METHODS FOR IMPROVING MACHINE LEARNING TECHNIQUES

SRIKANTH NANGINENI
Research Scholar
Dept of Computer Science & Engg
Arni University-Himachal Pradesh

DR. PRASADU PEDDI
Research Supervisor
Dept of Computer Science & Engg
Arni University-Himachal Pradesh

DR. S. RAMA SREE
Co-Supervisor
Professor
Dept of Computer Science & Engg
Aditya Engineering College-A.P

ABSTRACT

In the current age of the Fourth Industrial Revolution (4IR or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cyber security data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding smart and automated applications, the knowledge of artificial intelligence (AI), particularly, machine learning (ML) is the key. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the deep learning, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale. In this study, we present a comprehensive view on these machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application. Thus, this study's key contribution is explaining the principles of different machine learning techniques and their applicability in various real-world application domains, such as cyber security systems, smart cities, healthcare, e-commerce, agriculture, and many more. We also highlight the challenges and potential research directions based on our study.

Keywords: *Internet of Things (IoT) data, machine learning, Artificial Intelligence (AI), healthcare.*

INTRODUCTION

According to machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. In machine learning, algorithms are used to distinguish between meaningful and irrelevant patterns in data. Examples of machine learning applications include the provision of accurate medical diagnostics (e.g. breast cancer), real-time map-based monitoring of environmental disasters (e.g. forest fires) and sensory monitoring in the industrial process (e.g. mechanical failure). Machine learning as a kind of artificial intelligence (AI) which compose available computers with the efficiency to be trained without being veraciously programmed. ML learning interest on the extensions of computer programs which is capable enough to modify when unprotected to newfangled data. The evolution of machine learning is comparable to that of data mining. Both data mining and machine learning consider or explore from end to end data to assume for patterns. On the other hand, in choice to extracting data for human knowledge as is the case in data mining applications;

machine learning generate use of the data to identify patterns in data and fine-tune program actions.

LITERATURE REVIEW

Fernando Timoteo Fernandes (2023) Machine learning algorithms are being increasingly used in healthcare settings but their generalizability between different regions is still unknown. This study aims to identify the strategy that maximizes the predictive performance of identifying the risk of death by COVID-19 in different regions of a large and unequal country. Of all patients with a positive RT-PCR test during the period, 2356 (28%) died.

Kadiyala Ramana (2022) Huge amounts of data are circulating in the digital world in the era of the Industry 5.0 revolution. Machine learning is experiencing success in several sectors such as intelligent control, decision making, speech recognition, natural language processing, computer graphics, and computer vision, despite the requirement to analyze and interpret data. Due to their amazing performance, Deep Learning and Machine Learning Techniques have recently become extensively recognized and implemented by a variety of real-time engineering applications. **Mohamed LAZAAR (2022)** Machine learning performances always rely on relevant phase of pre-processing, that includes dataset cleaning, cleansing and extraction. In this study, we focus on feature selection using embedded-based methods in order to minimize computational time and complexity of ML models. Embedded-based methods combine advantages of both filter-based and wrapped-based methods, in terms of studying the importance of features while executing the model and their reduced time of execution.

Iqbal H. Sarker (2021) Deep learning (DL), a branch of machine learning (ML) and artificial intelligence (AI) is nowadays considered as a core technology of today's Fourth Industrial Revolution (4IR or Industry 4.0). However, building an appropriate DL model is a challenging task, due to the dynamic nature and variations in real-world problems and data. Moreover, the lack of core understanding turns DL methods into black-box machines that hamper development at the standard level.

Mohammed H. Alsharif (2020) Machine learning techniques will contribution towards making Internet of Things (IoT) symmetric applications among the most significant sources of new data in the future. In this context, network systems are endowed with the capacity to access varieties of experimental symmetric data across a plethora of network devices, study the data information, obtain knowledge, and make informed decisions based on the dataset at its disposal.

APPLICATIONS OF MACHINE LEARNING

In the current age of the Fourth Industrial Revolution (4IR), machine learning becomes popular in various application areas, because of its learning capabilities from the past and making intelligent decisions. In the following, we summarize and discuss ten popular application areas of machine learning technology.

Predictive analytics and intelligent decision-making: A major application field of machine learning is intelligent decision-making by data-driven predictive analytics. For instance, identifying suspects or criminals after a crime has been committed, or detecting credit card fraud as it happens.

Cybersecurity and threat intelligence: Cybersecurity is one of the most essential areas of Industry 4.0. Machine learning has become a crucial cybersecurity technology that constantly learns by analyzing data to identify patterns, better detect malware in encrypted traffic, find insider threats, predict where bad neighborhoods are online. For instance, clustering techniques can be used to identify cyber-anomalies, policy violations, etc.

Internet of things (IoT) and smart cities: Internet of Things (IoT) is another essential area of Industry 4.0., which turns everyday objects into smart objects by allowing them to transmit data and automate tasks without the need for human interaction.

Traffic prediction and transportation: Transportation systems have become a crucial component of every country's economic development.

Healthcare and COVID-19 pandemic: Machine learning can help to solve diagnostic and prognostic problems in a variety of medical domains, such as disease prediction, medical knowledge extraction, detecting regularities in data, patient management, etc.

CHALLENGES AND RESEARCH DIRECTIONS

Our study on machine learning algorithms for intelligent data analysis and applications opens several research issues in the area. In general, the effectiveness and the efficiency of a machine learning-based solution depend on the nature and characteristics of the data, and the performance of the learning algorithms. To collect the data in the relevant domain, such as cyber security, IoT, healthcare and agriculture discussed in Sect.

TYPES OF MACHINE LEARNING

Categorized machine learning algorithms into supervised, unsupervised and reinforcement learning algorithms: Figure one present the classification in a pictorial form:

i. Supervised Learning

Supervised learning is a core area of machine learning. In supervised learning the goal is to learn a mapping from the input to the output. The input is data that describes a collection of individual objects of interest and are commonly referred to as instances or examples. The output is some outcome or result provided by a supervisor. Classification is a form of supervised learning whereby a mapping (or discriminant function) separates different classes of the instances. The diff erent classes are specified by the output which, in machine learning, is termed as the class label.

ii. Unsupervised Learning

According to this machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

iii. Reinforcement machine learning algorithms

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

METHODOLOGY

Based on research on existing methods and metrics, an iterative knowledge discovery process will be started to answer the given research questions. This process includes the determination of quality criteria for translated documents, the implementation of needed metrics and algorithms as well as the optimization of the machine learning approaches to solve the given task optimally. It is important to note that this process is of iterative nature, since the criteria and attributes as well as their impact on translation quality and classification possibilities will be determined by evaluating the algorithms' results using a database of technical documents and their translations. The used data set will range from automated translations of technical documents using computerized translation systems to manual and professional translations. Furthermore, during this iterative process, the methods and algorithms used will be continually changed and optimized to achieve the best possible results. Finally, the process and results will be critically reviewed, evaluated and compared to one another. The limits of automated translations with the current state of the art will be pointed out and a prospect for possible further developments and studies on this topic will be given.

RESULTS

The results of this study have been shown in this section. The performances of four machine learning techniques on all the combinations of clinical attributes were examined one by one. Performance metrics namely accuracy, specificity, sensitivity have been tabulated in Tables 1-3 respectively. Among all the possible combinations of clinical attributes, the combination of features which accounted for the highest performance was identified.

Table 1: Highest Classification accuracy achieved by various ML methods

ML technique	Highest Classification Accuracy	Combination of attributes
Logistic Regression	86.2%	Age, Diabetes, TC, HDL, EX, FH, HT, HR, AL, SM, Gender, WC
k-NN	87%	Diabetes, HT, TC, SM, HD, TG, Gender, ST, AL, SF, BMI
Support Vector Machine	86.8%	TC, HDL, EX, FH, gender, BMI, FBS, CRP,CP, TG
Random Forest	90.1%	Age, Diabetes, LDL, SM, HD, BMI, ST, AL, SC, Gender

It is clear from Table 1 that Logistic Regression based system attained a maximum accuracy of accuracy of 86.2% when trained on input attributes like age, diabetes, total cholesterol, HDL, Exercise, family history, hypertension, heart rate, alcohol, smoking, gender, and waist circumference.

SVM performed better than Logistic Regression attaining an accuracy of 86.8%when trained on total cholesterol, HDL, exercise, family history, gender, BMI, Fasting blood sugar, CRP, and chest pain.

Table 2: Highest Sensitivity achieved by various ML methods

ML technique	Highest Sensitivity	Combination of attributes
Logistic Regression	87.2%	Age, Diabetes, TG, EX, FH, ST, CP, HD, BMI, SF
k-NN	85.4%	Age, Gender, TC, HDL, LDL, FH, AL, SM, ST, CRP, HR, WC
SVM	83.5%	Gender, Diabetes, TC, HT, EX, FH, HD, WC, FBS,GGT, LDL
Random Forest	91%	Age, Gender, Diabetes, HT, EX, AL, SM, BMI, HDL, HR, SC

Highest sensitivity attained using SVM was 83.5% while that of k-NN was observed to be 85.4%. Logistic regression when fed with input attributes age, HbA1c, Triglycerides, exercise, family history, stress, chest pain, diet habits, BMI, and serum fibrinogen attained the highest sensitivity of 87.2%. The highest sensitivity attained by Random Forest was observed to be maximum at 91% when the combination of age, gender, HbA1c, hypertension, exercise, alcohol, smoking, BMI, HDL, heart rate, serum creatinine was fed as input features. The highest specificity attained using k-NN and SVM were nearly 86% and 86.2% respectively. The highest specificity scored by Logistic regression was 88.7% when the input attributes were age, hypertension, HbA1c, diet habits, BMI, family history and stress/anxiety, AST/ALT ratio, total cholesterol, triglycerides, exercise, and smoking. The highest specificity attained using RF was 93%.

Table 3 Highest specificity achieved by various ML methods

ML technique	Highest specificity	Combination of attributes
Logistic Regression	88.7 %	Age, HT, Diabetes, HD, BMI, FH, ST AST/ALT, TC, TG, EX, SM
k-NN	86.2 %	Gender, TC, FH, SM, AL, BMI, FBS, LDL, CRP, FEV1, Diabetes, CP, WC
SVM	86 %	Age, Gender, HT, FH, BMI, HDL, HR, EX, AL, LDL,

		FBS
Random Forest	93 %	Age, Gender, Diabetes, HD, SM, AL, ST, EX, TC, CP,TG,

It is clear from Table 4 that accuracy was greatly dependent on attributes like gender, BMI, Total cholesterol, Diabetes (HbA1c>7) and alcohol consumption habits. Specificity was majorly affected by age, gender, BMI, total cholesterol, Diabetes (HbA1c>7), alcohol consumption, family history and exercise. The attributes which tend to increase the sensitivity were family history, exercise, age, gender, and Diabetes (HbA1c>7).

Table 4: Role of clinical attributes on performance

Table 8 Role of clinical attributes on performance

Attributes	**Attribute Code**	**Occurrence Highest Accuracy**	**Occurrence Sensitivity**	**Occurrence Specificity**	**Total Frequency**
Age	Age		3		8
Gender	Gender		3		10
Body Mass Index	BMI		2		8
Waist Circumference	WC		1		4
Cholesterol Levels	TC		2		8
HDL cholesterol	HDL		2		5
LDL cholesterol	LDL		2		5
Triglycerides	TG		1		5
Hypertension	HT		2		6
Diabetes	Diabetes		3		9
Fasting Blood Sugar	FBS		1		3
Heart Rate	HR		1		3
FEV1	FEV		0		1
gamma GT	GGT		1		1
C-reactive protein (CRP)	CRP		1		3
Serum fibrinogen	SF		1		2
Serum creatinine	SC		1		2

AST/ALT Ratio	AST/ ALT		(		1
Chest Pain	CP		1		4
Alcohol	AL		2		8
Smoking (last5 years)	SM		2		6
Exercise (Weekly 3 Hours)	EX		3		8
Stress	ST		2		6
Family History CVD	FH		3		8
Healthy Diet	HD		2		6

The significant noninvasive clinical attributes identified in this study for heart disease prediction are gender, age, body mass index, hypertension, Diabetes (HbA1c>7), alcohol consumption, family history, total cholesterol, exercise, smoking, intake of healthy diet and stress/anxiety in life.

CONCLUSION

In this study, we have conducted a comprehensive overview of machine learning algorithms for intelligent data analysis and applications. According to our goal, we have briefly discussed how various types of machine learning methods can be used for making solutions to various real-world issues. A successful machine learning model depends on both the data and the performance of the learning algorithms. The sophisticated learning algorithms then need to be trained through the collected real-world data and knowledge related to the target application before the system can assist with intelligent decision-making. We also discussed several popular application areas based on machine learning techniques to highlight their applicability in various real-world issues. Finally, we have summarized and discussed the challenges faced and the potential research opportunities and future directions in the area.

Reference

1. *Alex Ng (2020),"Cybersecurity data science: an overview from machine learning perspective",Journal of Big Data,ISSNno:2196-1115,Vol.7,https:// doi.org/10.1186/s40537-020-00318-5*
2. *Asif Irshad Khan (2020),"IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model",Symmetry,ISSNno:2073-8994,Vol.12 (754).DOI:10.3390/sym12050754*
3. *Fernando Timoteo Fernandes (2023),"Improving the performance of machine learning algorithms for health outcomes predictions in multicentric cohorts", Scientific Reports,ISSNno:2045-2322,Vol.13,https://doi.org/10.1038/s41598-022-26467-6*
4. *Iqbal H. Sarker (2021),"Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions",SN Computer Science,ISSNno:2661-8907,Vol.2,https://doi.org/10.1007/s42979-021-00815-1*
5. *Kadiyala Ramana (2022),"A Review on Machine Learning Strategies for Real-World Engineering Applications",Mobile Information Systems,ISSNno: 1875-905X,Vol.2022,https://doi.org/10.1155/2022/1833507*
6. *Mohamed LAZAAR (2022),"Improving Machine Learning Models for Malware Detection Using Embedded Feature Selection Method",IFAC-Papers OnLine,ISSNno:2405-8963,Vol.55,Issue.12,Pages.771-776.https://doi.org/10. 1016/j.ifacol.2022.07.406*

7. *Mohammed H. Alsharif (2020),"Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends",Symmetry,ISSNno:2073-8994,Vol.12(1),Pages.88.https://doi.org/10.3 390/sym12010088*
8. *Raza Nowrozy (2021),"AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions",SN Computer Science, ISSNno:2661-8907,Vol.2,https://doi.org/10.1007/s42979-021-00557-0*
9. *Vinay Chamola (2022),"Uniting cyber security and machine learning: Advantages, challenges and future research",ICT Express,ISSNno:2405-9595, Vol.8,Issue.3,Pages.313-321.*
10. *Wissam Mallouli (2023),"The Role of Machine Learning in Cybersecurity", Digital Threats: Research and Practice,ISSNno:2576-5337,Vol.4,Issue.1, Pages.1–38.https://doi.org/10.1145/3545574*