporal trends. It has introduced a multi-resolution convolutional neural network for early detection of multiple diseases from irregularly measured sparse lab values.

[5] V. Lebedev, E. Westman, G. J. P. Van Westen, M. G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, and A. Simmons."Random Forest ensembles for detection and prediction of Alzheimer's. Computer-aided diagnosis of Alzheimer's disease (AD) is a rapidly developing field of neuro imaging with strong potential to be used in practice. In this context, assessment of models' robustness to noise and imaging protocol differences together with post-processing and tuning strategies are key tasks to be addressed in order to move towards successful clinical applications. In this study, investigated the efficiency of Random Forest classifiers trained using different structural MRI measures, with and without neuroanatomical constraints in the detection and prediction of AD in terms of accuracy and between-cohort robustness.

Structural 1.5-T MRI-scans were processed using free surfer segmentation and cortical reconstruction. Using the resulting output, AD/HC classifiers were trained. Training included model tuning and performance assessment using out of bag estimation. Subsequently the classifiers were validated on the AD/HC test set and for the ability to predict MCI-to-AD conversion. Models between cohort robustness was additionally assessed using the Add Neuro Med dataset acquired with harmonized clinical and imaging protocols.

III. PROBLEM STATEMENT

Many of the machine learning based techniques are help us to detect the disease with the help of collected patient's health record. But prediction of the disease before its occurrence will be the better solution to improve the health status of the patient.

s3.1 EXISTING SYSTEM

Various methods are presents for prediction of disease such as support vector machines, logistic regression, random forests, neural networks and time series. A predictive model for progression of chronic kidney disease to kidney failure is an existing system which predicts the possibility of the all kind of Chronicle disease. In this model statistical analysis model is used to predict the CKD by analysing the factors such as demo graphic variables (age and sex), physical examination variables (bp and weight) comorbid conditions, including diabetes, hypertension, and aetiology of kidney disease; and laboratory variables from serum and urine collected at the initial nephrology visit.

Drawback

- Statistical analysis process is a time consuming task.
- This model designed for identification of multiple disease.

3.2 PROPOSED SYSTEM

In this approach it is aimed to develop a disease prediction model which identifies the multiple disease possibility by analysing the health record of the patient. XGBoost classifier is used to predict the diseases. This proposed model is trained for identification of heart failure (HF), type II diabetes mellitus (DM) and chronic kidney disease (CKD). Database of affected patient's health is collected and will be compared with current patient health data to identify the probability of the disease.

Advantages

- Multiple disease prediction.
- Execution time is less.

IV. METHODOLOGIES

1. KNN [K- Nearest Neighbour]

KNN can be non-parametric method used for both classification and regression predictive problems. The output depends on whether *k*-NN is used for classification or regression. However, it is more widely used in classification problems in the industry. To evaluate any techniques we generally look at 3 important aspects:

1. Ease to interpret output
2. Calculation time
3. Predictive Power

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

Algorithm

A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.

Euclidean:

$$d(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$$

Manhattan/city-block:

$$d(x,y)=\sum_{i=1}^{m}|x_i-y_i|$$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

2. Naïve Bayer's Classification

Naïve Bayer's classifier is the classification algorithm based on the Baye's theorem of probability. Given the set of attributes, posterior probability of the event is calculated using the Baye's theorem.

Posterior probabilityequation is given by ***Posterior(potable)=P(potable)P(potable|ph)P(potable|ec)…/evidence)***

Evidence is the sum of numerators in Posterior(potable) and Posterior(non-potable).

Given the Standard deviation σ, and mean μ, of the parameter p, calculation of P(potable | p) can be done as,

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i$$ Mean

$$\sigma=\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2\right]^{0.5}$$ Standard deviation

$$p(potable\mid p)=\frac{1}{\sqrt{2\pi\sigma\text{^}2}}\text{^}\left(\frac{-(val-\mu)\text{^}2}{2\sigma\text{^}2}\right)$$

V. SYSTEM DESIGN

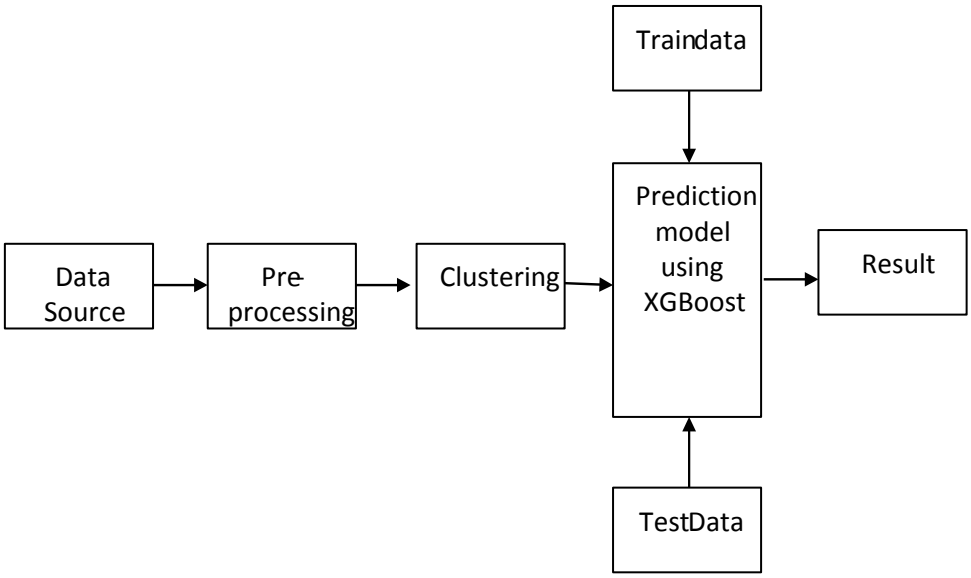


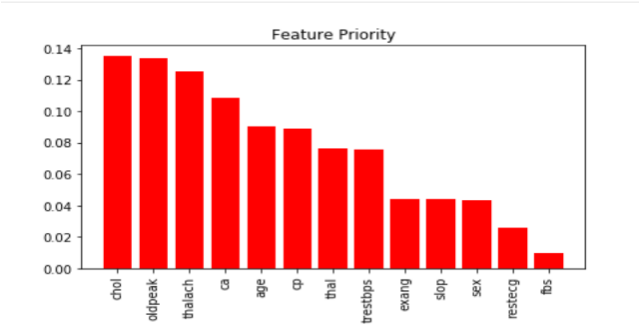
Fig 1. Data Flow Diagram

In the Data Source, the information about the patient's history of multiple diseases will be stored. The pre-processing is done to remove the unwanted data from the data source if it exists.After pre-processing clustering of the dataset based on the parameters of the disease will be processed. Total number of clusters will be identified as available number of disease count in the data source. Later testing phase started with user's health parameter input to identify the disease presence in his health. Given health parameter will be compared with existing trained set of data source using machine algorithm's such as Naïve bayes and KNN algorithms. Based on the result of ML algorithms decision on the health status will be takes place.

VI. RESULT

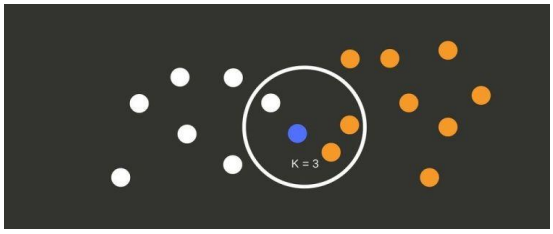
The result of our experiment shows that which is the best algorithm for predicting disease for multiple datasets. And KNN algorithm gives the best performance based on the raw diagnosis, procedure. Moreover, in the part of implementation it takes the dataset as single data point for different clusters. Then which values are near to the data point those points takes as cluster. Then we tried with the naïve Bayesian algorithm here the dataset values are taken as mean and variance so by probability of the training data set compare with the given resultant values of mean and variance. Here in this experiment we have taken XGBoostclassifier for disease predictions.XGBoost is a highly interpretable model which has the ability to give the result about how and why it makes the prediction. In this implementation the gradient boost decision tree is designed for speed performance. Here all of the three datasets are predicted with the different clustering and those points are classified with the KNN algorithm, this completely indicates the disease occurrence.

Now the dataset we are use in this prediction of 14 parameters.



Here the cluster k=3 we taken as level third clustering of the matching ratio method.

And the similar way we are taking k=1, k=2 and k=3



Here we are selecting the 3 different points of data from the individual disease. We got the accuracy level above 50%. This rate will be predicted from 3 dataset samples.

VII. CONCLUSION

This paper addresses the prediction of multiple datasets. It was very interesting to see different datasets (ie. Heart disease, Diabetes, Chronic kidney disease) to identify disease in an individual health issue. So we used KNN to avoid misclassification rate. By feature selection measures to select small featuresthat increase the classification performance. From the simulation result we applied the XGBoost classifier, here we are going with the matching ratio method. The XGBoost gives comparison of KNN algorithmand Naïve Bayesian algorithm. The XGBoost provides the comparison of the testing datasets and training datasets through the algorithms clustering process. KNN clustering that gives the result on the basis of individual parameter of the disease compares with the patient health record, and these all takes on different K-Values as different stages. Then prediction of the disease is shown by comparing values.

REFERENCES

[1] D. Kartchner, T. Christensen, J. Humpherys, and S. Wade, “Code2vec: Embedding and clustering medical diagnosis data,” in 2017 IEEE International Conference on Healthcare Informatics (ICHI), Aug 2017.

[2] Y. Choi, C. Y.-I. Chiu, and D. Sontag, “Learning Low-Dimensional Representations of Medical Concepts,” AMIA Summits on Translational Science Proceedings, vol. 2016.

[3] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in Proceedings of the 22Nd ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: ACM, 2016.

[4] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors," Big Data, vol. 3, no. 4, pp. 277–287, Dec. 2015.

[5] A. V. Lebedev, E. Westman, G. J. P. Van Westen, M. G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, and A. Simmons, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," NeuroImage: Clinical, vol. 6, pp. 115–125, 2014.