International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

A Framework for Disease Risk Prediction

Aswathy T[1], Anaswara R[2]

[1]*M. Tech. Student, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.*
[2]*M. Tech. Student, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.*

Abstract - *The big amount of information development in biomedical and healthcare companies, precise medical information analysis benefits from early identification, patient care, and community facilities. If the quality of medical information is incomplete, the precision of the assessment will be decreased. Different areas have distinctive features of certain regional illnesses, which may weaken the prediction of disease outbreaks. A neural network based disease risk prediction system proposed. The classification is done by using convolutional neural network. In the disease risk prediction model the class labels are identified based on the training set. Using convolutional neural network, disease prediction system to effectively predict the type of disease. The disease prediction system gives improved accuracy.*

Key Words: **Disease Prediction, Neural Networks, Machine Learning Algorithms, Convolutional Neural Network, Medical Data.**

1. INTRODUCTION

In healthcare industry, machine plays an important role for predicting diseases [1]. Data mining in health care has become increasingly popular because improved patient care by early detecting of disease. In machine learning, to make predictions on data is a common task. Such type of task is performed in data-driven predictions or decisions through building a mathematical model from input data. Usually the information used to construct the final model comes from various datasets. In particular, three types of data sets are commonly used for creation of the model. The model is initially fit on a training dataset that is a set of examples used to fit the parameters of the mode. A test dataset is a dataset that is autonomous of the training dataset but follows the same distribution of probability as the training dataset. Using training dataset, if a model fit that will also fit the test dataset. Different data mining techniques like Naive Bayes, Decision Tree and K-nearest neighbor's algorithm are used for disease prediction.

In data mining techniques classification is used for prediction. Classification is a form of data analysis that extracts models describing important data classes and Such Classifiers predict categorical class labels. Data classification consists of two step processes that are learning step and a training step. To predict and stop disease, it is essential to first define risk factors integrated in unstructured clinical documents. Over the past century, many studies have been conducted to define these risk factors, leading in the development of publicly accessible devices. Effective and efficient automated disease prediction systems can be helpful for predicting disease in the healthcare industry.

1.1 Objective

The main objective of "A Framework for Disease Risk Prediction" project is to develop a Disease Prediction System using machine learning techniques. The machine learning techniques are convolutional neural network based learning. The data mining tool Weka 3.8 is used for classification. Weka is a set of algorithms for knowledge mining functions in machine learning. The algorithms can either be applied directly to a dataset or called from the code. The objective is to predict diseases using data mining applications is challenging task but it will reduce the human efforts and increase the diagnostic accuracy.

2. RELATED WORK

Health monitoring through traditional wearable devices is difficult to sustainable. To get big data from healthcare through the author's sustainable health monitoring [2] design smart clothing, facilitating unobtrusive collection of multiple human body physiological signals. Mobile healthcare cloud platform is built using mobile web, cloud computing and large information analytics to provide pervasive intelligence for smart clothing design. The smart clothing system is constructed by a variety of biosensors into flexible textile clothing and to collect many important physiological indexes of human body. Smart clothing also provides comfortable wearing experiences and useful applications for user groups such as the elderly, children, suffered from chronic disease and mental illness.

Models based on individual patient features to predict the risk of cardiovascular events are significant tools for managing patient care. Using thoroughly chosen epidemiological cohorts, most present and frequently used risk prediction models were constructed. The homogeneity and restricted size of such cohorts, however, prevents these risk models predictive authority and generalizability to other populations. Electronic health data (EHD) from large health care systems provide access to data on large, heterogeneous, and contemporaneous patient populations. EHD's distinctive characteristics and difficulties, including missing data about risk factors, non-linear relationships between risk factors and cardiovascular event results, and distinct impacts from distinct subgroups of patients, require new approaches to

machine learning to develop risk models. The author [3] proposed a machine learning approach based on Bayesian networks trained on EHD to predict the cardiovascular risk from electronic health record data.

The survey [4] describes a data coherence protocol for the PHR-based distributed system. A flow estimating algorithm for the telehealth cloud system is proposed. Use several techniques of predicting future bandwidth consumption. A telehealth framework for bandwidth balance on emergency is presented. A telehealth scheme includes both clinical and non-clinical uses, not only providing data storage and forward services to be studied offline by appropriate experts, but also monitoring physiological information in real-time through omnipresent sensors to help remote telemedicine. However, the current telehealth systems do not consider the velocity and veracity of the big-data system in the medical context. Emergency incidents produce a big quantity of information in real time that should be stored in the data center and transmitted to distant hospitals. In addition, the information of patients is dispersed in the distributed data center, which cannot provide a highly efficient remote real-time service. A probability-based bandwidth model in a telehealth cloud system is proposed, which helps cloud broker to provide a high performance allocation of computing nodes and links. This brokering system considers Personal Health Record (PHR) location protocol in the cloud and schedules real-time signals with low transfer of data between various hosts. In a telehealth context, the broker utilizes several techniques of assessing bandwidth to predict the near future use of bandwidth. The findings of the simulation indicate that the model is efficient in determining the highest performing service and the service inserted validates the usefulness of the strategy.

3. SYSTEM ARCHITECTURE

System design is the method of defining a system's architecture, modules, interfaces, elements, and information to meet specific demands. It is seen as applying the theory of systems to the growth of products. There is some overlap with systems analysis, systems engineering. System design is high level strategy for solving a problem and building a solution. It includes the decision on system organization into subsystems, the distribution of subsystems to parts of hardware and software, and significant conceptual and policy choices that form the basis for comprehensive design. A system's general organization is called architecture of the system. The proposed system mainly consists of two modules:

- Admin Module
- DRP Module

In dataset management, it contains medical data. The data is uploaded and it is stored in a table. The Convolutional Neural Network classifier is applied to this dataset. It can be done by Disease Risk Prediction (DRP) module and the Convolutional Neural Network will produce the prediction result. Fig.1 shows the proposed system architecture.

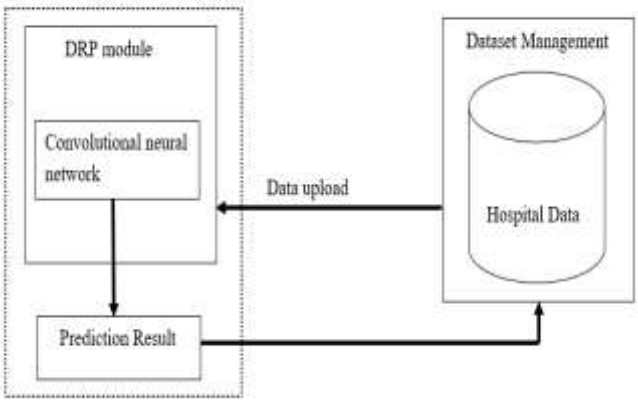

Fig -1: Proposed System of Disease Prediction

The main functionalities of the admin module are:

Table -1: Admin module functions and description.

Admin Home page	Functionalities added and performed in that will be displayed.
Dataset Management	Admin can view the entire document which is uploaded and stored in the table.
File Conversion	The dataset upload is comma separated value (CSV) files and this CSV file is converted into ARFF files.
Prediction	Using the ARFF file admin can predict the disease class labels.

The main functionalities of the DRP module are:

Table -2: DRP module functions and description.

Test set Management	User can upload details.
File Generation	Using test set CSV file is generated.
File Conversion	The CSV file is converted into ARFF files.
Prediction	Using ARFF file prediction is performed.

4. IMPLEMENTATION

Implementation is one of a project's most significant functions. It is the stage in which one must be careful, because all the attempts taken during this project will only be fruitful if the software is correctly implemented in accordance with the plans produced. Implementation is the phase in the project that transforms the theoretical design into a working scheme. The crucial stage is achieving successful new system and giving the users confidence in that the system will work effectively and efficiently.

Implementation is the carrying out, execution or exercise of a plan, process or design, concept, model, specification, standard or policy to do something.

It includes careful planning, investigating the present scheme and its limitations on implementing and designing techniques to bring about change. Apart from these, the major task of preparing for implementation is education and training of users.

The proposed system contains following steps:

- Dataset management
- File conversion
- Classification
- Prediction

4.1 Dataset Management

The data set used in this project is collected from UCI machine learning repository which is a repository of databases, domain theories and data generators. These are the names of the attributes that are the input provided for the record of the patient. In dataset management, the dataset is uploaded to the Disease Risk Prediction module. It is done by the admin. For disease risk prediction, the dataset used is electronic heath related data. The dataset used in this project is, breast cancer dataset and hepatitis data. These datasets are stored in the database and it is used for classification. There are two types of dataset is used for prediction. The dataset used are test set and training set. Identifying the correct and sufficient features to represent the data for the predictive models.

4.2 File Conversion

The dataset uploaded is Comma Separated Value (CSV) files. This CSV file is converted into Attribute Relation File Format (ARFF). The ARFF file is used for classification. The file conversion process is done by the admin.

4.3 Classification

The prediction of disease will be executed with the help of a tool known as Weka. Weka classifiers are models for nominal or numerical quantity prediction. Implemented learning schemes include decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptron's, logistic regression, naive bayes, neural network. Meta classifiers include bagging, boosting, stacking, output codes that correct errors, learning that is weighted locally. A Convolution Neural Network (CNN) consists of one or more convolution layers, followed as in a standard multilayer neural network by one or more fully connected layers. It is a kinds of artificial neural network specifically intended to process pixel information for image recognition and processing. Convolution Neural Network utilizes a scheme similar to a multilayer perceptron intended for decreased processing demands.
Neural network classifier is based upon back propagation algorithm to classify instances. The network is created by an MLP algorithm. During training time, the network can also be tracked and updated. The nodes in this network are all sigmoid. The neural network of back propagation is fundamentally a network of easy processing components that work together to generate a complicated output. The back propagation algorithm performs learning on a multilayer feed-forward neural network. It learns iteratively a set of weights to predict the class tuple label. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.

Each layer is made up of units. The network inputs match the measured characteristics for each training tuple. The inputs are fed into the units that make up the input layer concurrently. These inputs pass through the input layer and are then weighted and concurrently fed as units into a second layer of neuron, known as a hidden layer. The hidden layer unit's outputs can be input into another hidden layer, etc. The number of hidden layers is arbitrary, usually only one is used. At the core, back propagation is simply an efficient and exact method for calculating all the derivatives of a single target quantity with respect to a large set of input quantities. To improve the classification accuracy should reduce the training time of neural network and reduce the number of input units of the network. After classification, the class labels and predicted class are identified and this class labels are used for disease prediction.

4.4 Prediction

The convolutional Neural Network classifier is used for prediction. In the prediction step, test set training is performed. The users upload details of disease and using this details comma separated value file (CSV) is generated. The comma separated value file is converted into attribute relation file format (ARFF). The identified class labels and attribute relation file are used for prediction. Firstly, the training set is converted into training set by using convolutional neural network. Then this test set is used for prediction. After prediction the type of disease is understand.

To find predictive association rules in medical dataset the algorithm has three steps:

(i) In medical dataset both the categorical and numeric attribute are transformed into transaction dataset.
(ii) To find the predictive association rules with medically relevant attributes.
(iii) To validate the association rules the train and test approach should be used.

5. CONCLUSION

The patient should need a number of trials to detect a disease. But using data mining technique the number of test should be reduced Disease Risk Prediction System is developed using Convolutional Neural Network Classification technique. The system identifies categorical type of disease. This is the most effective model to predict class labels of disease. By using convolutional neural network classifier, the prediction accuracy and efficiency is improved. The prediction system normally uses a training set where all the objects are already associated with known class labels. The project determines the disease categorization using diverse machine learning algorithms by Weka tool. Dataset used from UCI Machine Learning Repository has been used as a source data. The disease prediction system gives improved accuracy.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities," in *IEEE Access*, vol. 5, pp. 8869-8879, 2017.

[2] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu," Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications, Vol. 21, No. 5, pp.825C845, 2016.

[3] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1033–1069, 2015.

[4] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," Journal of Systems Architecture, vol. 72, pp., 2017.

BIOGRAPHIES

Aswathy T, she is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Elavumthitta, Kerala, India. Her research area of interest includes the field of data mining, internet security and technologies.

Anaswara R, she is currently pursing Master's Degree in Computer Science and Engineering from Sree Buddha college of Engineering, Elavumthitta, India. Her research area of interest includes the field Data mining.