DESIGN OF MACHINE LEARNING TOOLS FOR BIG DATA ANALYTICS

Abstract

Big data has increased in use in healthcare over the past few years as a result of three main factors: the large amount of data that is already available, rising healthcare costs, and an emphasis on individualized care. Big data processing in healthcare relates to the development, gathering, analysis, and storage of too huge or complex clinical data inferred by conventional data processing techniques. Examples of big data sources for the healthcare industry include the Internet of Things (IoT), Electronic Medical Record/Electronic Health Record (EMR/EHR), which holds patient's medical history, diagnoses, medications, treatment schedules, sensitivities, laboratory test results, genomic sequencing, medical images, insurances, etc. Many techniques related to machine learning approaches relevant to various healthcare data sets are discussed in this paper. Moreover, the challenges associated with handling massive data, their applications, and their processing are dealt in this study.

Keywords: Big Data, Machine Learning, Healthcare, Genomic Sequencing, Data Processing.

Authors

Mrs. A Vaideghy
Assistant Professor
Department of Computer Science
PSG College of Arts & Science
Coimbatore, Tamilnadu, India.

Dr. C Thiyagarajan
Associate Professor
Department of Computer Science
PSG College of Arts & Science
Coimbatore, Tamilnadu, India.

I. INTRODUCTION

Machine Learning (ML) is a subfield of artificial intelligence that bring up the skill of IT systems to discover resolutions to problems on their own by recognising patterns in databases. Machine Learning enables intelligence systems to recognise patterns and build appropriate solution concepts based on current algorithms and data sets. As a result, artificial knowledge is formed in machine learning experience. To learn from data sets, statistical and mathematical methods are utilised in machine learning. There are two types of approaches: symbolic approaches and sub-symbolic approaches. Sub-symbolic systems are artificial neural networks, whereas symbolic systems are propositional systems in which the knowledge content, is openly recorded. These operate on the basis of the human brain, in which the knowledge contents are implicitly represented. The critical issues of machine learning for big data are large scale data, different types of data, high speed of streaming data, uncertain and incomplete data [1]. Machine learning classifications are supervised, unsupervised, and reinforcement learning.

II. LITERATURE SURVEY

Article compares the assessments made by many writers in tackling a clinical challenge in the domain of machine learning.

Machine Learning, as a part of AI, teaches the system from previously collected data to recognise patterns and make judgements with minimal human intervention. Support Vector Machine (SVM), logistic regression, clustering, and other such techniques are examples.

Kai Hwang et al. [2] suggested a big data-applicable Convolution Neural Network-based multimodal disease risk prediction (CNN-MDRP) algorithm. The accuracy of the disease risk model is determined by combining structured and unstructured variables.

Ya Zhang and Tao Zheng[3] suggested a method based on machine learning that makes use of a large data EHR database. A supervised learning algorithm served as the foundation for the framework. Because raw EHR is frequently unstructured and sparse, feature engineering was required to correctly structure it. A total of 16 features were mined and built for machine learning structure. ML methods such as Random Forest, Logistic Regression, and AdaBoost are applied, for improved results. Algorithms additionally optimise the filtering criteria to boost recall while minimising false-positives.
Jyotishman Pathak et al.,[4] shared the enactment of SHFM (Seattle Heart Failure Model) using EHR, for risk prediction using ML techniques.

Andy Schuetz et al.,[5] suggested recurrent neural network (RNN) models based on gated recurrent units (GRUs) to find relationships between time-stamped events during a 12 to 18-month observation window of patients and controls.

Shulong Zhang et al. [6] proposed an LSTM (long short term memory) prediction model framework for HF diagnosis.

In order to predict sepsis using the retrospective Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III dataset, Jana Hoffman et al. [7] proposed a

machine-learning classification system that uses multivariable combinations of readily available patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age). This dataset was restricted to intensive care unit (ICU) patients aged 15 years or more.

Big data-driven machine learning methodology was proposed by Arjun K. Venkatesh et al. [8] to predict in-hospital mortality of ED (emergency department) patients with sepsis.

With the help of EHR, Susan E. Clare et al. [9] suggested a unique concept-based filter and prediction model to help breast cancer patients find local recurrences.

The many machine learning techniques for identifying diabetes levels have been detailed by Thiyagarajan C., Anandha Kumar K., and Bharathi K. [10].

With the help of CT scan pictures, Barstugan M., Ozkaya U., and Ozturk S.[11] have implemented machine learning techniques for early-stage coronavirus identification. 53 cases from 150 numbered abdomen CT scan images and 150 numbered chest CT scan images have been treated using the suggested procedure. In this study, images were cropped, and several feature extraction techniques were used to obtain the necessary information. Support Vector Machine, a classifier, categorizes the extracted characteristics. The GLSZM feature extraction techniques produced the best results, with an accuracy rate of 99.68%.

A precise, lively, and intelligent M-Health system has been proposed by Naseer Qreshi K, Din S., and Jeon G [12]. Their work uses a machine learning-based prediction model that supports data collecting, pre-processing, data segmentation, algorithm learning, and decision-making using trained datasets.

The use of several machine learning techniques for the situation of a coronavirus outbreak is highlighted by Lalmuanawma, Hussain, and Chhakchhuak [13]. Their research suggests several ways to combat the pandemic.

The application of machine learning techniques in emergency care is covered by Shafaf N., Malek H., and others [14]. Their research gathers and assesses the methods used in earlier studies.

ML has numerous critical uses in musculoskeletal medicine, according to Christopher Tack[15].

According to Schwartz, J. M., Moy, A. J., Rossetti, S. C., Elhadad, & Cato, K. D.[16], machine learning systems assist physicians in interpreting data from electronic medical records and carrying out their diagnosis and treatment plans.

A. Garg, V. Mago, and others [17] present different machine learning techniques helpful to the medical industry.

In a comparison study of algorithms like GBM and Logistic Regression models, authors Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao [18] have found that they outperform techniques like random forest and decision tree models.

The papers mentioned above by multiple authors offer numerous suggestions for diagnosing the illness using various machine learning techniques. Despite the volume of study, there are currently no reliable ways for diagnosing the disease, which motivates us to develop new ones.

III. MACHINE LEARNING APPROACHES

1. **Support Vector Machine (SVM):** SVM is a ML algorithm that comes under the group of supervised learning and is useful for solving classification and regression-related issues. In order to categorize fresh data points and divide n-dimensional space into classes, the method constructs the optimum decision boundary, or hyper plane. By selecting the extreme vector known as the support vector, a hyperplane is produced.

2. **Random Forest Classifier:** A machine learning technique called Random Forest Classifier supports supervised learning while preferring classification and regression-related problems. This approach assumes the property of ensemble learning to integrate many classifiers for handling complicated tasks. Additionally, this strategy enhances the system's functionality.

3. **K-Nearest Neighbour Classifier:** This technique is best suited for classification issues in supervised learning. The categorization process of the algorithm is based on the notion that new and existing cases are comparable, and it finds the new example that most closely resembles the existing categories. This approach classifies the new data point from the existing data based on similarity. Due to the fact that the algorithms do not instantly learn from the training set, it is correspondingly known as a lazy learner algorithm. As an alternative, it implements action during classification using the dataset that was previously stored. The technique just saves the dataset while it is in the training phase. Data will be categorized and classed when exposed to fresh data depending on how well it matches the new data.

4. **Gradient Boost Algorithm (GBM):** Final predictions are produced using the Gradient Boosting Machine algorithm by merging predictions from various decision trees. The weak learners are used to build decision trees. In order to choose the appropriate split, each node in the decision tree considers the characteristics of several subsets. This function enables extracting various signals from the data. Additionally, each new tree creates a succeeding decision tree by counting the errors made by the preceding trees. The trees are constructed in this way in order.

5. **Logistic Regression Model:** This approach supports supervised learning-related machine learning principles. It predicts the output of a categorical dependent variable from the independent factors.

6. **Grey Wolf Optimization (GWO):** This population-based, meta-heuristic algorithm mimics the natural leadership structure and hunting strategy of grey wolves. The top of the food chain is snatched up by grey wolves, which are classed as apex predators. Grey wolves prefer to live in packs, which typically have 5 to 12 members apiece. Each member of the group adheres to a strict hierarchy of social authority.

7. **Bat Algorithm (BA):** This is a modern global optimization meta-heuristic algorithm based on the echolocation ability of microbats. This approach can be used to solve estimation issues with lacking data.

8. **Firefly Algorithm (FA):** A novel meta-heuristic algorithm called FA was developed based on the characteristic of fireflies and their flashing patterns. The algorithm can be used to solve estimation issues with lacking data.

IV. PERFORMACE MEASURES IN MACHINE LEARNING

1. **Accuracy:** The ratio of the overall number of true positive predictions to the overall number of true negative predictions that a model correctly classified.

2. **Calibration:** A measurement, such as the Brier score, of how well anticipated probabilities for a result contest the observed result in test data.

3. **Discrimination:** Area under the receiver operator curve is a common way to assess a model's ability to distinguish between true positive and true negative cases that were chosen at random.

4. **Negative predictive value:** Sum of genuine negatives divided by sum of false negatives

5. **Precision:** Sum of all correctly classified positive events divided by the sum of all positively classified events.

6. **Recall:** Total correctly classified positive data divided by the total number of positive class members in the data.

7. **Specificity:** Sum of all correctly classified negative data to the total negative class members

V. CONCLUSION AND FUTURE SCOPE

Big data supports better understanding of each patient's health with more accurate predictive models to yield disease analysis and treatment. Big data combines increased amount of information on many scales of what makes up an illness, including DNA, proteins, and metabolites as well as cells, tissues, organs, organisms, and ecosystems. We covered large data applications, processing, and handling utilizing a variety of machine learning approaches in this study. Additionally, big data is employed to support the metrics used to assess the machine learning models' success. By using various methodologies to forecast diseases and make prompt diagnoses, which can improve a patient's health, machine learning also aids in effective decision-making. Information can be predicted in advance, and early disease prevention is possible. The quick adoption of EHR has produced a wealth of fresh patient data that is a goldmine for deepening our understanding of human health. In the near future, the healthcare sector and the healthcare organization will quickly and widely incorporate and exploit big data and machine learning. Concerns about preserving security, setting standards, ensuring privacy, ensuring governance, and continually advancing the tools and technology will gain focus as big data analytics becomes more widespread.

REFERENCE

[1] Xu Y, Qiu J, Wu Q, et al. (2016) "A survey of machine learning for big data processing", EURASIP J Adv Signal Proc 2016: 67
[2] Hao Y, Hwang MCK, Wang L, et al. (2017) Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5: 8869-8879.
[3] Zhang Y and Zheng T, (2017) A big data application of machine learning-based framework to identify type 2 diabetes through electronic health records, In: International Conference on Knowledge Management in Organizations, Springer, 451-458.
[4] Pereira N, Taslimitehrani V, Pathak J, et al. (2015) Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inf 216: 40.
[5] Stewart WF, Sun J, Choi E, et al. (2017) Using recurrent neural network models for early detection of heart failure onset. J Am Med Inf Assoc 24: 361-370.
[6] Liu Z, Zhang S, Jin B, et al. (2018) Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access 6: 9256-9261.
[7] Calvert J, Hoffman J, Jay M, et al. (2016) Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. JMIR Med Inf 4: e28.
[8] Hall MK, Pare JR, Venkatesh AK, et al. (2015) Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 23: 269-278.
[9] Neapolitan R, Zexian SE, Roy A, et al. (2018) Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinfor 19: 65-74.
[10] C. Thiyagarajan, K. Anandha Kumar, A. Bharathi "A Survey on Diabetes Mellitus Prediction using Machine Learning Techniques," International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 3 (2016) pp 1810-1814
[11] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," 5, pp. 1–10, 2020.
[12] Naseer Qureshi K, Din S, Jeon G, Piccialli F. An accurate and dynamic predictive model for a smart M-Health system using machine learning. Inf. Sci. 2020;538:486–502. doi: 10.1016/j.ins.2020.06.025.
[13] Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals. 2020;139:110059. doi: 10.1016/j.chaos.2020.110059.
[14] Shafaf, N., Malek, H.: Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch. Acad. Emerg. Med. 7, (2019).
[15] Tack C. Artificial intelligence and machine learning | applications in musculoskeletal physiotherapy. Musculoskelet. Sci. Pract. 2019;39:164–169. doi: 10.1016/j.msksp.2018.11.012.
[16] Schwartz JM, Moy AJ, Rossetti SC, Elhadad N, Cato KD. Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review. J. Am. Med. Inform. Assoc. 2021;28:653–663. doi: 10.1093/jamia/ocaa296.
[17] Garg A, Mago V. Role of machine learning in medical research: A survey. Comput. Sci. Rev. 2021;40:100370. doi: 10.1016/j.cosrev.2021.100370.
[18] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi & Xin Gao, "Predictive models for diabetes mellitus using machine learning techniques", BMC Endocrine Disorders, Article Number. 101, 2019.

