

A literature survey on Machine Learning Algorithms

Rekha Nagar^{1*}, Yudhvir Singh^{2*}
U.I.E.T (M.D.U), India

Abstract:

Machine Learning (ML) has unfolded from the Artificial Intelligence, a field of computer science. Machine Learning (ML) is a multidisciplinary field, a combination of statistics and computer science algorithms which is widely used in predictive analyses and classification. The second section of the paper focuses on influencing the basic machine learning methods and algorithms. This paper will go through the various machine learning tools needed to run the machine learning projects. The main concern of the concerned paper is, the study of the main approaches and case studies of using machine learning for forecasting in different areas such as stock price forecasting, tourism demand forecasting, solar irradiation forecasting, supply chain demand and consideration of neural network in machine learning methods.

Keywords: machine learning, unsupervised learning, support vector machine, supervised learning

1. Introduction

Over the past decades, Artificial Intelligence (AI) stream has become the broad and exciting field in computer science as it prepares the machines to perform the tasks that human beings may do, and it aims to train the computers to solve real-world problems with the maximum success rate. As perceiving scientific growth and advancement in technology, AI systems are now capable to learn and improve through past experiences without explicit assistance code if they are exposed to new data. Eventually, it leads to the technology of Machine Learning (ML) which uses learning algorithms to learn from the data available [1]. Machine Learning uses data mining techniques to extract information from huge size datasets. ML and Data Mining techniques explore data from end to end to find the hidden patterns inside the dataset [2]. Machine Learning and data mining algorithms have been deployed in various fields such as Computer networking, travel and tourism industry, finance forecasting, telecommunication industry and electric load forecasting and so on [2].

2. Preliminaries

2.1 Methods used in Machine Learning

Over the past years, an enormous number of ML algorithms were introduced. Only some of them were able to solve the problem, so they were replaced by another one [3]. There are three ML algorithms, for example, unsupervised learning and reinforcement learning, supervised learning, which are displayed in the following figure 1.

2.1.1 Supervised learning: It consists of a given set of input variables (training data) which are pre-labeled and target data [5]. Using the input variables, it generates a mapping function to map inputs to required outputs. Parameter adjustment procedure continues until the system acquires a suitable accuracy extent regarding the training data.

2.1.2. Unsupervised learning : In this algorithm, we only have training data rather than outcome data. That input data is not previously labeled. It is used in classifiers by recognizing existing patterns or clusters in the input datasets [4].

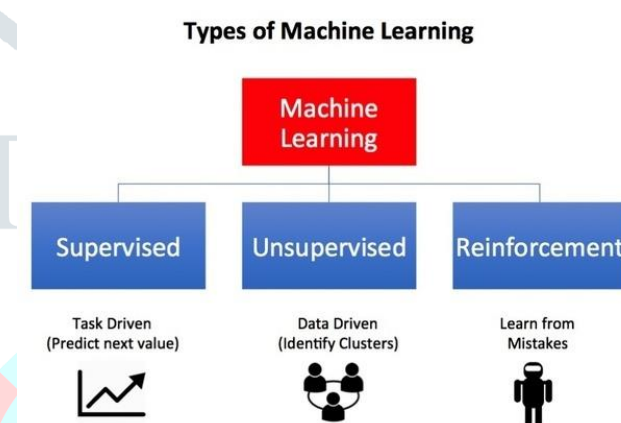


Fig 1

2.1.3 Reinforcement learning : Applying this algorithm, a machine is trained to map action to a specific decision, hence the reward or feedback.

Signals are generated. The machine trained itself to find the most rewarding actions by reward and punishment using past experience.

2.2 Algorithm of Machine Learning

There are a massive number of algorithms used by machine learning, designed to erect models of machine learning and implemented in it [4]. All algorithms can be grouped by their learning methodology, as follows:

2.2.1. Regression algorithms

In Regression algorithms, predictions are made by the model with modeling the relationship between variables using a measure of error [25]. Continuously varying value is predicted by the Regression technique. The variable can be a price, a temperature. The favored regression algorithms are as follows:

- **Linear Regression algorithm**
- **Ordinary Least Squares Regression**
- **Multivariate Adaptive Regression Splines**
- **Logistic Regression**
- **Locally Estimated Scatter plot Smoothing**
- **Stepwise Regression**

2.2.2. Instance based learning algorithms

In the algorithms which are based on Instance, a decision problem is an issue with illustration of training data, build up a database and compare test data, then form a prediction. Instance-based learning

method is famous as lazy learner. The most well known algorithms based on instance learning algorithms are:

- Learning Vector Quantization
- Self-Organizing Map
- k-Nearest Neighbor
- Locally Weighted Learning

2.2.3. Algorithms using Decision Tree

Algorithms using Decision trees are used mainly in classification problem. They splits attributes in two or more groups by sorting them using their values. Each tree have nodes and branches[4]. Attributes of the groups are represented by each node and each value represented by branch [5]. An example of decision tree is given in Fig. 3.

The most well known algorithms using decision tree are:

- Iterative Dichotomized 3
- M5
- Chi squared Automatic Interaction Detection
- C5.0 and C4.5 (different versions of a powerful approach)
- Decision Stump
- Classification and Regression Tree
- Conditional Decision Trees

Decision Tree

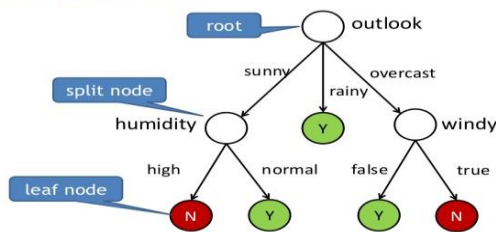


Fig. 2.

2.2.4. Bayesian algorithms

Machine Learning is multidisciplinary field of Computer Science like Statistics and algorithm. Statistics manages and quantifies the uncertainty and are represented by bayesian algorithms based on probability theory and Bayes' Theorem. The most famous Bayesian algorithms are:

- Bayesian Belief Network (BBN)
- Multinomial Naive Bayes Bayesian Network (BN)
- Averaged One-Dependence Estimators (AODE)
- Gaussian Naive Bayes
- Naive Bayes

2.2.5. Data Clustering algorithms

This algorithm split items into different types of batches. It groups the item set into clusters in which each subset share some similarity. It is unsupervised learning method and its methods are categorized as hierarchical or network clustering and partitioned clustering. The most well known algorithms for clustering are:

- K Means
- Expectation Maximisation (EM)
- K Medians
- Hierarchical-Clustering

2.2.6. Learning algorithms using Association Rule

Learning algorithms using Association rule are generally utilized by the organization commercially when multidimensional datasets are huge in size. They are used as extraction methods that can

explore observed relationships between variables and data. The most well known learning algorithms using association rule are:

- Eclat algorithm
- Apriori algorithm

2.2.6. Algorithms using Artificial Neural Network

Artificial neural networks models are based on the biological neuron structure and uses supervised learning. It consists of artificial neurons which have weighted interconnections among units. They are also well known by parallel distributed processing networks. The most famous or well known algorithms for artificial neural network are:

- Radial Basis Function Network (RBFN)
- Back-Propagation
- Perceptron
- Hopfield Network

2.2.7. Deep Learning algorithms

Deep Learning methods upgraded the artificial neural networks They are more complex neural networks are large in size. The most Famous algorithms for deep learning are:

- Deep Belief Networks
- Stacked Auto Encoders
- Deep Boltzmann Machine (DBM)
- Convolution Neural Network (CNN)

2.2.8. Algorithms using Dimensionality Reduction

Dimensionality reduction method is widely used in case of large nuber of dimensions, large volume of space concerned. Then that problem requires a statistical significance. Dimensionality reduction methods used for minimizing the number of dimensions outlined the item and removes unrelated and unessentaial data which lessen the computational cost. Some of these methods are used in classifying and regression. The algorithms using reduction in dimensionality are as follows:

- Partial Least Squares Regression
- Multidimensional Scaling
- Principal Component Analysis
- Flexible Discriminant Analysis
- Mixture Discriminant Analysis
- Sammon Mapping
- Projection Pursuit
- Linear Discriminant Analysis
- Principal Component Regression
- Quadratic Discriminant Analysis

2.2.9 Ensemble Algorithms

They are based on unsupervised Learning. It groups the teaching data into many types of classes of data. self-supporting models for learning are built for those groups. To make correct hypothesis all learning models are combined . The well known ensemble algorithms are

- Gradient Boosting Machines
- Boosting
- Gradient Boosted Regression Trees
- Bagging
- Bootstrapped Aggregation
- Stacked Generalization (blending)
- AdaBoost
- Random Forest

3. Literature review

Rob Law (1998) [7] applies neural networks to forecasts occupancy rates for the rooms of Hong Kong hotels and finds that neural networks outperforms naïve extrapolation model and also superior to multiple regression. This research studied the feasibility incorporating the neural network to predict the rate of occupancy of rooms in Hong Kong hotel industry.

Authors **Hua et al. (2006) [8]** described support vector machines approach to predict occurrences of non zero demand or load time demand of spare parts which used in petrochemical enterprise in china for inventory management. They used a integrated procedure for establishing a correlation of explanatory variables and autocorrelation of time series of demand with demand of spare parts. On performing the comparison the performance of SVM method with this LRSVM model, Croston's model, exponential smoothing model, IFM method and Markov bootstrapping procedure., it performs best across others.

Authors **Vahidov et al. (2008) [9]** compares the methods of predicting demand in the last of a supply chain, the naïve forecasting and linear regression and trend moving average with advanced machine learning methods such as neural networks and support vector machines, recurrent neural networks finds that recurrent neural networks and support vector machines show the best performance.

Wang (2007) [10] describes the machine learning method with genetic algorithm (GA)-SVR with real value GAs. The experimental findings investigates this, SVR outshines the ARIMA models and BPNN regarding the base the normalized mean square error and mean absolute percentage error.

Authors **Chen et al. (2011) [11]** presents a method forecast the tourism demands that is SVR built using chaotic genetic algorithm (CGA), like SVRCGA, which overcome premature local optimum problem. This paper reveal that suggested SVRCGA model outclass other methodologies reviewed in the research paper.

Turksen et al. (2012) [12], presents next-day stock price prediction model which is based on a four layer fuzzy multi agent system (FMAS) structure. This artificial intelligence model used the coordination of intelligent agents for this task. Authors investigates that FMAS is a suitable tool for stock price prediction problems as it outperforms all previous methods.

Shahrabi et al. (2013) [13] proposed a method for estimating tourism demand which is a new combined intelligent model i.e. Modular Genetic-Fuzzy Forecasting System using a genetic fuzzy expert systems and finds that accuracy of predicting power of MGFFS is better than approaches like Classical Time Series models, so it is suitable estimating tool in tourism demand prediction problems.

Chen Hung et al. (2014) [14] proposes forecasting model for tourists arrival of Taiwan and Hong Kong named as LLSSVR or logarithm least-squares support vector regression technologies. In combinations with fuzzy c-means (FCM) and Genetic algorithms (GA) were optimally used and indicates that method explains a better performance to other methods in terms of prediction.

Guang-Bin Huang et al. (2015) [15] explores the basic features of ELMs such as kernels, random features and random neurons, compares the performance of ELMs and shows it tend to outshine classification, support vector machine and regression applications

Wang et al. (2016) [16] proposed a novel forecasting method CMCSGM based Markov-chain grey model which used algorithm of

Cuckoo search optimization to make better the performance of the Markov chain grey model. The resultant study indicates that the given model is systematic and fine than the traditional MCGM models.

Barzegar et al. (2017) [17] demonstrates model predict multi-step ahead electrical conductivity i.e. indicator of water quality which is needed for estimating the mineralization, purification and salinity of water based on wavelet extreme learning machine hybrid or WAELM models and extreme learning machine which exploiting the boosting ensemble method. The findings showed that upgrading multi WAELM and multi WAANFIS ensemble models outshines the individual WAELM and WAANFIS constructions.

Authors **Fouillo et al. (2018) [18]** suggested a statistical method employing machine learning model and to analyze and applied it to solar irradiation prediction working hourly. This methodology used the high, low and medium meteorological variability like Ajacio, Odeillo, Tilos. They compared model with auto regressive moving average and multi-layer preceptor.

Makridakis et al. (2018) [19] presents Machine Learning methods to statistical time series forecasting and compared the correctness of those methods with the correctness of conventional statistical methods and found that the first one is better and outtop using the both measures of accuracy. They provide the reason for the accuracy of learning models is less that of statistical models and suggested some other achievable ways.

Zhang et al. (2018) [20] suggests a design of multi kernel ELM or MKELM method for segregation of motor imagery electroencephalogram or EEG and investigate performance of kernel ELM and impacts of two different functions of kernel such as polynomial and Gaussian kernel. Compares MKELM method gives greater segregation accuracy than other algorithms indicates betterment of the suggested MKELM based.

4. Applications of Machine-Learning

In the research paper we studied various Machine-learning techniques such as supervised and unsupervised learning. Supervised learning is applied in classification problems like face recognition, medical diagnosis, pattern recognition, character recognition, web advertizing [22].

Unsupervised learning can be applied in clustering, association analysis, CRM, summarization, image compression, bioinformatics. reinforcement learning is widely applied in game playing and robot control [23].

5. Tools used in Machine Learning

Tools makes machine learning swift and rapid. Machine learning tools provides interface to the machine learning programming language. They provide best practices for process and implementation [23]. Machine learning tools contains platforms which provides capabilities to run a module or project. Examples of platforms of machine learning are:

- **Python SciPy subparts such as scikit-learn, Panda**
- **R Platform.**
- **WEKA Machine Learning Workbench.**
-

Machine learning tools contains various libraries which provides all capabilities to complete a project and libraries provides various algorithms. Some of libraries are:

- **JSAT in Java.**
- **scikit-learn in Python.**

- Accord Framework in .NET

6. Conclusion

Machine learning methods and algorithms have been reviewed in this paper. This paper also reviewed algorithms describing the various types of machine learning techniques, algorithms and methodology. Various applications of Machine learning and many tools needed for processing are also being reviewed. In the Literature review section we studied various machine learning algorithms implemented in past years in different areas in combination with the tradition methods and studied how they outperformed the previous models.

References:

- Mariette Awad, Rahul Khanna. "Efficient Learning machines: Concepts and Applications". *Aspress Publishers*, 2015.
- Teng Xiuyi, Gong Yuxia. "Research on Application of Machine Learning in Data Mining". *IOP Conf. Series: Materials Science and Engineering*, 2018.
- M. Praveena, V. Jaiganesh, "Literature Review on Supervised Machine Learning Algorithms and Boosting Process". *International Journal of Computer Applications*, ISSN No. 0975 – 8887, vol. 169, 2017.
- Kajaree Das, Rabi Narayan Behera. "A Survey on Machine learning: Concept, Algorithms and Applications", *International Journal of Innovative Research in Computer and communication Engineering*. vol. 5, 2017.
- S.B. Kotsiantis. "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*. pp 249-268, 2007.
- Rob Law, "Room occupancy rate forecasting: a neural network approach", *International Journal of Contemporary Hospitality Management*, vol. 10 Issue 6, pp 234 – 239, 1998.
- Zhongsheng Hua, Bin Zhang, "A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts" *Applied Mathematics and Computation* 181, pp 1035–1048, 2006.
- Real Carbonneau, Kevin Laframboise, Rustam Vahidov, "Application of machine learning techniques for supply chain demand forecasting", *European Journal of Operational Research* 184, pp 1140–1154, 2008.
- Kuan-Yu Chen, Cheng-Hua Wang, "Support vector regression with genetic algorithms in forecasting tourism demand", *Tourism Management* 28, pp 215–226, 2007.
- Wei-Chiang Hong, Yucheng Dong, Li-Yueh Chen, Shih-Yung Wei, "SVR with hybrid chaotic genetic algorithms for tourism demand forecasting", *Applied Soft Computing* 11, pp 1881–1890, 2011.
- M. H. Fazel Zarandi, Esmail Hadavandi, B. Turksen, "A Hybrid Fuzzy Intelligent Agent-Based System for Stock Price Prediction", *International Journal Of Intelligent Systems*, Vol. 00, pp 1–23, 2012.
- Jamal Shahrabi, Esmail Hadavandi, Shahrokh Asadi, "Developing a hybrid intelligent model for forecasting problems: Case study of tourism demand time series", *Knowledge-Based Systems* 43, pp 112–122, 2013.
- Ping-Feng Pai, Kuo-Chen Hung, Kuo-Ping Lin, "Tourism demand forecasting using novel hybrid system", *Expert Systems with Applications* 41, pp 3691–3702, 2014.
- Guang-Bin Huang, "An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels", *Springer*, 2014.
- Xu Sun, Wangshu Sun, Jianzhou Wang, Yixin Zhang, Yining Gao, "Using a Greye Markov model optimized by Cuckoo search algorithm to forecast the annual foreign tourist arrivals to China", *Tourism Management* 52, 2016.
- Rahim Barzegar, Asghar Asghari Moghaddam, Jan Adamowski, Bogdan Ozga-Zielinski, "Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model", *Springer*, ISBN 00477-017-1394, 2017.
- Alexis Fouilloy, Cyril Voyant, Gilles Notton, Fabrice Motte, Christophe Paoli, Marie-Laure Nivet, Emmanuel Guillot, Jean-Laurent Duchaud, "Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability", *Energy* 165, 2018.
- Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward", *PLoS ONE* 13, 2018.
- Yu Zhang, Yu Wang, Guoxu Zhou, Jing Jin, Bei Wang, Xingyu Wang, Andrzej Cichocki, "Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces", *Expert Systems With Applications* 96, pp 302–310, 2018.
- Pedro Domingos, "A Few useful Things to Know about Machin Learning". 2012.
- Yogesh Singh, Pradeep Kumar Bhatia & Omprakash Sangwan "A review of studies in machine learning technique". *International Journal of Computer Science and Security*, vol.1, pp 70 – 84, 2007.
- Petersp, "The Need for Machine Learning is Everywhere" March 10, 2015.
- Jason Brownlee, "A Tour of Machine Learning algorithms" November 25, 2013.
- Jason Brownlee, "Machine Learning Tools", December 28, 2015.
- Taiwo Ayodele, "Types of Machine Learning algorithms", 2010.
- Teng Xiuyi, Gong Yuxia, "Research on Application of Machine Learning in Data Mining, IOP Conf. Series: Materials Science and Engineering 392 (2018)