Shodhsamhita : Journal of Fundamental & Comparative Research
Vol. VII, No. 11 : 2021
ISSN: 2277-7067

MACHINE LEARNING: CONCEPTS, ALGORITHMS AND APPLICATIONS - A SURVEY

A. Muruganandham *Assistant Professor of Computer Applications (BCA)*
Nallamuthu Gounder Mahalingam College, Pollachi, India.
E-mail: muruganandham24@gmail.com

Abstract

Machine Learning has unfold from the Artificial Intelligence, a ground of computer science .It is multidisciplinary field ,a mixture of statistics and computer science algorithms which is widely used in predictive analyses and classification. The second part of the paper focus to influence the necessary machine learning methods and algorithms. This paper will go through the variety of machine learning tools needed to run the machine learning projects. The main concern of this paper is to study the main approaches and case studies 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.

I. INTRODUCTION

Over the past decades, Artificial intelligence flow has become the broad and exciting field in science as it prepare the machines to perform the tasks that human being may do and it aims to train the computers to solve real world troubles with the utmost success rate. As perceiving scientific growth and progress in technology AI systems are now capable to learn and improve through past experiences without explicitly assistance code if they uncovered to new data. Eventually it leads to technology of Machine learning which uses learning algorithms to learn from the data available [1]. ML uses data mining techniques to extract the information from the huge size datasets. ML and Data Mining techniques explore data from end to end to find the hidden patterns inside dataset [2]. Machine Learning and data mining algorithms has 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].

II. METHODS USED IN MACHINE LEARNING

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

2.1 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 generate a mapping function to map inputs to required outputs. Parameter adjustment process continues until the system acquired a suitable accuracy extent regarding the teaching data.

2.2 Unsupervised Learning:

In this algorithm we only have training data rather a outcome data. That input data is not earlier labeled. It is used in classifiers by recognizing existing patterns or cluster in the input datasets [4].

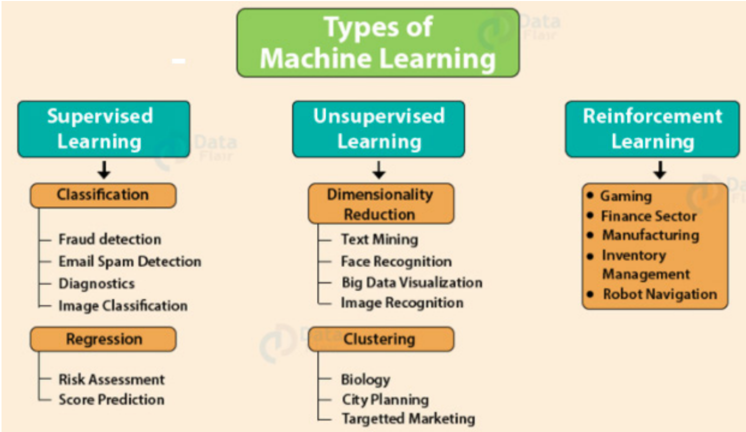

Fig.1 Types of Machine Learning

2.3 Reinforcement learning:

Applying this algorithm 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

III. ALGORITHM OF MACHINE LEARNING

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

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms
    1) GBM
    2) XGBoost
    3) LightGBM
    4) CatBoost

1. Linear Regression:

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.

2. Logistic Regression:

Don’t get confused by its name! It is a classification not a regression algorithm. It is used to estimate discrete values based on given set of independent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

3. Decision Tree:

This is one of my favorite algorithms and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

4. SVM (Support Vector Machine):

It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

5. Naive Bayes:

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

6. kNN (k- Nearest Neighbors):

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function.

7. K-Means:

It is a type of unsupervised algorithm which solves the clustering problem. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups.

8. Random Forest:

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we've collection of decision trees (so known as "Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class.

9. Dimensionality Reduction Algorithms:

In the last 4-5 years, there has been an exponential increase in data capturing at every possible stages. Corporate/ Government Agencies/ Research organizations are not only coming with new sources but also they are capturing data in great detail.

10. Gradient Boosting Algorithms

10.1. GBM:

GBM is a boosting algorithm used when we deal with plenty of data to make a prediction with high prediction power. Boosting is actually an ensemble of learning algorithms which combines the calculation of several base estimators in order to improve robustness over a single estimator. It combines multiple weak or average predictors to a build strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix.

10.2. XG Boost:

Another classic gradient boosting algorithm that's known to be the decisive choice between winning and losing in some Kaggle competitions.

The XGBoost has an immensely high predictive power which makes it the best choice for accuracy in events as it possesses both linear model and the tree learning algorithm, making the algorithm almost 10x faster than existing gradient booster techniques.

10.3. Light GBM:

Light GBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages Faster training speed and higher efficiency, Lower memory usage, Better accuracy, Parallel and GPU learning supported, Capable of handling large-scale data.

The framework is a fast and high-performance gradient boosting one based on decision tree algorithms, used for ranking, classification and many other machine learning tasks. It was developed under the Distributed Machine Learning Toolkit Project of Microsoft.

10.4. Catboost:

CatBoost is a recently open-sourced machine learning algorithm from Yandex. It can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML.

IV. LITERATURE REVIEW

Hua et al. (2006) [8] described support vector machines approach to predict occurrences of non zero demand or load time insist of spare parts which used in petrochemical enterprise in china for register management. They used a included procedure for establishing a correlation of explanatory variables and autocorrelation of time series of demand with demand of spare parts. On performing the judgment 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.

Vahidov et al. (2008) [9] compares the methods of predicting demand in the last of a supply chain, the naive 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 .

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 WA ELM and multi WAANFIS ensemble models outshines the individual WAELM and WA ANFIS constructions.

Fouilloy 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.

V. 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].

V1. TOOLS USED IN MACHINE LEARNING

Tools make machine learning swift and rapid. ML tools provide interface to the ML programming language. They provide best practice for process and implementation [23]. ML tools contain platforms which provide 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.

ML tools contain 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

VII. 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. Different applications of Machine learning and many tools needed for processing are also being reviewed. In the Literature review section, we come across a variety of machine learning algorithms implemented in past years in different areas in combination with the traditional methods and learnt how they outperformed the previous models.

REFERENCES

[1] Mariette Awad, Rahul Khanna. “Efficient Learning machines: Concepts and Applications”. Aspress Publishers, 2015.

[2] Teng Xiuyi1,Gong Yuxia1. “Research on Application of Machine Learning in Data Mining”. IOP Conf. Series: Materials Science and Engineering, 2018.

[3] 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.

[4] 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.

[5] S.B. Kotsiantis."Supervised Machine Learning: A Review of Classification Techniques", Informatica. pp 249-268, 2007.

[6] Rob Law,"Room occupancy rate forecasting: a neural network approach", International Journal of Contemporary Hospitality Management, vol. 10 Issue 6, pp 234 – 239, 1998.

[7] 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.

[8] 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.

[9] Kuan-Yu Chen, Cheng-Hua Wang, "Support vector regression with genetic algorithms in forecasting tourism demand", Tourism Management 28, pp 215–226, 2007.

[10] 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.

[11] M. H. Fazel Zarandi, Esmaeil 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.

[12] Jamal Shahrabi , Esmaeil 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.

[13] 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.

[14] Guang-Bin Huang, "An Insight into Extreme Learning Machines: Random Neurons,Random Features and Kernels", Springer, 2014.

[15] 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.

[16] 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.

[17] 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.

[18] Spyros Makridakis, Evangelos Spiliotis , Vassilios Assimakopoulos, " Statistical and Machine Learning forecasting methods: Concerns and ways forward", PLoS ONE 13,2018.

[19] 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.

[20] Pedro Domingos, "A Few useeful Things to Know about Machin Learning".2012.

[21] 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.

[22] Petersp,"The Need for Machine Learning is Everywhere" March 10, 2015.

[23] Jason Brownlee,"A Tour of Machine Learning algorithms" November 25, 2013.

[24] Jason Brownlee,"Machine Learning Tools", December 28, 2015.

[25] Taiwo Ayodele,"Types of Machine Learning algorithms", 2010.

[26] Teng Xiuyi1, Gong Yuxia1, Research on Application of Machine Learning in Data Mining,IOP Conf. Series: Materials Science and Engineering(2018)