A Comparative Theoretical and Empirical Analysis of Machine Learning Algorithms

Shailja Gupta
Manav Rachna University. E-mail: shailja@mru.edu.in

Manpreet Kaur
Manav Rachna University. E-mail: manpreet@mru.edu.in

Sachin Lakra
Manav Rachna University. E-mail: sachin@mru.edu.in

Yogesh Dixit
Manav Rachna University. E-mail: yogesh087dixit@gmail.com

Received April 07, 2020; Accepted May 26, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V17I1/WEB17011

Abstract

With the explosion of data in recent times, Machine learning has emerged as one of the most important methodical approaches to observe significant insights from the vast amount of data. Particularly, it is witnessed that with the alarming rise in the volume of unstructured data on the world wide web, machine learning algorithms can be applied in a wide number of domains to solve various problems related to understanding humans. At the onset, this paper introduces the field of machine learning, classic learning approaches, and machine learning algorithms. A theoretical comparison study of state of the art algorithms is carried based on their logic, characteristics, weaknesses, strengths, and kind of applications in which these algorithms can be used. The study is expected to help buddy researchers who are in the beginning to work in this area.

Keywords

Machine Learning, Supervised Machine Learning, Unsupervised Machine Learning.

Introduction

Machine Learning refers to learning from experience to make future predictions [1]. It is a sub-field of artificial intelligence (AI). Without explicit programming, artificial intelligence allows the researchers to develop methods with self-learning capabilities whose accuracy is expected to improve with time [2]. For example, machine learning

techniques have been implemented in the field of optical communication to deal with unexpected growth in network complexity in the last few years. According to a survey [9][10], 80 percent of worldwide data will be unstructured by 2025.

With the revolution of data, machine learning has become important and is being employed in almost every domain of computer science, statistics, biology, psychology, finance, language processing, transportation, and many other domains particularly in the field of computer science, sub-fields like information retrieval [3], computer vision [4], pattern recognition [5], sentence recognition [6], handwriting recognition [7], computational biology [8] are a few among numerous fields that are using machine learning algorithms. The accuracy has been improved in all these domains due to the presence of massive data which leads to improved decision making in business and problem-solving. We are in the early days of a "data revolution", with the increasing data storage capacity and availability of computation power at a low cost. It has been observed that people have started expressing themselves on online platforms like social networking sites, blogging websites, consumer goods online platforms more than ever before. This has lead to the accumulation of different types of data (text, audio, video, images) in an unstructured form which has got the attention of the researchers towards finding useful patterns in it.

Various machine learning algorithms can be used to find solutions for research problems. For example: using logistic regression, naïve bayes, random forest, recurrent neural networks for understanding human expression in the text, image, and speech. These forms of human expression have always been found complex in terms of analysing due to the complexity of the language. The machine learning algorithms can be used to determine abusive and offensive words in sentences, translators (like Google translator), creating chatbots like Alexa and Google Home.

In this paper, we will be exploring machine learning algorithms i.e. supervised and unsupervised machine learning algorithms. In supervised Algorithms, models are built and trained using labelled data. Labelled data refer meaningful numerical or textual tags that are applied to the group of samples. The algorithms that fall in the category of supervised machine learning algorithms namely, logistic regression, naïve bayes, random forest, decision trees etc. are used to solve the classification problems like spam detection, detection of disease, and regression problems like predicting house prices or property prices, predicting stock market prices and so on. in unsupervised machine learning algorithms, models are built using unlabelled data. It considers a scenario where the data is huge and it becomes difficult to analyse this huge data. The algorithms that fall in the

category of unsupervised Machine Learning Algorithms namely Hierarchical clustering, k-means clustering, etc., can be used for solving problems such as grouping customers according to their purchasing behaviour, determining what products a customer buys together, and similar other problems.

The contributions of the paper are:

a. An introductory overview of the most popular supervised and unsupervised algorithms is presented and the advantages and disadvantages of using these algorithms have been discussed.
b. An overview of the applications of machine learning algorithms in different fields of research is then discussed. This will help the reader to develop an understanding of how to select a suitable algorithm to solve s particular problem. The study is expected to help buddy researchers who are in the beginning to work in this area.

In this paper, we have presented a comparative study of many popular algorithms and applications of machine learning on a single platform. Due to the vast expanse of this field, it is not feasible to carry out a complete study of the field. Therefore, we have tried our best to present a survey which would be beneficial for any beginner in this field.

The following paper is divided into sections as follows. Section 2 gives a brief literature survey in the field of machine learning. We discuss the classification of machine learning algorithms with their advantages and disadvantages in section 3. Section 4 discusses various applications of the field of the machine learning algorithm.

Related Work

Before diving into the related work of machine learning, we would like to introduce machine learning for budding researchers.

Types of Learning:
A. Supervised/ Task Driven Machine Learning Algorithms

In supervised machine learning the data given for the training of the algorithm is labeled data, which is used to produce a machine learning model. The dataset is split into two parts: a training set and a test set. With the help of the training set, a classifier is used to learn the patterns to predict the output. The performance of the classifier is evaluated using the test set. For example, logistic regression is a supervised learning technique that is used to predict the outcome in binary classification. Some of the advantages of using supervised learning over rule-based learning are:

a. *Adaptive Intelligence:* With the availability of new knowledge, old knowledge can be discarded or changed i.e., building rules is spontaneous.
b. *Automation*: Automated learning techniques are giving competitive results when compared to human intelligence. For example, spam filtering automatically classifies incoming emails as spam or not without any human intervention.
c. *Continuous Improvement:* Machine Learning (ML)models continue to learn with experience and more data. With the availability of more data, it is capable of making more accurate and efficient predictions. For example: as the data for weather prediction increases, the weather prediction results are more accurate and fast.
d. *Identification of patterns*: ML is capable of handling large volumes of data and finding patterns in the same that is otherwise beyond the capacity of a human to do manually. For example, finding buying patterns of a user based on buying history of a customer, on an e-commerce site like Amazon.
e. *Handling dynamic data*: ML is good at handling data that is available on the fly. It is capable of handling a variety of data like weather data and data with high dimensionality like gene expression.

Supervised Learning Algorithms can be further classified into classification modelling and regression modelling. Fig. 1 represents a flow diagram of the classification of machine learning algorithms into classification and regression algorithms.

Classification Modelling: Classification modelling methods are used to predict a discrete class label. A classification problem with two class labels is called binary classification. For example: Labelling an email as spam or not, predicting high-risk patients from low-risk patients and more. These modelling methods are evaluated by estimating their accuracy using a confusion matrix as explained in section (V-B). Multi-class classification is a problem with more than two classes. For example, to predict rainfall as low, medium or high.

Regression Modelling: Regression modelling methods are used to predict a continuous value and are considered as more suitable for the problem statements having continuous values for input and output. These problems can be analysed by generating a hyperplane (line in case of a 2D matrix). Multiple regression is a problem having multiple input variables. These modelling methods are evaluated by estimating mean square error (MSE), root mean square error (RMSE) and so on as described in section (V-B).

Machine Learning is used as an integral part of artificial intelligence. It has been considered suitable for problems with an ideal number of results and observations and insufficient theoretical knowledge [9]. It has been used to design algorithms, features and hidden patterns by past learning and by finding trends in data. Kubat et al., [10] have discussed various activities that are involved in machine learning algorithms using a

dataset on the detection of oil spills at sea. The authors have tried to resolve the common problem of unbalanced datasets by proposing two algorithms: one-sided selection method to reduce the number of negative examples and the SHRINK algorithm to remove noisy data in negative examples. Ben-Hur et al., [11] presents a new concept of boundary support vector which is based on Support vector machines.

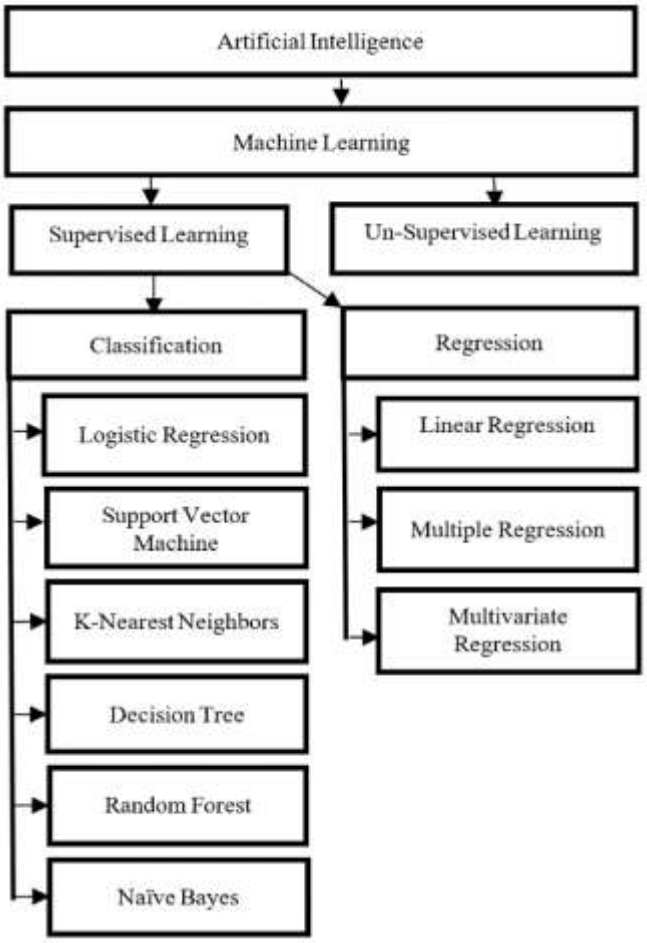

Fig. 1. Supervised Machine Learning Algorithms

The authors have utilized clustering parameters like Gaussian kernel and soft margin to present various clustered solutions. They have also tried to evolve from the common geometric shapes to arbitrary cluster shapes using the iris dataset. Schapire, Robert E [12] has focused primarily on the AdaBoost algorithm by reviewing some recent work on boosting algorithms. They discuss the contribution of AdaBoost on Logistic regression, loss minimization problem, multiclass problems, and regularization problems. They further compare the error rates for AdaBoost on text classification algorithms like Naïve Bayes, probabilistic tf-idf and so on using Reuters news wire and headlines corpora.

Jerez et al., [13] have performed experiments on missing data imputations using statistical methods and machine learning methods. The imputation data is one where due to criticality of data it becomes mandatory to use all the data without leaving out the data that has missing values. The experiments performed as a part of the El-Almano-I (breast cancer dataset from thirty-two hospitals) project showed that the machine learning algorithms like multiple layer perceptron, self-organization maps, and KNN outperform imputation statistical methods like mean, hot deck, and multiple imputations. Otukei et al., [14] have performed experiments on determining land coverage change which is one of the most important problems in remotely sensed data. They compared the performance of machine learning algorithms like Support vector machines, decision trees and maximum likelihood class on 1986 and 2001 Landsat TM and ETM+ data. The results of the comparison among the three showed that decision trees provide higher accuracy on pixel-based land classification problems.

Caesarendra et al., [27] has proposed a combination between logistic regression and relevance vector machine (RVM) to assess the degradation of the machine. The authors evaluate the performance using normal to failure data of the bearing present in the machine. The method has been evaluated on RMSE (root mean square error) and correlation to show its effectiveness as a machine degradation detection model. Chen et al., [15] has proposed a novel method Fuzzy Rough SVM for improving hard margins of SVM and improving noise in the datasets. The authors have considered eight datasets (Iris, Pima, Bupa, Ionosphere, New_Thyroid, Tae, Sonar, Wdbc) from UCI and compare them on hard margin SVM, soft margin SVM, Fuzzy SVM, and Fuzzy Rough SVM. The experiments conclude that FRSVM can be used over hard margin SVM, soft margin SVM, and Fuzzy SVM.

Mandal, Indrajit [16] has proposed a clinical healthcare recommender enterprise system based on multiple classifier systems. Enterprise systems are ones that provide clinical recommendations, nurse training and clinical quality control. They have implemented a hybrid machine learning ensemble model based on random subspace and random forest algorithms to overcome traditional association rule-based methods for the same. The experimental setup using five benchmark data shows that the proposed system performs promisingly against the present state of art approaches.

Mandal, Indrajit, and N. Sairam [17] have proposed a novel inference system for the detection of Parkinson's disease (PD). The inference system proposed used support vector machines and ranker algorithm for feature selection. Other methods like LogitBoost, Furia, Pregasor, Bayesian network, SVM, ANN, Boosting methods have been applied to

PD dataset to conclude that F-measure, kappa value and accuracy are efficient methods to check the accuracy of classification algorithms over ROC, main entropy gain and test region coverage. Škrinárová, J., Huraj, L., Siládi, V [18] have proposed a novice model of neural tree architecture for the classification of the resources of a computer grid. They have used a tree for the classification of the hardware portion of the dataset and for identifying patterns of software identifiers. They have also proposed a Particle Swarm Optimizer for task scheduling in a computer grid for effective mapping of resources to large tasks.

Min, Fan, Qinghua Hu, and William Zhu [19] have proposed a feature selection model over the existing methods for the problem of test cost constraint. Feature selection is a problem to identify those features that can find effective results from the dataset.

Table 1. Classification Algorithms in Supervised Learning

Supervised Algorithm (Classification)			
Algorithm	**Characteristics**	**Advantages**	**Disadvantages**
Logistic regression (C)	• Used for binary classification. • Prediction can be done on both categorical and numerical. • It gives a logistic curve as an output with the values limited between 0 and 1 [21].	• Efficient for numerical and categorical classification problems. • Don't require scaled input features. • Easy to regularize. • Handles non-linear effects.	• Boolean values only. • Not suitable for predicting the values of a binary value.
Naïve Bayes (C+R)	• It is a statistical method for classification which is based on the Bayes theorem method. • It is used in text classification, spam filtering, etc. [22].	• Easy to implement. • Efficient results in most of the applications. • Less training data is required • Used for classification of binary and multiple class.	• A lot of hyper-parameters are required. • It assumes features are independent. • It takes strong assumptions.
K-Nearest Neighbors (C+R)	• Also known as lazy learning, K-Nearest Neighbors simply stores the instances of the training data and does not construct a general internal model. • It classifies new cases based on distance functions or similarity measures [23] [24].	• Robust to noisy training data. • Effective for large training data. • Simple to implement. • No training phase is required. • Easily handles complex models.	• Need to determine the value of k. • High computation cost. • Difficult to apply for large dimension problem.
Decisions tree (C+R)	• It builds a hierarchical structure consisting of nodes and branches for the classification models. • It is an incremental process where the dataset is broken down to a smaller part at each level. • A root node determines the best predictor and leaf nodes determine classification or decision from the given dataset [25].	• Simple to understand and visualize. • Categorical and numerical data can be handled. • Variable feature selection is available. • Date preparation is easy.	• Create complex trees. • Decision trees can be unstable. • Highly prone to sampling error due to overfitting of data.
Random forest (C+R)	Random forest is a meta-estimator that fits several decision trees on various sub-samples of the datasets [26].	• Reduction in over-fitting. • More accurate than a decision tree.	• Slow in real-time prediction. • Difficult to implement. • Complex algorithm.
Support vector	It classifies the data by finding a hyperplane between the	• Effective in high	• Does not directly

machine (C+R)	margins of two or more classes. The training data in support vector machine is represented as points on the plane [26].	dimensional spaces. • Memory efficient. • Multiple kernels functions are available for various decision problems.	estimate the probability. • Calculated using expensive validation. • Feature size is greater than the sample size.
Artificial Neural Network	• An artificial neural network is designing a machine having intelligence similar to a human being. • The human brain is made of neurons. In ANN each neuron is connected to another neuron with certain coefficients [31][32].	• The ability to handle datasets with high dimensionality features. • Ability to handle documents with contradictory and noisy data. • Have Fault tolerance. • It has distributed memory and parallel processing capabilities.	• Consume high physical memory and CPU. • Not easy to understand for everyone. • It requires parallel processing processors. • The behavior of the network is still unexplained.

For this task, the authors have used datasets from zoology, society (voting), botany and gaming. Using these datasets, they have developed a heuristic method that has been proved to be effective using an experimental setup for the classification problem over backtracking methods. Shu et al., [20] took forward the research of fake news detection by measuring user sharing behaviour and grouping users who share fake and real news.

The authors in [20] have then analysed explicit and implicit profile features like age, profile image, location of the user, association to political parties and so on to determine if the user profile contributes to posting fake news online. They have estimated the accuracy using parameters like Recall and Precision of the confusion matrix. In this section, we are going to discuss related work in classification and regression modelling. Supervised machine learning algorithms are being applied to solve many classification and regression problems. Logistic Regression has been applied for multiclass classification the on Amazon-product review dataset [21]. The results of this NLP (natural language processing) problem have shown that the part of speech feature extraction methods give better results in terms of accuracy than the n-gram method on the same.

Naïve Bayes method has been used by the author in to find name ambiguity while downloading research papers [22]. They have collected citations with attributes like co-author name, title word of research paper and title of journal or conference proceedings from the homepage of "J Anderson" and "J Smith" and from dblp computer science bibliography. The results showed that NB has better accuracy, standard deviation, and p-value than the k-means method.

K-Nearest neighbor algorithm based on ML has been applied document mining on the Reuters-21578 dataset where the dataset is split into 9603 training data documents and

3299 test documents [23]. The authors have applied KNN, Term Graph and Naïve Bayes algorithms on the dataset after preprocessing the data using bag of words, stop-word removal, tf-idf, case folding and normalization. The results generated showed that the accuracy obtained by using KNN is much higher compared to the other two methods.

Another attempt has been made for text classification using the K-Nearest Neighbor (KNN) model, a machine learning model depends upon the problem of selecting a good value for k. It is considered less efficient when it comes to classifying large datasets. As a solution to this problem, a new model for KNN has been presented [24] where the detection of k has been done automatically. They have also tested this model on various public UCI datasets to conclude that the new model shows optimal accuracy than the original KNN model. The new model makes the classifier faster by reducing the dependency of the model on k value.

The authors compare the efficiency and complexity of k-means and decision tree algorithms using a student dataset [25]. The dataset is a small sample dataset and consist of eight student attributes. The results generated shows that with the increasing number of attributes in the train set, the accuracy and complexity increases, which leads to an increase in the level of prediction. As the complexity of the decision, trees is higher, so the prediction of decision trees is better in comparison to the k-means algorithm.

A comparison between SVM (support vector machine) and RF (random forest) has been done on Intrusion detection system using KDD'99 dataset [26]. The results has been based on precision and false-negative rate (FNR). FNR is important along with precision as misclassification of an intrusion with a non-intrusion leads to non-tolerance. The results showed that RF performs better over SVM in terms of precision and false-negative result. It has been found that SVM performs better than RF in terms of training accuracy and RF performs better than SVM in terms of testing accuracy. The time taken by Random forest algorithm to classify a problem is less as compared to SVM. Most of the machine learning algorithms are solved using supervised learning models. The authors have discussed the issues faced by supervised learning methods and have tried to provide a solution for the same. They have discussed artificial intelligence-based models, support vector machines and elaborate the ensemble classification model. They have published results based on the accuracy of the above said machine learning algorithms on 45 different datasets from the UCI repository.

An artificial neural network (ANN) is an electronic model that provides a biological neuron-like structure i.e., it is based on the structure of neurons in the brain. The authors

have presented a study of artificial neural networks and have discussed theoretically their working and training [31]. They have further described supervised and unsupervised training in the neural network along and have presented some real-time applications in which neural networks can be used and listed out the advantages of using an ANN.

Table 1 and Table 2 in this paper presents the characteristics of machine learning supervised algorithms and mention certain benefits and disadvantages of using them. This has been done in order to help the reader to understand the basics of all supervised algorithms. It will further help the users to understand which algorithm will be suitable for a particular problem. Further, advantages and disadvantages will help the reader in selecting an algorithm to solve a problem.

Table 2. Regression Algorithms in Supervised Learning

Supervised Algorithm (Regression)			
Algorithm	**Working**	**Advantages**	**Disadvantages**
Simple Linear Regression (R)	It is a model that assumes a linear relationship between independent input variables and a single output variable[35].	• Extremely simple method. • It can be used to find the nature of the relationship between two variables.	• Its assumption that the relation between the variables is a straight line is not correct sometimes. • Not good for a large number of parameters.
Multivariate Regression	Multivariate Regression is a method to measure the relationship between two or more response variables. In each situation or decision where there is involvement of more than a single factor, it attempts to model the reality [29].	• Gives a relationship between variables in an overarching. • To determine the link between independent and dependent variable this Introduces other variables.	• Complex. • Involves high level of mathematics. • Outputs are not easy to interpret.
Multiple Linear Regression	Multiple Regression is used to find the relationship between several independent variables and one dependent variable [28].	• Can predict the influence of one or more predictor variables on the creation value. • Can identify the outliers.	• Co-linearity and multiple co-linearity occur in multiple regression.

Table 3. Algorithms for Unsupervised Learning

Unsupervised Machine Learning Algorithms			
Algorithm	**Working**	**Advantage**	**Disadvantage**
K-means clustering	• K-means is a popular clustering method in data mining. • It aims to partition n observation into k cluster [30].	• Faster than hierarchical, if variables are huge. • Produce tighter clusters.	• Difficult to predict k-value. • With the global cluster, it doesn't work well.
Hierarchical clustering	Hierarchical clustering is a method of cluster analysis. It seeks to build a hierarchy of clusters [33].	• From all four clustering algorithms. The hierarchical method is easiest to understand.	• It rarely provides the best solution. • It does not work with missing data.

Regression problems can be solved using simple, multiple and multivariate linear regression models. In the above table i.e. Table 2, presents the characteristics of machine learning supervised algorithm in regression problems and mention certain benefits and disadvantages of using them.

Regression can be used to find the co-relationship between variables. In univariate regression, there is only one dependant and one independant variable, while multivariate regression has more than one dependant and independant variable. If one independent and more than one dependant variables are present it is called multiple regression. Analysis of multilinear regression on a faculty-student dataset of 240 undergraduate students on parameters like teaching methodology, guidance, educational psychology, curriculum developments has been selected. A scatter diagram has been presented by the authors to check the Z scores of tall the variables [28]. KPSS and ANOVA statistics have been used to analyse statistics to predict independant and dependant variables.

Multivariate linear regression has been applied to the understanding of a child's psychology using a Connecticut child study [29]. The authors have discussed disadvantage of using regression analysis for an individual label and pooling strategies. The authors have used multivariate linear regression for the variables having continuous outcomes. They have concluded at the end that the continuous data carry more information which helps in the precision of the data.

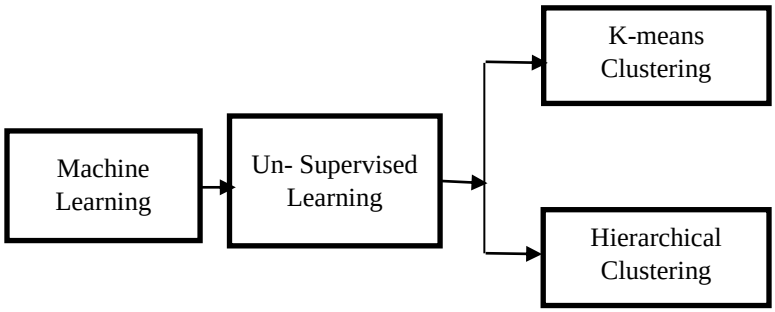

Fig. 2. Unsupervised Machine Learning Algorithm

B. Un-Supervised/ Task Driven Machine Learning Algorithms

The unsupervised machine learning uses unlabeled data for the training of the unsupervised algorithms. A classifier is used to learn patterns in the data to label and classify it.

The following are the advantages of unsupervised algorithms over supervised machine learning algorithms.

a. It understands the data to identify structures or patterns.
b. Evaluation is done indirectly or qualitatively.
c. It does not predict or find anything.

The authors have presented an incremental approach for clustering via a global k-means algorithm [30]. They have also proposed a modification over the current methods to reduce the computational load without affecting the quality of the solution. They have conducted experiments using iris, synthetic and image segmentation datasets by considering only the feature vectors and concluded that the global k-means clustering algorithm performs better than k-means algorithms. Fig.2 presents the unsupervised machine learning algorithms.

Hierarchal clustering is a concept of clustering analysis that deals with the creation of hierarchy clusters. The authors [33] have presented an overview of clustering algorithms like CURE (clustering using representatives), BIRCH (balanced iterative reducing and clustering using hierarchies), ROCK (robust clustering using links), Chameleon, Linkage, leaders-subleaders, and bisecting k-means. They have discussed on how to improve the quality of clustering in hierarchal methods by integrating them with multiple phase clustering.

Recent Work in Applications of Supervised Machine Learning

The advancements of machine learning can be seen in its important real-life applications. This section gives detailed applications of machine learning in day to day life as well as in various scientific fields. The following Table 4 classifies the applications based on supervised algorithms. It gives a brief description of the applications along with the algorithms that give prominent results in solving that problem.

Table 4. Applications of Supervised Machine Learning Algorithms

Applications of Supervised Machine Learning Algorithms		
Application	**Characteristics**	**Algorithms**
Email Spam Filtering	• It uses supervised machine learning to filter junk, spam, unsolicited commercial email (UCE) from genuine e-mails. • It prevents the user from opening spam emails [34].	Bayes' theorem and some heuristics.
Handwriting recognition	• It is an application of AI to recognize characters from handwritten sentences. • It is highly used in segregating envelopes in post offices thereby making distribution of mails easy [7].	Genetic algorithm, Artificial Neural Networks (ANNs).
Face Recognition	• To recognize the facial features as the identification of human beings. • The idea is exploited as a security feature over places like ATMs, areas under surveillance, to unlock mobile phones, and so on [100].	Deep convolutional neural network.

Speech recognition	• To recognize the words in human speech into characters and designing text or speech models. • A Speech recognition system trained as a microphone independent system trained by the producer of that system or on the users' voice model [6].	Hidden Markov Models, Neural Network (recurrent neural network).
Information retrieval	• This application is about finding information from a large collection of the database based on requirements given by the user. • The information could be extracted as a keyword, a list of keywords or a document [3].	Document Ranking Optimization (DROPT).
Computer vision	• This application involves learning via vision. • It recognises images, things, patterns, handwriting, etc. [4].	Convolutional networks.
Text filtering	• The classification of incoming documents dispatched by an information producer in an asynchronous way to an information consumer [36].	Vector space model, Naïve-Bayes.
Operation systems	• ML is being used by operating systems to enhance their performance by learning the app usage behaviour of the user. • The system loads the frequently used application in local memory which results in the speedy start-up of the application [37] [38].	KNN (K-Nearest Neighbor).
Natural language processing (NLP)	• It implies the processing of natural human language. • It can be in any form, either written or spoken. • It is the latest field in research where work is done in processing and finding patterns [39].	Support Vector Machine (SVM), Random Forest, Bayesian Network, Maximum Entropy, Deep learning.
Intrusion detection	• The detection of the events occurring in the system for the possible violations or threats to the security policies. • The machine learning algorithm is being used for an analysis process to detect attacks and is quite a challenging field [40] [41].	Support Vector Machine, Naïve Bayesian, KNN.
Anomaly detection or recognizing anomalies	The identification of an unusual sequence of events. Because these anomalous events have the potential of getting translated into some errors or frauds. For example, fraud sim credit card transactions, error in sensor reading in a nuclear power plant [42].	KNN.
Signature-based detection	• It is the detection of predetermined attack patterns. • These patterns form a signature and are used to determine further network attacks. • ML is used in this area to examine the network traffic with a predefined signature and each time the signatures are updated. • SNORT is an example of Signature-based detection [43].	Support vector machine.
Epileptic Seizure Detection	• Epilepsy is a CNS disorder, in which the patient can result in a lapse of attention or a whole-body convulsion. • ML can be used to construct detectors capable of detecting Epileptic Seizure onset with high accuracy [44].	Support vector machine.
Automated Text Categorization	• It is the process of categorizing the text document into categories. • ML automates the task of categorization using different algorithms.	Deep learning.

	• Text categorization is based on Document organization, Text filtering, word sense disambiguation [45] [46] [47].	
Data Center Optimization	• A data center is a complex interaction of multiple mechanical, electrical control systems. • The most complex challenge faced in data center optimization is power management. • The objective of this problem is to optimize the power and performance of the data center and can be obtained using ML algorithms [48].	Neural networks.
Cognitive Radio	• Dynamic programming options are available in cognitive radio. • It detects the best wireless channel available in the area and uses it to provide wireless communication [49] [50] [51] [52].	Dimensionality reduction, Support Vector Machine (SVM).
Classification of Software Engineering Artifacts Using Machine Learning	• A network can be developed using machine learning for training itself for the task of classification, which uses the defining properties of existing artifacts and then carries on with the task of classification of the artifacts by itself [53].	Support Vector Machine (SVM).
Finance Computation	• The finance market is most unstable and unpredictable. • The prediction of future stock prices can be predicted using machine learning. The agent is given with a buy/sell signal if a pattern that has been seen before is recognized [54] [55] [56] [57].	Random Forest.
Semantic Scene Classification	• This system scans a picture and categorises the image to a specific class. • In a situation where a scene can be categorized into more than one classes, multi-label machine learning provides efficient results [58].	Support Vector Machine (SVM).
Music Information Retrieval	• ML algorithms can be used in music classification, transcription, instrument classification, beat detection, etc.[59].	Deep learning, Support Vector Machine (SVM).
Brain-Computer Interfaces (BCI)	• The collaboration between the brain and a device that reads an electrical signal from the brain is called Brain-Computer Interface. • The brain acts as a controller and the device acts as an interface for the external object to be controlled by the brain [60].	Artificial Neural Networks (ANNs).
Acoustic Environment Identification (AEI) and Audio Forensics	• Possible distortions like background noise, acoustic reverberation, etc. can help investigate important clues. • It creates the disturbances for us but by using various techniques like Acoustic Environment Identification (AEI), Audio Forensics this can put to an advantage. Because that background sound can investigate many important clues [61].	Artificial Neural Networks (ANNs).
Document Classification	• It is the task of categorizing the documents into one or more classes or under different labels [62].	Back-Propagation and Modified Back Propagation (BPNN) and (MBPNN).
Patent Document Classification	• The increase in the number of patent applications requires a need to classify these documents in different classes [63].	Artificial Neural Networks (ANN).

Recent Work in Applications of Un-Supervised Machine Learning

The problems in which the labels have to be determined from the textual data are solved using unsupervised algorithms. Applications like DNA classification, market segmentation, astronomy, cancer detection, etc., where the data changes continuously and new patterns occur frequently are solved using unsupervised algorithms. This set of continuously changing data needs to be learned by a machine to extract patterns. Although these problems can be solved using supervised algorithms, unsupervised algorithms help in bringing accuracy to the pattern recognition task by bringing the error value to a minimum. Clustering Algorithms, Bayesian Networks, Gaussian Mixture Model, Deep Neural Networks, Convolution Neural Networks, Recurrent Neural Networks are some of the unsupervised algorithms used for the process of classification and regression. . The following Table 5 classifies the applications based on un- supervised algorithms. It gives a brief description of the applications along with the algorithms that give prominent results in solving such problems.

Conclusion

In the current paper, we have explored various areas in which machine learning is providing better solutions. We also find that text classification is one such task in the area of natural language processing has a great scope in different domains. Through this survey, we have found that in recent years machine learning algorithms have got the attention of the research community strongly but still there is lot of scope for applying machine learning techniques to obtain better and better solutions.

Furthermore, this paper presents an empirical study of machine learning algorithms in multiple research domains. The aim of carrying out this analysis is to foster discussion among beginners that will help them to kickstart their learning towards problem-solving in machine learning. In our next work, we plan to study deep neural networks for classification techniques. We will be using other standard datasets as well to better judge the performance of the proposed algorithms in the field of natural language processing.

Table 5. Applications of Unsupervised Machine Learning Algorithms

Applications of Unsupervised Machine Learning Algorithms		
Application	**Working**	**Algorithms**
DNA classification	• To find patterns in DNA sequences for the identification of rare diseases, ancestry, etc. • Different DNA structures show different patterns of development, reproduction, and functioning of a human being • The aim is to categorize them such that each of them has a certain gene [64].	Clustering algorithm

Market segmentation	• To discover some patterns based on different characteristics of the customer and making various strategies for the customers. • These strategies can be based upon various traits of the customers like locations, interests, buying and selling patterns and so on. • The customers are divided into segments according to their traits to provide different solutions [65].	Segmentation algorithm
Astronomy	• A large amount of astronomical data is available to predict the theories on the formation and destruction of galaxies. • It is a research area in unsupervised learning where some patters have to be created and analyzed [66].	Clustering techniques
Cancer diagnosis	• It is the field of diagnosis of different types of cancer. • Machine learning is applied for the analysis of the datasets and classifying the types of cancer [67] [68] [69] [70].	Bayesian networks, neural trees, and radial basis function networks
Speech activity detection (SAD)	• It is the process of identifying the segments of speech in the audio signal. • It is important for speeding up manual transcription and reduces error rates for speech recognition [71].	Support vector machines (SVM), gaussian mixture models (GMMs), multi-layer perceptron (MLP)
Computational biology	• Also known as bioinformatics. • It is the application of biological data to develop algorithms and establish relations among various biological systems. • Machine Learning offers some of the most cost-effective tools for building predictive models with the use of biological data [8] [72] [73] [74] [75].	Learning from highly unbalanced datasets
Organizing large computer clusters	• Analysis of the tendency of machines to work together is done by creating clusters using machine patterns. • The clusters are optimized to find nodes that are compatible with each other. • It has helped in boosting the efficiency of data centers [76].	K-means, Random forest
Social network analysis	• The increase in storage capabilities and availability of the internet has led to an increase in social networking platforms. • The analysis of this content from a different perspective in different fields have become a focus of ML and natural language processing society [77].	Classification algorithms like clustering
The cocktail party problem	• A famous cocktail party problem of AI can be solved using unsupervised learning [78].	Cocktail party algorithm
Medical records	• EHR or electronic health records are records converted digitally from handwritten medical records. • EHR's helps in providing medical knowledge to understand the disease in a better way. It helps in making the diagnosis easy [79] [80].	Random forest
Acoustic factor analysis for robust speaker verification	• Background noise, handset variability, transmission channel, etc. are the mismatches that take place during verification in speech recognition. • To overcome the problem, acoustic factors are analyzed to suppress unwanted channel components [81].	Deep neural network

References

Richert W. Building machine learning systems with Python. Packt Publishing Ltd., 2013.

Welling M. A first encounter with Machine Learning. Irvine, CA. University of California 2011.

Cunningham SJ, Littin J, Witten IH. Applications of machine learning in information retrieval 1997.

LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In Proceedings of IEEE international symposium on circuits and systems 2010: 253-256.

Prasad JR. Pattern recognition: possible research areas and issues. International Journal of Computer Science and Network 2014; 3(5).

Mitchell TM. The discipline of machine learning. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department 2006.

Ball GR, Srihari SN. Semi-supervised learning for handwriting recognition. In 10th International Conference on Document Analysis and Recognition 2009; 26-30.

Caragea C, Vasant H. Machine Learning in Computational Biology. Encyclopedia of Database Systems 2009: 1663-1667.

Data management solution review. 2019. Retrieved from:https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

IDC-Seagate. 2018. The reference can be found at https://www.aparavi.com/data-growth-statistics-blow-your-mind/

Ben-Hur A, David H, Hava TS, Vladimir V. Support vector clustering. Journal of machine learning research 2001; 2: 125-137.

Schapire RE. The boosting approach to machine learning: An overview. In Nonlinear estimation and classification 2003; 149-171.

Jerez JM, Ignacio M, Pedro JGL, Emilio A, Nuria R, Miguel M, Leonardo F. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine 2010; 50(2): 105-115.

Otukei JR, Blaschke T. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. International Journal of Applied Earth Observation and Geoinformation 2010; 12: S27-S31.

Chen D, Qiang H, Xizhao W. FRSVMs: Fuzzy rough set based support vector machines. Fuzzy Sets and Systems 2010; 161(4): 596-607.

Mandal I, Sairam N. Enhanced classification performance using computational intelligence. Communications in Computer and Information Science 2011: 384-391.

Mandal I, Sairam N. Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system. International Journal of Medical Informatics 2013; 82(5), 359-377.

Škrinárová J, Huraj L, Siládi V. A neural tree model for classification of computing grid resources using pso tasks scheduling. Neural Network World 2013; 23(3): 223-241.

Min F, Qinghua H, William Z. Feature selection with test cost constraint. International Journal of Approximate Reasoning 2014; 55(1): 167-179.

Shu K, Zhou X, Wang S, Zafarani R, Liu H. The role of user profiles for fake news detection. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019; 436-439.

Pranckevičius T, Virginijus M. Application of logistic regression with part-of-the-speech tagging for multi-class text classification. In IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE) 2016: 1-5.

Han H, Wei X, Hongyuan Z, Lee Giles C. A hierarchical naive Bayes mixture model for name disambiguation in author citations. In Proceedings of the ACM symposium on Applied computing 2005; 1065-1069.

Bijalwan V, Kumar V, Kumari P, Pascual J. KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application 2014; 7(1): 61-70.

Guo G, Hui W, David B, Yaxin B, Kieran G. KNN model-based approach in classification. In OTM Confederated International Conferences On the Move to Meaningful Internet Systems, Springer, Berlin, Heidelberg 2003: 986-996.

Patel BN, Satish GP, Kamaljit IL. Efficient classification of data using a decision tree. Bonfring International Journal of Data Mining 2012; 2(1): 6-12.

Hasan MAM, Nasser M, Pal B, Ahmad S. Support vector machine and random forest modeling for intrusion detection system (IDS). Journal of Intelligent Learning Systems and Applications 2014; 6(1).

Caesarendra W, Achmad W, Bo-Suk Y. Application of relevance vector machine and logistic regression for machine degradation assessment. Mechanical Systems and Signal Processing 2010; 24(4): 1161-1171.

Uyanık GK, Neşe G. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences 2013; 106: 234-240.

Goldwasser MA, Garrett MF. Multivariate linear regression analysis of childhood psychopathology using multiple informant data. International Journal of Methods in Psychiatric Research 2001; 10(1): 1-10.

Likas A, Nikos V, Jakob JV. The global k-means clustering algorithm. Pattern recognition 2003; 36(2): 451-461.

Maind SB, Priyanka W. Research paper on basic of artificial neural network. International Journal on Recent and Innovation Trends in Computing and Communication 2014; 2(1): 96-100.

Petri M, Henry T. Bayesian Case-Based Reasoning with Neural Network. In Proceeding of the IEEE International Conference on Neural Network'93 1993; 1: 422-427.

Rani Y, Harish R. A study of hierarchical clustering algorithm. TERS & on Te SIT-2 2013.

Tzanis G, Katakis I, Partalas I, Vlahavas I. Modern applications of machine learning. In Proceedings of the 1st Annual SEERC Doctoral Student Conference–DSC 2006; 1(1): 1-10.

Asai HTSUK, Tanaka S, Uegima K. Linear regression analysis with fuzzy model. IEEE Trans. Systems Man Cybern 1982; 12: 903-907.

Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR) 2002; 34(1): 1-47.

Horvitz E. Machine learning, reasoning, and intelligence in daily life: Directions and challenges. Proceedings 2006.

Negi A, Kumar PK. Applying machine learning techniques to improve linux process scheduling. In TENCON 2005-2005 IEEE Region 10 Conference 2005; 1-6.

Collobert R, Jason W, Léon B, Michael K, Koray K, Pavel K. Natural language processing (almost) from scratch. Journal of machine learning research 2011; 12: 2493-2537.

Kaur H, Singh G, Minhas J. A review of machine learning based anomaly detection techniques. International Journal of Computer Applications Technology and Research 2013; 2(2): 185 – 187.

Othman SM, Ba-Alwi FM, Alsohybe NT, Al-Hashida AY. Intrusion detection model using machine learning algorithm on Big Data environment. Journal of Big Data 2018; 5(1).

Wiese B, Omlin C. Credit card transactions, fraud detection, and machine learning: Modelling time with LSTM recurrent neural networks. In Innovations in neural information paradigms and applications, Springer, Berlin, Heidelberg 2009; 231-268.

Kumar V, Sangwan OP. Signature based intrusion detection system using SNORT. International Journal of Computer Applications & Information Technology 2012; 1(3): 35-41.

Shoeb AH, Guttag JV. Application of machine learning to epileptic seizure detection. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) 2010; 975-982.

Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR) 2002; 34(1): 1-47.

Bratko A. Spam filtering using statistical data compression models. The Journal of Machine Learning Research 2006; 7: 2673-2698.

Guyon I, André E. An introduction to variable and feature selection. The Journal of Machine Learning Research 2003; 3: 1157-1182.

Gao J, Ratnesh J. Machine Learning Applications for Data Center Optimization. Google White Paper 2014.

Hou S, Qiu RC, Chen Z, Hu Z. SVM and dimensionality reduction in cognitive radio with experimental validation 2011.

Tsagkaris K, Apostolos K, Panagiotis D. Neural network-based learning schemes for cognitive radio systems. Computer Communications 2008; 31(14): 3394-3404.

Tabaković Ž. A survey of cognitive radio systems. Post and Electronic Communications Agency, Jurišićeva 2011.

Hosey N, Bergin S, Macaluso I, O'Donoghue D. Q-learning for cognitive radios. In Proceedings of the China-Ireland Information and Communications Technology Conference (CIICT 2009). National University of Ireland Maynooth 2009.

Bruegge B, David J, Helming J, Koegel M. Classification of Software Engineering Artifacts Using Machine Learning.

Boyarshinov V. Machine learning in computational finance. Diss. Rensselaer Polytechnic Institute 2005.

Pawar P. Machine Learning applications in financial markets. Diss. Indian Institute of Technology, Bombay Mumbai.

Stephan C. A Machine-Learning View of Quantitative Finance¸ con - Institut Mines Telecom LTCI UMR Telecom Paris Tech.

Shen S, Haomiao J, Tongda Z. Stock market forecasting using machine learning algorithms 2012.

Shen X. Multilabel machine learning and its application to semantic scene classification. Electronic Imaging 2004. International Society for Optics and Photonics 2003.

Øland A. Machine Learning and its Applications to Music. Machine Learning report. e IT University of Copenhagen (ITU) 2011.

Makeig S, Kothe C, Mullen T, Shamlo NB, Zhang Z, Kreutz-Delgado K. Evolving Signal Processing for Brain-Computer Interfaces. Proceedings of the IEEE 100 (Centennial-Issue) 2012; 1567- 1584.

Hafiz M. Acoustic Environment Identification and Its Applications to Audio Forensics. IEEE Transactions on Information Forensics and Security 2013; 8(11): 1827-1837.

Wang TY, Chiang HM. One-against-one fuzzy support vector machine classifier: An approach to text categorization. Expert Systems with Applications 2009; 36(6): 10030-10034.

Trappey AJC, Hsu FC, Trappey CV, Lin CI. Development of a patent document classification and search platform using a back-propagation network. Expert Systems with Applications 2006; 755–765.

Cho, SB, Hong HW. Machine learning in DNA microarray analysis for cancer classification. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19. Australian Computer Society, Inc., 2003.

Haider P, Luca C, Ulf B. Discriminative clustering for market segmentation. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012.

Xiong L, Poczos B, Connolly A, Schneider J. Anomaly detection for astronomical data. Data Analysis Project, Machine Learning Department, Carnegie Mellon University 2010.

Hwang KB. Applying machine learning techniques to the analysis of gene expression data: cancer diagnosis. Methods of Microarray Data Analysis. Springer US 2002: 167-182.

Luca S. Machine Learning in Biology. Universitadeglistudi Di Padova.

Zararsiz G, Elmali F, Ozturk A. Bagging support vector machines for leukemia classification. International Journal of Computer Science Issues (IJCSI) 2012; 9(6): 355-358

Wang Y. Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry 2005; 29(1): 37- 46.

Sadjadi SO, Hansen JH. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Processing Letters 2013; 20(3): 197-200.

Guyon I, André E. An introduction to variable and feature selection. The Journal of Machine Learning Research 2003; 3: 1157-1182.

Hou S, Qiu RC, Chen Z, Hu Z. SVM and dimensionality reduction in cognitive radio with experimental validation 2011.

Tarca Adi L. Machine learning and its applications to biology. PLoS computational biology 2007; 3(6).

Sajda P. Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng., 2006; 8: 537-565.

Liao SW, Hung TH, Nguyen D, Chou C, Tu C, Zhou H. Machine learning-based prefetch optimization for data center applications. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis 2009; 1-10.

Haider P, Chiarandini L, Brefeld U. Discriminative clustering for market segmentation. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012; 417-425.

Haykin S, Zhe C. The cocktail party problem. Neural computation 2005; 17(9): 1875-1902.

Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine 2001; 23(1): 89-109.

Magoulas GD, Andriana P. Machine learning in medical applications. Machine Learning and its applications. Springer Berlin Heidelberg 2001; 300-307.

Hasan T, Hansen JH. Acoustic factor analysis for robust speaker verification. IEEE Transactions on audio, speech, and language processing 2012; 21(4): 842-853.

Aruna S, Abinaya P. A Web service recommendation system based on QoS attribute based collaborative filtering. Webology 2020; 17(1): 246-254.

Gholampour S, Noruzi A, Gholampour B, Elahi A. Research trends and bibliometric analysis of a journal: Sport management review. Webology 2019; 16(2): 223-241.