A COMPREHENSIVE REVIEW ON MACHINE LEARNING AND GENETIC ALGORITHMS IN NETWORK BEHAVIOR OPTIMIZATION

NALLA AKHILA
Research scholar, Department of Computer Science & Engineering, University of Techonology, Vatika, Jaipur, E-mail: akhila.uot@gmail.com

Prof. SUNEEL PAPPALA,
Department of Computer Science & Engineering, University of Technology, Vatika, Jaipur, E-mail: suneelpappala@gmail.com

ABSTRACT

In machine learning the process of selecting features is an important task. This is a method of selecting a subset of relevant or significant variables and features. Neural networks and genetic algorithms are the two sophisticated machine learning techniques presently attracting attention from scientists, engineers, and statisticians, among others. They have gained popularity in recent years. Optimization is aimed toward deviating from the limitations attributed to machine learning in order to solve complex and challenging problems. Feature selection is applicable in multiple areas such as Diabetes Prediction, anomaly detection, Bioinformatics, image processing, etc. where high dimensional data is generated. This study gives a literature review on different feature selection methods. Emphasizing the need for standardized data, low-cost ML applications for smallholder farmers and interdisciplinary collaboration. Future research also needs to work on data integration features, real-time data monitoring, and eliminating socio-economic barriers to ML usage for that sustainable and responsible solution for the agricultural sector. The genomes often contain repetitive sequences and duplicated regions, which can lead to ambiguities during assembly. Thus, the process of reconstructing a complete genome from a set of reads necessitates the use of efficient assembly programs.

Keywords: *machine learning, genetic algorithms, Diabetes Prediction, data integration features, nature-inspired algorithms, Bioinformatics.*

INTRODUCTION

Combined with simultaneous increases in computing power, the explosion in Big Health Data has driven a resurgence in Machine Learning (ML) research including data-hungry Deep Learning algorithms. More computationally efficient algorithms with improved performance now offer huge potential for improved diagnosis, risk prediction and more personalised approaches to clinical management. This is particularly relevant for the management of the older patient population which is typically characterised by complex multimorbidity phenotypes and significant interindividual variability in homeostatic capacity, organ function, and response to treatment. Clinical tools that utilise ML algorithms to support clinicians in determining the optimal choice of treatment are slowly gaining the necessary approval from governing bodies and being implemented into healthcare, particularly in the fields of radiology, pathology and imaging, and it is expected that almost all medical disciplines will likely be affected during the next phase of digital medicine. Beyond obtaining regulatory approval, a crucial element in the implementation of these tools into healthcare is the trust and support of the people that use them. An increased understanding by clinicians of artificial intelligence (AI) and ML algorithms provides an appreciation of the benefits, risks, and uncertainties they may bring, and improves the chances for successful adoption. Advances in DNA sequencing technologies have revolutionized biological research, enabling the analysis of DNA sequences on

a large scale. DNA sequencing helps unravel the genetic code, identify mutations, study genetic variations, trace evolutionary relationships, and diagnose genetic diseases. DNA plays a central role in fields such as genomics, evolutionary biology, genetic engineering, forensic science, and medicine. It serves as a foundation for understanding the complexities of life, exploring the diversity of species, and developing innovative approaches for disease treatment and prevention.

LITERATURE REVIEW

Elchin Asgarov (2024) heart disease is one of the most important problems the world faces. It is an ongoing problem and it is leading to the cause of death globally. To solve this issue, predicting early heart disease is important. This research focuses on supervised machine learning techniques as a potential tool for heart disease prediction. This study has done a comprehensive review of 30 articles published between 1997 to 2023 about machine learning techniques to predict heart disease. The common problem is that authors use different data sets, and different numbers of parameters to train and test these models. These two factors could affect the model's accuracy. To compare different models, I only used articles that analyse more than one method using the same data to prevent bias.

Ghada S. El-Taweel (2024) Analyzing big data, especially medical data, helps to provide good health care to patients and face the risks of death. The COVID-19 pandemic has had a significant impact on public health worldwide, emphasizing the need for effective risk prediction models. Machine learning (ML) techniques have shown promise in analysing complex data patterns and predicting disease outcomes. The accuracy of these techniques is greatly affected by changing their parameters. Hyperparameter optimization plays a crucial role in improving model performance. In this work, the Particle Swarm Optimization (PSO) algorithm was used to effectively search the hyperparameter space and improve the predictive power of the machine learning models by identifying the optimal hyperparameters that can provide the highest accuracy. A dataset with a variety of clinical and epidemiological characteristics linked to COVID-19 cases was used in this study.

Arduino A Mangoni (2023) The increasing access to health data worldwide is driving a resurgence in machine learning research, including data-hungry deep learning algorithms. More computationally efficient algorithms now offer unique opportunities to enhance diagnosis, risk stratification, and individualised approaches to patient management. Such opportunities are particularly relevant for the management of older patients, a group that is characterised by complex multimorbidity patterns and significant interindividual variability in homeostatic capacity, organ function, and response to treatment. Clinical tools that utilise machine learning algorithms to determine the optimal choice of treatment are slowly gaining the necessary approval from governing bodies and being implemented into healthcare, with significant implications for virtually all medical disciplines during the next phase of digital medicine.

Sahar Faezi (2023) These days, Internet coverage and technologies are growing rapidly, hence, it makes the network more complex and heterogeneous. Software defined network (SDN) revolutionized the network architecture and simplified the

network by separating the control and data plane. On the other hand, machine learning (ML) and its derivations have made the systems more intelligent. Many pieces of research papers have addressed ML and SDN. In this survey, we collected the papers published in Springer, Elsevier, IEEE, and ACM and addressed SDN and ML between 2016-2023. The research papers are organized based on the solutions, evaluation parameters, and evaluation environments to help those working on SDN and ML for improving the target functional or non-functional parameters. The research papers will be analyzed to extract the solutions, evaluation parameters and environments.

Aqil Tariq (2022) The development of smart network infrastructure of the Internet of Things (IoT) faces the immense threat of sophisticated Distributed Denial-of-Services (DDoS) security attacks. The existing network security solutions of enterprise networks are significantly expensive and unscalable for IoT. The integration of recently developed Software Defined Networking (SDN) reduces a significant amount of computational overhead for IoT network devices and enables additional security measurements. At the prelude stage of SDN-enabled IoT network infrastructure, the sampling-based security approach currently results in low accuracy and low DDoS attack detection. In this study, we propose an Adaptive Machine Learning based SDN-enabled Distributed Denial-of-Services attacks Detection and Mitigation (AMLSDM) framework.

Tong Cheng (2021) As an information-rich collective, there are always some people who choose to take risks for some ulterior purpose and others are committed to finding ways to deal with database security threats. The purpose of database security research is to prevent the database from being illegally used or destroyed. This study introduces the main literature in the field of database security research in recent years. First of all, we classify this study, the classification criteria are the influencing factors of database security. Compared with the traditional and machine learning (ML) methods, some explanations of concepts are interspersed to make these methods easier to understand. Secondly, we find that the related research has achieved some gratifying results, but there are also some shortcomings, such as weak generalization, deviation from reality.

Padmanabhan Nair (2021) In a previous report we had reported on the discovery of a novel bispecific immunoglobulin expressed by colonic epithelial cells as they transform into immunomimetic cells during exfoliation (Albaugh et al. (2020) Open Journal of Preventive Medicine, 10, 126-150). Colonic cells isolated from 0.5 gm aliquots of fresh stools (SCSR-10, Fecal Cell Isolation Kit, NonInvasive Technologies, Elkridge, MD) preserved at room temperature for up to one week, with viability of >85% were used to determine the number of cells expressing this novel bispecific immunoglobulin. Over the course of this period (18 years) we recognized that these cells opened the opportunity to investigate the expression of cell membrane biomarkers. As the applications grew, we introduced a new terminology, termed COPROCYTOBIOLOGY*. In this study, we surveyed a cohort of 58 free-living adults for the expression of the newly discovered bi-specific chimeric antibody. Almost all of the subjects showed a strong signal during flow-cytometric evaluation of their stool samples; averaging around 65%.

Lu Lu (2021) Despite great progress in simulating multiphysics problems using the numerical discretization of partial differential equations (PDEs), one still cannot seamlessly incorporate noisy data into existing algorithms, mesh generation remains complex, and high-dimensional problems governed by parameterized PDEs cannot be tackled. Moreover, solving inverse problems with hidden physics is often prohibitively expensive and requires different formulations and elaborate computer codes. Machine learning has emerged as a promising alternative, but training deep neural networks requires big data, not always available for scientific problems. Instead, such networks can be trained from additional information obtained by enforcing the physical laws.

Anuja Badeti (2020) In this study, we introduce a novel interpreting framework that learns an interpretable model based on an ontology-based sampling technique to explain agnostic prediction models. Different from existing approaches, our algorithm considers contextual correlation among words, described in domain knowledge ontologies, to generate semantic explanations. To narrow down the search space for explanations, which is a major problem of long and complicated text data, we design a learnable anchor algorithm, to better extract explanations locally. A set of regulations is further introduced, regarding combining learned interpretable representations with anchors to generate comprehensible semantic explanations. An extensive experiment conducted on two real-world datasets shows that our approach generates more precise and in sightful explanations compared with baseline approaches.

Maruf Pasha (2017) In medical imaging, Computer Aided Diagnosis (CAD) is a rapidly growing dynamic area of research. In recent years, significant attempts are made for the enhancement of computer aided diagnosis applications because errors in medical diagnostic systems can result in seriously misleading medical treatments. Machine learning is important in Computer Aided Diagnosis. After using an easy equation, objects such as organs may not be indicated accurately. So, pattern recognition fundamentally involves learning from examples. In the field of bio-medical, pattern recognition and machine learning promise the improved accuracy of perception and diagnosis of disease. They also promote the objectivity of decision-making process. For the analysis of high-dimensional and multimodal bio-medical data, machine learning offers a worthy approach for making classy and automatic algorithms.

Nature-Inspired Algorithms for Genome Assembly

Nature-inspired optimization algorithms are defined as a group of algorithms that are inspired by the behavior natural systems, including bio-inspired algorithms, swarm intelligence and evolutionary algorithms. Inspired by animal, insects' behaviors, biology and chemical reactions, those algorithms have provided many engineering, medical and bioinformatics solutions such as solving the DNA Fragment Assembly Problem. The genetic assembly is a critical step in any genomic project; it attempts in reconstructing a DNA sequence from a set of a large number of fragments taken obtained by biologists in the laboratory. DNA Fragment Assembly Problem is known to be an NP-hard combinatorial optimization problem; therefore, efficient approximate metaheuristics are required to solve such kinds of problems. The purpose of the study

presented in this section is to analyse and synthesize the existing nature inspired optimization algorithms in for genome assembly. Since the genome assembly is a particularly difficult problem in computational biology due to the problem's NP-hardness, the ideal solution cannot be found. So, it necessitates the use of metaheuristics and other computational techniques of intermediate complexity. Which their goal is to compare all possible solutions to an optimization problem in order to choose the best (feasible) one. They evaluate prospective solutions and perform a number of operations on them in an effort to find better alternatives in order to accomplish this.

Genetic Algorithm

In this section, we present a rudimentary of genetic algorithm

Genetic algorithm

The idea of GA (formerly called genetic plans) was conceived, as a method of searching centered on the principle of natural selection and natural genetics. Darwins theory was their inspiration, as they carefully learned the principle of evolution and applied the knowledge acquired to develop algorithms based on the selection process of biological genetic systems. The concept of GA was derived from evolutionary biology and survival of the fittest. Several parameters require the setting of values when implementing GA, but the most critical parameters are population size, mutation probability and crossover probability, and their interrelations.

Genetic Algorithm Operations

When a problem is given as an input, the fundamental idea of GA is that the pool of genetics specifically contains the population with a potential solution or better solution to the problem. GA use the principle of genetics and evolution to recurrently modify a population of artificial structures through the use of operators, including initialization, selection, crossover and mutation, in order to obtain an optimum solution. Normally, GA start with a randomly generated initial population represented by chromosomes. Solutions derived from one population are taken and used to form the next generation population. This is carried out with the expectation that solutions in a new population are better than those in the old population. The solution used to generate the next solution is selected based on its fitness value; solutions with a higher fitness value have higher chances of being selected for reproduction, while solutions with lower fitness values have a lower chance of being selected for reproduction. This evolution process is repeated several times until a set criterion for termination is satisfied. For instance, the criterion could be the number in the population or the satisfaction of the improvement of the best solutions.

Genetic algorithms in machine learning

AI Overview Genetic algorithms for feature selection in machine learning Genetic algorithms (GAs) are search heuristics inspired by the process of natural selection and evolution. In machine learning, they are used to find optimal or near-optimal solutions to complex problems by iteratively improving candidate solutions. GAs is particularly useful for optimization and search problems, especially when dealing with a large solution space where traditional methods struggle. Here's a breakdown of how they work:

1. **Initialization:** A population of potential solutions is created, often randomly.
2. **Fitness Evaluation:** Each solution (often called an individual) is assessed using a

fitness function, which measures how well it performs based on the problem's criteria.

3. **Selection:** The fittest individuals are selected for reproduction, meaning they are more likely to be chosen to create new solutions.

4. **Crossover (Recombination):** Selected individuals are combined to create new offspring, inheriting characteristics from both parents.

5. **Mutation:** Random changes are introduced into the offspring to maintain diversity within the population and prevent premature convergence to a local optimum.

6. **New Generation:** The offspring become the new generation, and the process of selection, crossover, and mutation is repeated until a satisfactory solution is found or a maximum number of generations is reached.

Genetic Entities and Initialization

A random population of individuals is initialized. An individual is an entity that is composed of a genome and a fitness. A population is a collection of individuals. The number of individuals in the population is specified by the "population size" hyperparameter, which remains constant along generations. The genome is the general structure of the solution that the algorithm is searching for. In this thesis, GAs is used to evolve FFNNs of a fixed architecture. In this case, the genome of each individual can be any network of the specified architecture. The evolution process aims at finding an individual whose genome is the network with the optimal parameters given the fitness metric. To do so, GAs efficiently searches the parameters space by selecting high fitness individuals and recombining their genomes piece-wise. At the beginning of the procedure, individuals of the population are initialized with random network weights.

Artificial intelligence and machine learning

AI includes the domains of ML, machine reasoning, and robotics. Whilst there is no formal definition of ML it can be considered as the subfield of AI which focuses on the development of algorithms that allow computers to automatically discover patterns in the data and improve with experience, without being given a set of explicit instructions. Rather than being developed or borrowing from one specific scientific field, ML sits at the intersection of statistics, mathematics and computer science, with analytic tools that transcend the boundaries across the three disciplines. A distinct feature of ML algorithms is their data-driven approach to learning, in contrast to rule-based models that rely on domain knowledge. ML algorithms include supervised learning (SL), unsupervised learning (UL), and reinforcement learning (RL) for diagnosis and prediction, phenotyping, and treatment recommendation, respectively. As the name implies, ML algorithms can learn (or improve) by receiving additional data either during training or after deployment. RL algorithms self-learn by using a trial-and-error approach to determine the best policy including decision making under conditions of uncertainty such as in the context of treatment recommendations.

Model complexity versus clinical interpretability

Earlier approaches to clinical decision support relied on using observational epidemiological data and traditional statistical techniques to develop regression-based risk prediction models such as the Framingham Risk score for the triaging of future cardiovascular events. Although results from these models provide high interpretability by virtue of the model's

coefficients, they also have important limitations. Such limitations include assumptions of the underlying data distributions, assumed linear and additive effects, an absence of interactions, a limited number of variables (features), and a dependence on domain expertise. These limitations also lead to a one-size-fits-all population level model designed mainly for identifying the predictive value of risk factors and the average risk for persons with a given combination of those risk factors. As such, the models are unsuitable for individualised risk prediction and more targeted (personalised) approaches to treatment recommendation, key features of the modern management of geriatric patients. Fully incorporating patient heterogeneity including a knowledge of existing disease, physical functioning, and intrinsic capacity requires moving away from previous domain-knowledge modelling approaches to data-driven models that adequately capture the full patient history, demographics, and clinical profile. Clinical decision support tools can, therefore, now be seen to exist on a continuous spectrum of model complexity from regression-based models to state-of-the-art Deep Learning (DL), whose better predictive accuracy is in general traded for their more limited interpretability.

The taxonomy of machine learning

There now exist many thousands of different ML algorithms that have been developed for prediction, pattern recognition or recommendation. The process of model selection, as it is commonly known, involves picking the best algorithm for the specific problem at hand. This is ultimately determined by the broad research goal (e.g., disease diagnosis, risk prediction, phenotyping, or treatment recommendation), the researchers' knowledge of algorithms that might each fit the purpose, and restrictions arising from the available data such as its volume and dimensionality. ML algorithms can be broadly divided into six different categories (SL, UL, semi-supervised learning, SSL, DL, RL, and Other) with numerous classes within each. It is important to note that the boundaries for many of these algorithms are fluid and each algorithm can potentially be classified under multiple subgroups. Therefore, used as much for illustration of the breadth and diversity of machine learning algorithms as it is for classification purposes. Supervised learning (SL) is used when predicting specified outcomes from a collection of predictors. The data require labelling of the outcome, and training with the labelled data. SL creates an automated system that determines whether items of interest (e.g., patient clinical features) belong to a specific class (e.g., presence or absence of disease). Unsupervised learning (UL) is used when only data without specific outcomes and labelling is available. The similarity of observations within the data is assessed to divide the data into distinct groups without previous labels.

CONCLUSIONS

Complementing these, the genetic algorithm provided a robust framework for optimizing complex problems that involve multiple variables and constraints — such as routing paths, scheduling, and resource allocation. Its bio-inspired approach of natural selection allowed for continuous improvement of network strategies by evolving better solutions over successive generations. The combination of GA with ML, such as in hyperparameter tuning, feature selection, or policy optimization, demonstrated a significant enhancement in both learning efficiency and network adaptability. This will be useful for

researchers working in the areas of health analytics, intrusion detection, disease prediction in crops, community detection and many more. For each algorithm, detailed insights and highlights were provided, outlining their characteristics, strengths, and potential applications. Recent advancements in the literature also were examined, considering how these algorithms have evolved to address the challenges of genome assembly problem. In the future, the suggested DNA fragment assembler could be further developed by leveraging both machine learning techniques and nature-inspired algorithms, having the potential to be adapted into a parallel version using parallel programming frameworks like MapReduce.

REFERENCES

1. *Arduino A Mangoni (2023), "A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future", Aging Clin Exp Res., issn: 1720-8319, vol. 35(11), pages. 2363-2397. doi: 10.1007/s40520-023-02552-2*
2. *Maruf Pasha (2017), "Survey of Machine Learning Algorithms for Disease Diagnostic", Journal of Intelligent Learning Systems and Applications, issn: 2150-8410, vol. 9, pages. 1-16.*
3. *Elchin Asgarov (2024), "A Comprehensive Analysis of Machine Learning Techniques for Heart Disease Prediction", Open Access Library Journal, issn:2333-9721, vol. 11, pages. 1-17.*
4. *Sahar Faezi (2023), "A Comprehensive Survey on Machine Learning using in Software Defined Networks (SDN)", Human-Centric Intelligent Systems, ISSN no:2667-1336, Vol.3(10), DOI:10.1007/s44230-023-00025-3*
5. *Ghada S. El-Taweel (2024), "Particle Swarm Optimization-Based Hyperparameters Tuning of Machine Learning Models for Big COVID-19 Data Analysis", Journal of Computer and Communications, issn: 2327-5227, vol. 12, pages. 160-183.*
6. *Tong Cheng (2021), "The Overview of Database Security Threats' Solutions: Traditional and Machine Learning", Journal of Information Security, issn: 2153-1242, vol. 12, pages. 34-55.*
7. *Anuja Badeti (2020), "Ontology-based Interpretable Machine Learning for Textual Data",2020 International Joint Conference on Neural Networks (IJCNN), ISSNno:2161-4407, DOI: 10.1109/IJCNN48605.2020.9206753*
8. *Lu Lu (2021), "Physics-informed machine learning", Nature Reviews Physics, ISSNno:2522-5820, DOI:10.1038/s42254-021-00314-5*
9. *Aqil Tariq (2022), "Adaptive Machine Learning Based Distributed Denial-of-Services Attacks Detection and Mitigation System for SDN-Enabled IoT", Sensors, ISSNno:1424-8220, Vol.22(7), Pages.28. DOI:10.3390/s22072697*
10. *Padmanabhan Nair (2021), "Germline Deletion of the Expression of a Human Bispecific Mucosal Immunoglobulin: Genetic Predisposition to Cancer and Communicable Diseases Predominantly among African-Americans", Open Journal of Preventive Medicine, ISSN no:2162-2485, Vol.11, No.10, Pages.383-390.*