ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

A Survey on Machine Learning: Algorithms, Performance and Applications

Sushanto Praharaj [1], Shivam Bhat[1], Ganashree K.C [2], Dr. Krishnappa H.K [2]

B.E. Student, Dept. of Computer Science and Engineering, R.V College of Engineering, Bangalore, India[1]

B.E. Student, Dept. of Computer Science and Engineering, R.V College of Engineering, Bangalore, India[1]

Assistant Professor, Dept. of Computer Science and Engineering, R.V College of Engineering, Bangalore, India[2]

Associate Professor, Dept. of Computer Science and Engineering, R.V College of Engineering, Bangalore, India[2]

ABSTRACT: Machine Learning and Artificial Intelligence are the most popular buzzwords today in the technological industry. Developing practical and accurate probabilistic Machine Learning models in today's computational environment is possible, largely due to the strides made in the field of High Performance Computing and massively parallelized hardware architectures. This paper conducts a survey on the most relevant Machine Learning Algorithms with respect to their applications in industry, discusses how they work, the performance of these algorithms across various hardware platforms, and finally, discusses their real-world applications.

KEYWORDS: Machine Learning, Artificial Intelligence, Algorithms, Training time, Validation Time

I. INTRODUCTION

Machine Learning has already found applications in several domains of Computer Science research. As the name suggests, the intuitive idea is to train a machine to learn how to implement a computational task and execute this computation with sufficient accuracy. Less intuitively, it usually involves the development of mathematical models with learnable parameters that are honed based on the numerical values of the training data, in order to predict probabilistic outcomes. Most generic machine learning algorithms used in the industry fall under supervised and unsupervised algorithms. Supervised algorithms include labelled data and are easier to train, when compared to unsupervised machine learning algorithms. The implementation of these algorithms has become feasible over the course of the last decade owing to the major strides made in computation power and development of GPUs with massive multithreading capabilities for high performance computing. As a result, machine learning algorithms have quickly found applications in niche areas where it is useful to have probabilistic mathematical models, such as classification and regression.

Apart from the groundbreaking results machine learning algorithms have produced in the field of image classification and object detection and in the field of Natural Language Processing, statistical machine learning techniques have found utility in Domains pertaining to Data Science, Visualization and Analytics. This involves the development of robust mathematical models to represent and map the data present. Once a suitable mapping function is chosen, predictions for ensuing categorical data can be carried out with reasonable accuracy.

A. *BACKGROUND : DEVELOPMENT OF MACHINE LEARNING ALGORITHMS*

The idea of sentient computers has been around since long before the hardware that could realise it was developed. ELIZA**[1]** was among the earliest chatbots built in the 1960s. While the algorithm supporting didn't support learning, it was coded to produce responses that would allow a conversation to keep going for most input sentences. Further fields where the concept of Artificial Intelligence saw successes, was in 1997 when IBM's 'Deep Blue' beat the world chess champion.

2006 was a landmark year for researchers in Machine Learning, with the emergent concept of Deep Learning[2]. The term was coined to denote Neural Network models that included multiple hidden layers. While these layers come at the cost of greater training time, the accuracy and results they produced were found to be reasonably better.

B.. *ENHANCEMENTS BROUGHT ABOUT BY DEEP LEARNING*

The emergence of deep learning and the rise of Neural Networks was heavily reliant on Biomimicry. The concept of a Deep Neural Network is based on the Central Nervous System in Mammals, wherein a neuron receives an input and passes the information on after a minor amount of processing to the next neuron. These models were built using cognitive science principles and the results have been outstanding. As a result, Artificial Neural Networks have been used for Image Processing and Language modelling and were instrumental in the development of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Artificial Neural Networks, however, come at the cost of increased computation time and learning parameters. The more intricate the neural network, the more learnable parameters there will be. As a result, the realisation of artificial neural networks on hardware was heavily reliant on the improvements made in GPU architecture for High Performance Computing.

C. *OTHER STATISTICAL TECHNIQUES*

While Deep Learning has produced outstanding results among learnable Mathematical Models, the massive number of learnable parameters in a typical neural network makes it difficult to train and as a result, Deep Learning models are not always scalable. There are other Statistical Modelling techniques that can be used to produce a reasonable fit to the data and allow predictions. Chief among these are Regression Techniques: Linear, Logistic, Polynomial, etc., Decision Tree models, Bayesian Classifiers, Support Vector Machines (SVMs), and Principal Component Analysis (PCAs). It is often necessary to change the dimensionality of the data in order to develop a mathematical model that produces a better fit. This is precisely what SVMs and PCA do.

II. PRESENT LEARNING METHODS AND RELATED WORK

A. *SUPERVISED LEARNING*

As mentioned previously, Supervised Learning models deal with learning techniques wherein labelled data is prevalent. That is, the model being developed has information of what it has to label test data as, based on the parameters it has learnt during the training process. The model learns a mapping, i.e, a relationship between the feature vectors of the input data and associates it with the labelling provided for that input data. It then tunes its learnable parameters according to this relationship in order to be able to produce similar labelling outputs for input data with similar feature vectors.

V.N Vapnik's New York's Springer Edition[3], suggests that in order to develop a machine learning model to solve a given supervised learning problem, the following steps need to be followed.

1. Determining the type of training examples
2. Gathering and arranging the training data set
3. Deterimining the input feature representation of learning functions
4. Understanding the learning function and learning algorithm and evaluating its appropriateness in mapping the data.

5. Proceeding with designing and and training the developed model on the training data-set
6. Evaluating the accuracy of the learning function and based on predetermined accuracy metrics. This accuracy needs to measured on the training data-set as well as a different data-set to check for over-fitting. (Validation Data-Set).

Generally, supervised learning models are developed to solve problems where training data is plentiful. As is, supervised learning models generally find application in solving the following types of problems:

1. Classification: These problems use supervised learning based machine learning models to classify input data samples into preset categories. These models are trained to identify specific features in the input data sample that would uniquely classify it into a particular category. Based on several similar inputs the model sees during its training period, it learns to classify such input samples into their corresponding categories. Popular examples of classification are using trained Convolutional Neural Networks to classify images into a preset class. Thus the output variable for a classification problem is categorical.

2. Regression: These problems use supervised learning to predict continuous quantity outputs. These predictions are made based on trends in the input data that the model learns. The major difference between Regression and classification problems is that classification produces categorical output variables, while regression produces a continuous quantitative output. Regression techniques have several methods, chief among which are Linear Regression, Logistic Regression, Polynomial regression,etc. The principal amongst these are Linear Regression which produces a Linear mapping in the data, and Logistic Regression.

B. *UNSUPERVISED LEARNING*

Unsupervised learning models are trained without labelled data-sets. This indubitably makes such problems more difficult to solve as there are no clear mapped examples of input training data to corresponding outputs. Thus the model cannot learn these relationships and practise them on other input data. The goal of this technique is therefore to develop an algorithm to train the model in drawing inferences from the training dataset without labelled responses.

The most common method used in training unsupervised machine learning models is Cluster Analysis - which looks for hidden patterns or groups in the data-set provided. The clusters are designated using a set of metrics such as Euclidean Distance, Mahalanobis Distance, Manhattan Distance, etc. Data items with similar metric values are usually clustered together, thereby allowing the model to learn to map features to clusters.

Types of Clustering Algorithms include:

1. Hierarchical Clustering
2. K-Means Clustering
3. Hidden Markov Models
4. Gaussian Mixture Models
5. Self Organizing-maps.

Another approach to developing training algorithms for unsupervised machine learning models is by incorporating a reward system to measure success and failure. I. Wittenet. et. al draws an analogy of this approach to real world examples wherein agents are rewarded for their actions which are favourable and punished for their actions that cause unfavorable outcomes.

Unsupervised learning techniques are also used when it is desirable to deal with non-linear data with millions of parameters. Deep Belief Networks (DBF) and sparse coding are two more popular unsupervised techniques for training on unlabelled data-sets. Rajat Raina et. al's paper on Large-scale Deep Unsupervised Learning using Graphics

Processors[4] suggests that unsupervised algorithms have a massive potential for parallelization using Graphics Processors. The paper suggests that the learning process of the suggested DBF networks was as much as 70 times faster for large models

III. DISCUSSED ALGORITHMS

A. REGRESSION

Regression models are often used in predictive analytics to build appropriate mappings between the data points and obtain trends and patterns in the data. It is used to model and analyse variables and determine relationships between them as to whether they are dependent or independent variables. Popular Regression models include

1. Linear Regression: This involves finding a best fit line for the input data. Error data for each training sample is usually calculated using the OLS (Ordinary Least Squares) method. The result is a line with minimal errors (often distance) for the training data points. The case of a single explanatory variable (independent variable) is referred to as Simple Linear regression, while linear regression involving multiple independent variables is called Multiple Regression.

2. Logistic Regression: Logistic Regression is used to map dichotomous (binary) dependent variables. There exist extensions of this model however, allowing the regression to map to non-binary variables as well (Multiple Logistic Regression).

B. DECISION TREE-BASED ALGORITHMS[5]

Decision Trees are supervised Machine Learning models and can be used to solve both regression as well as classification problems. Decision Trees start off with a single decision (the root) after which it forks into further decisions. The leaf nodes represent the labels of the classes, and the internal nodes are used to represent the attributes and decisions.

Popular Decision Tree algorithms include:

1. The ID3 Algorithm: The Iterative Dichotomiser 3 is a decision tree that was implemented under the premise of Hunt’s algorithm. An iterative top-down approach is followed in this algorithm, wherein we begin at the root node and continue based on a greedy approach. On each iteration of the algorithm, the attributes’ entropies (H(S)) are calculated and correspondingly, so are their Information Gains (IG(S)). The set is partitioned into smaller subsets and the algorithm continues recursively through these, until a label is reached.

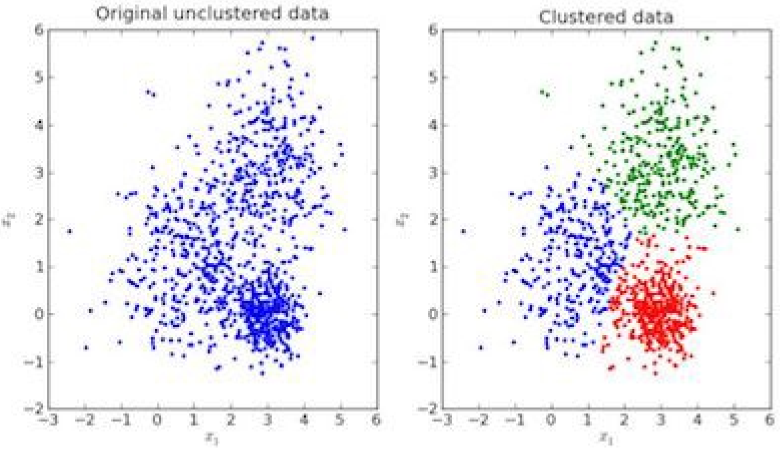
2. The C4.5 Algorithm: The C4.5 algorithm is an extension to the ID3 algorithm, and is used almost exclusively for classification problems. The C4.5 uses a more efficient way of splitting the attributes into its subsets as mentioned previously in the ID3 algorithm. C4.5 attempts to go back through the tree once it’s been created and attempts to prune the ineffective branches as well.

C. SUPPORT VECTOR MACHINE[6]

Support Vector Machines are majorly used in classification problems wherein the segregation of data into their classes doesn’t have an apparent relationship. The algorithm changes dimensionality accordingly and finds an appropriate hyper-plane to clearly segregate the data into their corresponding classes. SVMs have proved to be robust against outliers.

D. CLUSTERING ALGORITHMS[7]

Clustering algorithms are unsupervised algorithms that look for patterns and similarities amongst training data. The idea is to find relationships between the data and group data with similar characteristics. The K-Means clustering algorithm is one such algorithm used to group input data into K clusters using appropriate distance metrics. It involves a set of random centroids for each cluster and iteratively evaluates distances of each data point to these centroids. Thus based on the shortest distance metric (highest similarity) the data points are grouped into appropriate clusters.



E. DEEP LEARNING ALGORITHMS

Deep Learning Algorithms are difficult to train from a computational perspective because of how intricate most deep networks are and the millions of learnable parameters they possess. However, once models are trained with reasonable accuracy, they are capable of producing outstanding results and learning relationships in data.

Deep Learning is modelled based on the Central Nervous System, wherein a node (neuron) gets inputs from the previous layers, performs its own computations, and passes the computed output to further layers which finally generate a cumulative output based on the operations carried out by each hidden layer. Such ANNs (Artificial Neural Networks are often referred to as Multi Layer Perceptrons (MLPs) where in a neuron, is called a Perceptron.[8][10][11]

Deep learning has find massive use in the emerging fields of Computer Vision and Machine Learning. The more popular types of Neural Networks being

1. Convolutional Neural Networks: CNNs are used to extract features from input images. These usually comprise combined layers of Convolution Filters and Maxpooling Layers used to extract features from the input images. These are then fed into an ANN for classification. CNNs have produced exemplary results in Image Classification Problems.[9]

2. Recurrent Neural Networks: RNNs are used to hold data and reuse it. An improvement on RNNs is the concept of Attention[12]. A notable extension of the RNN was the concept of Long Short Term Memory networks (LSTMs) introduced by Hochreiter and Schmidhuber[13]. The concept of attention gives importance to context in sentences and this is crucial in coherent implementations of Natural Language Processing systems.

IV. APPLICATIONS AND PERFORMANCE ANALYSIS

1) Content Generation in Game:AI is today being used in gaming industry for two main areas: game design budget and upgrade in-game experience.procedural content generation(PCG) is an important task of game development and includes generation of game content like levels,game maps,rules,items,quest,music,vehicle,characters,etc.

In order to reduce labour cost and create content faster,gaming industry is adopting state of art artificial intelligence at a fast pace.AI is well suited for PCG problems as it can handle visual and audio data and learn patterns from large volumes of data.

An increasingly popular method in PCG research is generative adversarial network(GANs) which is a deep neural net architecture comprised of two nets contesting with one another They are highly capable of producing similar style content. Recently researchers at Georgia Institute of Technology were able to invent new games using using artificial intelligence[15]

2) Enhancing Gaming experience with AI:AI is being used to model a human player to understand user interaction with game.AI tries to understand player game play and his experiences during a game.This is achieved using machine learning techniques,such as supervised learning like support vector machines or neural networks to build player experience model[17].By model we refer to a mathematical representation which can consist of a rule set,a vector of parameters or a set of probabilities that captures the underlying function between the characteristics of player and her interaction with the game, and the player's response to that interaction.The training data here consists of some aspect of the game or player-game interaction, and the targets are labels derived from some assessment of player experience, gathered for example from physiological measurements or questionnaires.

3)Hyper-formalist Game Studies:AI methods can be applied to corpora of games in order to understand distributions of game characteristics[19]. For example, decision trees can be used to visualize patterns of resource systems in games. There are likely many other ways of using game AI for game studies that are still to be discovered.

4)Intention-driven user interfaces: In such interfaces,instead of responding to users' literal queries, search will use machine learning to take vague user input, discern precisely what was meant, and take action on the basis of those insights.This will allow products which use machine learning[16] to make user interfaces that can tolerate imprecision, while discerning and acting on the user's true intent.We're already seeing early examples of such intention-driven interfaces: Apple's Siri; Wolfram Alpha; IBM's Watson; systems which can annotate photos and videos; and much more.

5)Language Modelling and Prediction using RNNS: Recurrent Neural Networks are one of the most common Neural Networks used in Natural Language Processing because of its promising results[18]. The applications of RNN in language models consist of two main approaches. We can either make the model predict or guess the sentences for us and correct the error during prediction or we can train the model on particular genre and it can produce text similar to it, which is fascinating.The logic behind a RNN is to consider the sequence of the input. For us to predict the next word in the sentence we need to remember what word appeared in the previous time step.

In this method, the likelihood of a word in a sentence is considered. The probability of the output of a particular time-step is used to sample the words in the next iteration(memory). In Language Modelling, input is usually a sequence of words from the data and output will be a sequence of predicted word by the model.

6)Convolutional Neural Networks for Image Processing: Convolutional neural networks (CNNs) represent an interesting method for adaptive image processing, and form a link between general feed-forward neural networks and adaptive filters. Two dimensional CNNs are formed by one or more layers of two dimensional filters, with possible non-linear activation functions and/or down-sampling. CNNs possess key properties of translation invariance and spatially local connections (receptive fields).The images are passed through a series of convolutional, nonlinear, pooling layers and fully connected layers, and then generates the output.This has made self-driving cars, efficient web search, speech and image recognition a reality. **[9]**

Given below is a performance analysis of various algorithms implemented in various popular Machine Learning Libraries based on Pedregosa et. al's paper 'Scikit-Learn: Machine Learning in Python'. **[14]**

	Scikit-Learn	MLpy	pybrain	pymvpa	mdp	shogun
SVM	5.2	9.47	17.5	11.52	40.48	5.63
LARS	1.57	105.3	-	37.35	-	-
Elastic-Net	0.52	73.7	-	1.44	-	-
K Nearest Neighbours	0.57	1.41	-	0.56	0.58	1.36
PCA	0.18	-	-	8.93	0.47	0.33
K-Means	1.34	0.79	-	-	35.75	0.68

Table1: Time in seconds on the Madelon data set for various machine learning libraries exposed in Python: MLPy (Albanese et al., 2008), PyBrain (Schaul et al., 2010), pymvpa (Hanke et al., 2009), MDP (Zito et al., 2008), Shogun (Sonnenburg et al., 2010), and Scikit-Learn (Pedregosa et al.).

V. CONCLUSION AND INFERENCES

Machine Learning algorithms have indubitably proved to be reliable mathematical models that can solve complex computational tasks efficiently and accurately. Trained models are being used for several tasks that would be extremely difficult were it not for the existence of Machine Learning models. Self Driving Cars employ complex CNNs for object classification. Object Detection has been implemented with the help of CNNs using the YOLO and the RCNN models to provide exemplary results in this field. Given that Machine Learning algorithms provide such elegant solutions to problems that are intuitive to solve but difficult develop algorithms for, it looks to be at the helm of progress in the development of Artificial Intelligence.

Furthermore, given its firm statistical roots, Machine Learning Algorithms have produced impeccable results in the field of predictive analytics and trend surveys and predictions. Algorithmic trading employs several such techniques as does the field of Computational Science. Machine Learning techniques are also being embedded in everyday software use in order to provide users with an optimized experience pertaining to their likes and dislikes in order to render the software customized to serve the user optimally.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

REFERENCES

1. ELIZA—a computer program for the study of natural language communication between man and machine, Joseph Wizenbaum, MIT, Cambridge, 1966
2. A fast learning algorithm for Deep Belief Nets, Geoffrey. E. Hinton, Simon Osindero, Yee Whye Teh, 2006
3. V. Vapnik. Statistical Learning Theory. John Wiley & Sons, 1998
4. Large-scale Deep Unsupervised Learning using Graphics Processors, *NIPS 2008 Workshop on Parallel Implementations of Learning Algorithms*
5. *J Ross Quinlan, Machine Learning*, vol. 1, no. 1, 1975
6. Vikramaditya Jakkula, Tutorial on Support Vector Machine (SVM), Washington State University
7. Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEEAn Efficient k-Means Clustering Algorithm: Analysis and Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence,, VOL. 24, NO. 7, JULY 2002..
8. Hassan Ramchoun, Mohammed Amine Janati Idrissi, Youssef Ghanou, Mohamed Ettaouil - Multilayer Perceptron: Architecture Optimization and Training, International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, Nº1
9. Yann Lecun, Patrick Haffner, Leon Bottou, and Yoshua Bengio, Object recognition with Gradient Based Learning.
10. David E Rumelhart, Geoffrey E. Hinton, Ronal J Williams, Learning Representations by Backpropagating Errors.Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.
11. JA Hertz, Introduction to the theory of Neural Computation, 2018
12. Minh-Thang Luong Hieu Pham Christopher D. Manning Computer Science Department, Stanford University, Stanford, CA 94305, Effective Approaches to Attention-based Neural Machine Translation
13. Sepp Hochreiter, Jurgen Schmidhuber, Long Short Term Memory, Neural Computation 9(8):1735{1780, 1997]
14. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Scikit Learn - Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825-2830 Submitted 3/11; Revised 8/11; Published 10/11
15. Matthew Guzdial and Mark Riedl,'Automated Game Design via Conceptual Expansion',14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment on Nov. 13-17
16. Louridas, P., & Ebert, C. (2016). Machine Learning. IEEE Software, 33(5), 110–115
17. Matthew Guzdial, Boyang Li, and Mark Riedl.,'"Game Engine Learning from Video',International Joint Conference on Artificial Intelligence, Aug. 19-25.
18. Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, Sanjeev Khudanpur,'Recurrent Neural Network Based Language Model',11th Annual Conference of the International Speech Communication Association,Makuhari, Chiba, Japan,September 26-30. 2010
19. Georgios N. Yannakakis, Julian Togelius,'Artificial Intelligence and Games'.