ROLE OF MACHINE LEARNING ALGORITHM FOR BIG DATA PROCESSING IN HEALTHCARE SECTOR - A REVIEW

V DURGA DEVI
Research Scholar ,
Department of Computer Applications,
VISTAS, Pallavaram.

Dr R PRIYA
Professor,
Department of Computer Applications,
VISTAS, Pallavaram.

Abstract

Artificial intelligence and expert systems plays a key role in modern medicine sciences for disease prediction, surveillance interventions, cost efficiency and better quality of life etc. With the arrival of new web-based data sources and systematic data collection through surveys and medical reporting, there is a need of the hour to develop effective recommendation systems which can support practitioners for better decision-making process. Machine Learning Algorithms (MLA) is a powerful tool which enables computers to learn from data. While many novel developed MLA constantly evolves, there is need to develop more systematic, robust algorithm which can interpret with highest possible accuracy, sensitivity and specificity. This study reviews previously published series on different algorithms their advantages and limitations which shall help make future recommendations for researchers and experts seeking to develop an effective algorithm for predicting the likelihood of various diseases.

Keywords: Artificial Intelligence, Expert Systems, Machine Learning Algorithms, Disease prediction, Future Recommendations.

Introduction

The advent of Healthcare Information Systems (HIS) plays an imperative role in the field of medical sciences and technology as it assists medical practitioners in development of accurate methods of disease prediction, high-risk assessment and sustainable health monitoring [1, 2]. Healthcare information systems integrate IT with healthcare to meet the growing demands of quality and efficacy in healthcare systems across the globe[3]. Artificial Intelligence and deep learning techniques are currently in trend and several studies exclusively focus on analyzing its support towards modern medical decision models [4]. With the arrival of new web-based data sources and systematic data collection through surveys and medical reporting, there is a need of the hour to develop effective recommendation systems which can support practitioners in the better decision-making process [5].

Machine Learning Algorithms (MLA) is a powerful tool which enables computers to learn from data [6]. Over past decade, a number of machine learning classifiers have been developed which is broadly classified into the white and black box. While, white box MLA is simple and transparent which includes simple decision tree, black box models which are also known as deep learning models are often difficult to interpret their inner working [7]. Neural Networks are paradigmatic examples of deep learning algorithms [8] which also includes Random Forest model, the Support Vector Machine models etc. Along with machine learning algorithms, data mining techniques and statistical analysis provide major support to experts in prediction of disease [9]. Medical data in the form of electronic health records, sensors and monitors analyzed using traditional machine learning algorithms like logistic regression and regression analysis proves to be effective in disease prediction using structured clinical or hospital data. These algorithms were supervised and trained to classify characteristics based on past experiences [10]. With the development of new computational tools such as big data analytics technology, modern machine learning algorithms use unsupervised machine learning approach to select features or attributes automatically from larger datasets to improve the accuracy [11].

The objectives are To understand application of different machine learning algorithm in prediction of particular disease and to identify the advantages and limitations of the MLA for healthcare sector data. The methodology used for this study is a review methodology where several journal and peer-reviewed articles were reviewed with respect to machine learning algorithm for medical applications. In this review, machine learning methods used for the medical application is described. Also, a comparative review of techniques adopted in machine learning technique is adopted and analyzed. Potential uses of machine learning methods such as support vector machine, artificial neural networks and deep learning are also discussed for application in the field of medical science. The content of this paper flows as follows: Review of previous studies conducted in terms of disease management using various machine learning algorithms and conclusion.

Related Works

Mental Health related Diseases

Study by Schnack[12] used CNN to analyze sMRI image in prediction of Schizophrenia, Alzheimer's disease. The study tested the following classifiers, multiple (k) SVM classifiers; Decision tree such as Random Forest; fuzzy c means clustering; Gaussian distribution; multi-kernel learning. The study used Clustering technique based on data partitioning and concluded that fuzzy c means clustering helps to increase subgroup hence achieve higher accuracy in combination with SVM when compared to other models.

Study by Chen et al. [13] tested performance of different classifiers in prediction of cerebral infarction in regions of China. Data in the form of EHR, medical image and gene data both structured and unstructured was used. 31,919 hospitalized patient’s records were recovered. For text data CNN-UDRP classifier was used; For structure and text data, Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP) was used and for structured only data NB+ gaussian distribution, KNN and DT (CART) algorithm was used. CNN-based Unimodal Disease Risk Prediction (CNN-UDRP) algorithm showed accuracy rate of 94.8%. CNN-MDRP algorithm showed accuracy rate of 94.80%. For Structured data, although DT showed highest accuracy rate of 63%, overall NB classifier showed better performance in disease prediction.

Study by Chen et al. [14] predicted multiple sclerosis from clinical and MRI data. A total of 1600 subjects' record was used from CLIMB study. Support Vector Machine (SVM) classifier was tested for both clinical and MRI data which showed the following performance: Accuracy rate of 70%, sensitivity of 62% and 71%, specificity of 65% and 68% respectively in predicting multiple sclerosis. Study by Abos et al. [15] attempted to predict Parkinson's disease using fMRI images. 133 patients dataset was collected. Supervised machine learning technique was applied and the study tested performance of SVM. The study found that SVM had an accuracy rate of 80%. From the above studies we can conclude that for prediction of mental health-related diseases using MRI image, SVM classifier has outperformed other types of machine learning classifiers in terms of accuracy, sensitivity and specificity. The Below mentioned table1 provides insight into the management of mental health related diseases using different classifiers for different disease prediction.

Table 1. Mental Health-related Disease Management

Data input format	Classifiers used	Disease Prediction
sMRI, fMRI Clinical/ Hospital data HER, Demographical data	ridge, LASSO, elastic net, L0 norm regularized logistic regressions, a support vector classifier, regularized discriminant analysis, random forests, Gaussian process classifier and Support Vector Regression, CNN, NB+ gaussian distribution, KNN and DT (CART) algorithm	Psychosis, Alzeihmer's disease, Schizophrenia, cerebral infaraction, multiple sclerosis, Parkinson's disease

Lung Disease

Study by Le-Dong et al. [16] predicted interstitial lung disease based on combined data obtained from Pulmonary Function Test (PFT) and Hi-Res CT. The dataset consisted of 323 subjects of which 244 patients were identified with systemic sclerosis. The study reported that SVM with z score showed accuracy 84%, sensitivity 60% and specificity of 96% respectively.

Study by Le-Dong et al. [17] attempted to predict asthma based on patient's telemonitoring data. The dataset consisted of 7001 daily telemonitoring records of adult asthma patients for 7 days. Classifiers such a naive Bayesian classifier, adaptive Bayesian network, and support vector machines was tested for performance and found that adaptive Bayesian network showed accuracy, sensitivity and specificity of 100% each respectively. Study by Wiemken et al. [18]analysed statistical and machine learning algorithm in prediction re-hospitalisation within 30 days among pneumonia patients. Datasets was obtained from hospital which included 3249 patients suffering from pneumonia. The study tested the following classifiers: LR, LASSO regression, RF, RPT, CIT and NB. The study reported that it is a challenge to predict re-hospitalization among pneumonia patients using statistical and MLA. From the above-quoted studies, it is understood that neural network and SVM outperformed other prediction models. The below table 2 summarizes the machine learning classifiers used for prediction of various lung diseases.

Table 2. Lung Disease

Data Type Format	Classifiers used	Prediction
EHR Hospital data Tele-monitoring	DT, LR, SVM, NN, RF, kNN SVM with z score, NBC, ABN, LASSO, RPT, CIT, ID3 C4.5, NB and BN (K2).	Asthma COPD

Cancer

Study by Guadagni et al. [19] predicted breast cancer from web based source. Routinely collected clinical data was used for analysis. The study evaluated the performance of multiple kernel learning approach combining SVM and Random optimization but did not externally validate the result due to limitation of data.

Study by Murty and Babu[20] predicted lung cancer using NB. Dataset from UCI Machine Learning Repository of Lung Cancer Patients and Michigan Lung Cancer included 32 subjects of which 16 were identified as lung cancer patients. The initial 57 attributes were later developed to 7130 attributes as the number of subjects and cancer patients were increased to 96 and 86 respectively. The study reported that NB outperforms other prediction models. The below table 3 summarizes different classifiers used for cancer disease management.

Table 3. Cancer disease management

Data type format	Classifiers used	Prediction
Clinical, demographical data Ultrasound Web based source Open source data repository	multiple kernel learning approach combining SVM and Random optimization, NB, RBF Neural Network, MLP, C4.5 (J48) algorithm, trained NN, SVM, RF	Lung, breast, prostate, colorectal, reoccurrence, survivability.

Conclusion

In this review, the overview of research in big data analytics as specific towards machine learning approach is provided. Big data creates numerous challenges for traditional machine learning in terms of their scalability, adaptability and usability. The present study began to explain with a description of machine learning algorithm followed by the issues faced in the machine learning algorithm and its possible remedies. In future, we have planned to implement a solution for particular issue including uncertain and incomplete dataset using the solution learning with the use of Parkinson telecommunication dataset. It would be more interesting in concentrating towards the trend one or a greater amount of these issues frequently seen in big data, hence accumulation the machine learning and big data analytics research corpus.

Reference

1.Jamison D, Breman J, Measham A (2006) Disease Control Priorities in Developing Countries. Oxford University Press, New York

2.Winters-Miner LA (2014) "Seven ways predictive analytics can improve healthcare: Medical predictive analytics have the potential to revolutionize healthcare around the world". In: Elsevier. 22 Nov 2017

3.Omachonu VK, Einspruch NG (2010) Innovation in Healthcare Delivery Systems: A Conceptual Framework. Innov J Public Sect Innov J 15:1–20

4.Fatima M, Pasha M (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. J Intell Learn Syst Appl 09:1–16 .

5.Assistant Secretary for Planning and Evaluation (2011) Improving Data for Decision Making: HHS Data Collection Strategies for a Transformed Health System. 22 Nov 2017

6.Bhatt C, Dey N, Ashour AS (2017) Internet of Things and Big Data Technologies for Next Generation Healthcare. Springer, London

7.IBM (2012) Decision Tree Models. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm. Accessed 22 Nov 2017

8.Srinivas S, Sarvadevabhatla RK, Mopuri KR, et al (2016) A Taxonomy of Deep Convolutional Neural Nets for Computer Vision.

9. Danjuma K, Osofisan AO (2015) Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis

10. Dodek PM, Wiggs BR (1998) Logistic regression model to predict outcome after in-hospital cardiac arrest: validation, accuracy, sensitivity and specificity. Resuscitation 36:201–8.

11. L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM (2017) Machine Learning With Big Data: Challenges and Approaches. IEEE Access 5:7776–7797 .

12. Schnack HG (2017) Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases).

13. Chen M, Hao Y, Hwang K, et al (2017) Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE Access 5:8869–8879.

14. Zhao Y, Healy BC, Rotstein D, et al (2017) Exploration of machine learning techniques in predicting multiple sclerosis disease course. PLoS One 12:e0174866

15. Abos A, Baggio HC, Segura B, et al (2017) Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning. Sci Rep 7:45347

16. Le-Dong N-N, Hua-Huy T, Ngoc HMN, et al (2017) Detection of Interstitial Lung Disease in Systemic Sclerosis Using a Machine Learning Approach Based on Pulmonary Function Tests. Am J Respir Crit Care Med 195:A2531.

17. Finkelstein J, Jeong I cheol (2017) Machine learning approaches to personalize early prediction of asthma exacerbations. Ann N Y Acad Sci 1387:153–165 .

18. Wiemken TL, Furmanek SP, Mattingly WA, et al (2017) Predicting 30-day mortality in hospitalized patients with community-acquired pneumonia using statistical and machine learning approaches. J Respir Infect 1:50–56 .

19. Guadagni F, Zanzotto FM, Scarpato N, et al (2017) RISK: A Random Optimization Interactive System Based on Kernel Learning for Predicting Breast Cancer Disease Progression. In: Rojas I, Ortuño F (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2017. Lecture Notes in Computer Science. Springer, Cham, pp 189–196

20. Murty NVR, Babu PMSP (2017) A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis. nternational J Comput Intell Res 13:1041–1048.