Cancer Disease Prediction Using Machine Learning Over Big Data

Ankita D. Bele[1], Vrushali K. Suryawanshi[2], Rajeeta A. Sharma[3], Manasi N. Deore[4]

[1]Ankita D. Bele, Dept. of IT Engineering, A. C. Patil College of Engineering, Maharashtra, India
[2]Vrushali K. Suryawanshi, Dept. of IT Engineering, A. C. Patil College of Engineering, Maharashtra, India
[3]Rajeeta A. Sharma, Dept. of IT Engineering, A. C. Patil College of Engineering, Maharashtra, India
[4]Manasi N. Deore, Dept. of IT Engineering, A. C. Patil College of Engineering, Maharashtra, India

---***---

Abstract - *Now one day big data is commonly used in each and every sector. With the help of big data, the medical and health care sectors achieve their growth. Using the big data advantage of accurate medical data analysis, prediction of early illness, accurate patient data can be processed and used in a secure way. Due to various reasons the precision of an experiment can be that. Incomplete medical data, the reasons for the reduction in accuracy are some regional disease characteristics which may be outbreaks of the forecast. Precise analyzes of medical data support early detection of illnesses, patient care and community services. If medical data quality is insufficient then the accuracy of the study decreased.*

We can use a machine learning algorithm in this paper to predict exactly the disease. We may collect the hospital data of a given region for that reason. Machine learning with maximization (support) of the separation margin (vector), called learning with support vector machine (SVM). It is a strong classification tool which has been used to identify or subtype cancer genomics. Today, with advances in high-throughput technologies leading to the development of large amounts of genomic and epigenetic data, the classification function of SVMs is extending its use in genomics for cancer, leading to the discovery of new biomarkers, new drug targets and a better understanding of cancer-driver genes.

***Key Words*: Machine learning (ML), supporting vector machine (SVM), classifier, analytics of big data, data on health care.**

1. INTRODUCTION

Big data definition isn't a new concept it is consta ntly changing. Big data are nothing but data gathering. The re are three significant vs data which is speed, volume and variety. Healthcare is one of three best examples of data vs . Business data are distributed through multiple medical systems, business markets, and government hospitals with the advantages of a big data paying more attention to the prediction of disease. Amount of work to pick the characteristics of a disease prediction from a wide volume of data has been performed.

To predict future results, machine learning (ML) "learns" a pattern from the past data. Learning which is one of the artificial intelligences is the main method. The learning methods such as logistic regression, artificial neural networks (ANN), K-nearest neighbor (KNN), decision trees (DT) and Naive Bayes can be applied in many different mathematical, probabilistic, and optimization techniques. There are two main types of ML learning-supervised learning and learning without supervision. The supervised learning builds a model by learning from known classes (labelled training data). By contrast, unsupervised learning methods learn from unknown class data (unlabelled training data) the common features.

SVM learning is one of many approaches used in ML. SVM is very effective in detecting subtle trends in complex datasets as compared with the other ML methods. SVM can be used for recognizing handwriting, recognizing fraudulent credit cards, identifying a speaker and detecting face. Cancer is a genetic disease where the patterns of genomic features or function patterns that reflect the subtypes of cancer, the outcome prognosis, the prediction of drug benefits, tumorigenesis drivers, or a biological process unique to tumours. Hence, SVM's Artificial Intelligence will help us identify these trends in a range of applications.

The paper's structure is as follows. This section offers the related work focused on the activities of cloud computing research recently carried out. The theoretical approach is outlined in Section 2. The proposed algorithm is explained in Section 3. Section 4 ends with future work on the paper.

1.1 DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA FROM HEALTHCARE COMMUNITIES

With large data development in biomedical and healthcare sectors, detailed analyzes of medical data support early detection of illness, patient care and community services. However, when the consistency of the medical data is insufficient, the precision of the study is the. In addition, different regions exhibit unique features of certain regional diseases that may weaken the prediction of disease outbreaks. In this paper, we streamline machine learning algorithms to effectively predict chronic disease outbreak in populations that are recurrent with disease. Deep analyzes of medical data help early detection of illness, patient care, and community services through large data creation in the biomedical and healthcare sectors. If the accuracy of the medical data is lacking, however, the study's precision is the. Additionally, different regions have unique features of certain regional diseases that may hinder the prediction of outbreaks of disease. In this paper, we streamline machine learning algorithms to effectively predict outbreaks of chronic disease in populations with recurrent illnesses.

1.2 DISEASE PREDICTION BY MACHINE LEARNING OVER BIG DATA

Because of big data advances in scientific and healthcare fields, detailed analysis of medical data benefits from early detection of illness, patient care, and community services. The accuracy of the analysis is diminished when the consistency of the medical data is insufficient. In addition, different regions show unusual manifestations of certain regional diseases which may contribute to a weakening of disease outbreak prediction. It provides machine learning algorithms in the proposed framework for successful prediction of various disease occurrences in societies that are recurrent with disease. It is playing with the altered estimation models over data collected from real-life hospitals. To address the complexity of incomplete data, the missing data are retrieved using a latent factor model. It studies on a national chronic cerebral infarction condition. It uses Machine Learning Decision Tree algorithm and Map Reduce algorithm using structured and unstructured data from the hospital. None of the recent research focused on all forms of data to the best of our knowledge in the field of medical big data analytics. Compared to other standard estimation algorithms, our proposed algorithm's measurement accuracy hits 94.8 per cent with a convergence speed that is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

1.3 APPLICATIONS OF SUPPORT VECTOR MACHINE (SVM) LEARNING IN CANCER GENOMICS

Machine learning with maximization (support) of separation margin (vector) learning called Support Vector Machine (SVM) Learning is a powerful classification method used for the classification or subtyping of genomic cancer. Today, with advances in high throughput technologies leading to the development of large amounts of genomics and epigenomic data, the classification function of SVMs is extending its use in genomics for cancer, leading to the discovery of new bio makers, new drug targets and a better understanding of cancer-driver genes. Here we looked at recent progress of SVMs in genomic cancer. We plan to understand the impact of SVM learning in the cancer genomics applications and its future perspective.

1.4 CLASSIFICATION OF CANCEROUS PROFILES USING MACHINE LEARNING ALGORITHMS

There are many existing strategies for detecting lung cancer. The type of treatment prescribed for a patient is affected by various factors such as cancer type, cancer severity (stage) and the genetic heterogeneity most important. In such a complex environment it is possible that the targeted drug treatments will be irresponsive or respond differently. They need to consider the cancer profiles in order to study anticancer drug response. These cancer profiles carry information that can investigate the underlying factors responsible for the growth of cancer. Thus cancer data need to be analyzed to determine optimal treatment options. Examination of such profiles can aid in predicting and identifying possible drug and drug targets. The main aim in this paper is to provide classification technique based on machine learning for cancer profiles. Keywords-Cancer, Machine Learning, SVM, Random Forest, KNN, Genes, Prediction of Drugs.

2. RESEARCH METHOD

A literature review is conducted to identify an effective evaluation algorithm, and represents a recent approach that improved algorithm efficiency and accuracy in the distributed world. So if the disease continues, early treatment can be given. Prompt diagnosis can be given to patients who can reduce the risk of recovery and save lives of patients and can reduce the cost of disease care to some degree.

We have conducted analysis for our framework using the Support Vector Machine (SVM) algorithm using validation steps. By harnessing the forces of machine learning methods, the paper sets itself apart. The project proposes a framework with a strong predictive algorithm that implements powerful classification measures with an extensive module for report generation. This project aims to implement a self-learning protocol such that the past inputs of the disease outcomes determine the future possibilities of the Cancer to a particular use. The proposed algorithm is divided into two sections one is Dataset Pre-processing and Classification using Support Vector Machine. By harnessing the forces of machine learning methods, the paper sets itself apart. The project proposes a framework with a strong predictive algorithm that implements powerful classification measures with an extensive module for report generation.

3. PROPOSED ALGORITHM AND IMPLEMENTATION

This section describes the implementation plan which performs analysis using the Support Vector Machine (SVM) algorithms to implement validation steps. The paper distinguishes itself by drawing on the power of machine learning techniques. The project proposes a system with a strong predictive algorithm, which implements powerful classification measures with a comprehensive report generation module. This project seeks to integrate a self-learning method such that past experiences from disease results assess the potential possibilities for a particular application of the Cancer.

The project proposes a system with a strong prediction algorithm, with a detailed report generation module implementing efficient classification measures. This project seeks to integrate a self-learning method such that past experiences from disease outcomes assess the potential possibilities for a particular application of the Cancer. In this method, we consider the following steps, such as Load Dataset, after loading the dataset, Classify Features (Attributes) will then estimate Candidate Support Value based on class labels, since the state is While (instances! =null), Do condition if Support Value=Similarity in the attribute to each instance and then find the total error value. Assume the approximate value of the decision= Help Value\Total Error, repeated for all points until it is zero, if any, then < 0. Therefore, we primarily calculated the entropy and Gini index.

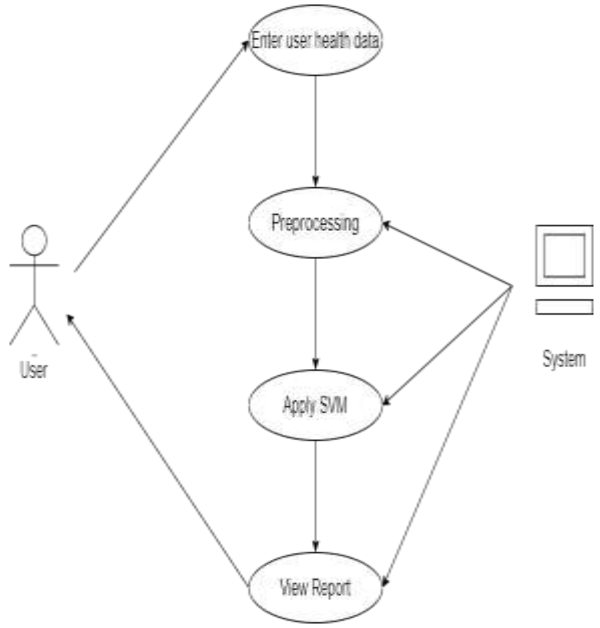

Figure 1: Block Diagram of Proposed Algorithm

EVALUATION METHODS

For the experiment success assessment. Next, we denote TP, FP, TN and FN as true positive (the number of instances correctly predicted as required), false positive (the number of instances incorrectly predicted as required), true negative (the number of instances correctly predicted as unnecessary), and false negative (the number of instances incorrectly predicted as unnecessary), respectively. We can then obtain four measurements as follows: precision, accuracy, recall and F1 calculation:

Accuracy = (TP + TN)/ (TP + FP + TN + FN)

Precision = TP/ TP + FP,

Recall = TP /TP + FN

F1-Measure = (2 Precision Recall) / (Precision + Recall)

4. CONCLUSIONS

In the report, we achieved a 73 percent accuracy rate. We propose a Support Vector (SVM) computer, which is a machine learning algorithm for classifying hospital data. None of the existing work centered, to the best of our knowledge, on both types of data in the area of medical big data analytics.

We have seen that clinical data is a recent area of research aimed at using data and machine learning capabilities to uncover the biological patterns. In addition, the oncogenomics research area seeks to classify and interpret genes related to cancer and thus assists in genotype-level diagnostics. Although different approaches have been suggested in the classification literature, selection of genes is still a big curse. Cancer is a heterogeneous disorder consisting of diverse subtypes. Therefore, there is an immediate need to build programs or approaches that can assist in early diagnosis and traditional cancer prognosis. Various new methods relating to cancer research have developed over the past decade. Scientists have used numerous biological and computational methods for early detection of types of cancer. The collection of vast collections of cancer data has hiked work in this area. Various approaches to machine-learning were used to predict cancer. The proposed strategy is to solve the problem of classifying cancer genomic profiles. Our technique is based on principle of using SVM, the algorithm for machine learning. Result offers comparative output analysis of the model when the sample size is varied. As the sample size increases, model output also improves, showing a positive aspect towards the model's robustness and adaptability.

REFERENCES

[1] in Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, Disease Prediction by Machine Learning over Big Data from Healthcare Communities, 2169-3536 (c) 2016 IEEE

[2] rof. Dhomse Kanchan B. Assistant Professor of IT department METS BKC IOE, Nasik Nasik, India kdhomse@gmail.com , Mr. Mahale Kishor M. Technical Assistant of IT department METS BKC IOE, Nasik, India kishu2006.kishor@gmail.com, Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analy-sis,2016 IEEE.

[3] Vinitha S, Sweetlin S, Vinusha H and Sajini S, “Disease prediction using machine learning over big data”, February 2018 CSEIJ

[4] hahab Tayeb*, Matin Pirouz*, Johann Sun1, Kaylee Hall1, Andrew Chang1, Jessica Li1, Connor Song1, Apoorva Chauhan2, Michael Ferra3, Theresa Sager3, Justin Zhan*, Shahram Latifi, Toward Predicting Med-ical Conditions Using k-Nearest Neighbours, 2017 IEEE International Conference on Big Data.

[5]. J. Senthil Kumar, S. Appavu. “The Personalized Disease Prediction Care from Harm using Big Data Analytics in Healthcare”. Indian Journal of Science and Technology, vol 9(8), DOI: 10.17485/ijst/2016/v9i8/87846, [2016]. ISSN (Print): 0974-6846, ISSN (Online): 0974-5645.

[6]Gakwaya Nkundimana Joel, S. Manju Priya. “Improved Ant Colony on Feature Selection and Weighted Ensemble to Neural Network Based Multimodal Disease Risk Prediction (WENN-MDRP) Classifier for Disease Prediction Over Big Data”. International Journal of Engineering & Technology, 7(3.27) (2018) 56-61.