Available online at: http://www.ijmtst.com/vol5issue07.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 05, Issue No: 07, July 2019

A Study on Machine Learning and Its Working

B Jayaram[1] | Dr. M Kalimuthu[2]

[1] Assistant Professor, CSE Department, Malla Reddy Institute of Technology, Hyderabad – 500100
[2] Associate Professor, CSE Department, Malla Reddy Institute of Technology, Hyderabad – 500100

To Cite this Article

B Jayaram and Dr. M Kalimuthu, "A Study on Machine Learning and Its Working", International Journal for Modern Trends in Science and Technology, Vol. 05, Issue 07, July 2019, pp: 44-48.

Article Info

Received on 16-June-2019, Revised on 14-July-2019, Accepted on 19-July-2019, Published on 23-July-2019.

ABSTRACT

Machine learning (ML) is widely popular these days and used in wide variety of domains for prediction of outcomes. In machine learning lot of algorithms exists for predicting the outcomes. But choosing the right algorithm according to the domain plays a very important in deciding the performance of the algorithm. This paper consists of five sections is organized in following way first section deals about collection of data from various resources, second section deals about data cleaning, third part deals about choosing the correct ML algorithm, fourth part deals about gaining knowledge from models and final part deals about data visualization

I. INTRODUCTION

Machine learning is a category of algorithms that allows software applications to be more precise in predicting outcomes without being explicitly programmed. Machine learning can also be used to build algorithms that can receive input data and use statistical analysis to predict an output or outcomes [1] and create a model for new model to be developed.

The machine learning process is generally a collection of data mining and predictive modeling concepts where both require searching through data to look for patterns and adjusting program actions accordingly [1].

Some of the well-known examples for machine learning is Facebook's News Feed. The News Feed uses machine learning to personalize each member's feed [1]. And you tube for searching videos, Google search engine and other web source also uses machine learning algorithms for prediction of each member preference in searching their data

Machine learning is also used in Customer relationship management (CRM) systems for learning models to analyze email and prompt sales team members to respond to the most important messages first.

Machine learning also plays a very important role in various fields of Artificial Intelligence (AI), Internet of Things (IOT), health care etc.

Generally machine learning algorithms can be classified in to two parts

- Supervised learning algorithms.
- Unsupervised learning algorithms.
- Reinforcement algorithms.

Supervised learning algorithms requires, a data analyst with learning skills to provide both input and desired output. And also provide details about accuracy of predicted data by providing feedback. Data analyst determines which variables, or features, the machine learning model should analyze and use to develop predictions.

Knowledge obtained from this output can be used to train other models.

In supervised learning model there are two parts in the first part referred as training, it starts with raw data the input is passed to a feature extraction which contains the feature extraction matrix and train the models and evaluate them and labels. In the second part called as prediction the new data is used for feature extraction through feature vector and predict the labels which gets the input from the training part as shown in figure 1.

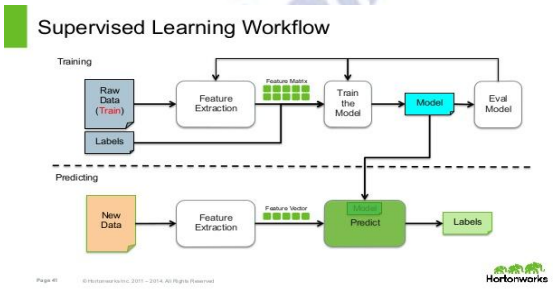

Fig 1:Supervised Learning Workflow

The lists of few supervised algorithms are listed below.

- Decision Trees
- Naive Bayes Classification
- Support vector machines for classification problems
- Random forest for classification and regression problems
- Linear regression for regression problems
- Ordinary Least Squares Regression
- Logistic Regression
- Ensemble Methods

Unsupervised learning algorithms do not need to be trained with desired outcome data, but it uses deep learning [1] approach to review data and come to conclusions. Unsupervised learning algorithms are also called as neural networks. Unsupervised learning algorithms are used in various applications such as image processing and speech to text conversion.

In unsupervised learning from the raw data we get a scaled data to build a model and validate it according to the requirements and validate the model to produce an effective prediction as shown in figure 2.

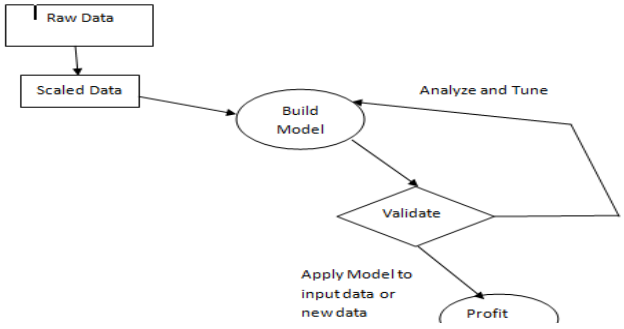

Fig 2: Unsupervised Learning Workflow

Some common unsupervised algorithms are listed below

- K-means for clustering problems
- Apriori algorithm for association rule learning problems
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis

Reinforcement learning algorithm [6]is a type of algorithm where the machine works by exposing the system to environment where the machine trains itself continuously using trial an error method. Here the machine learns from previous learning models and gains knowledge to make correct decisions.

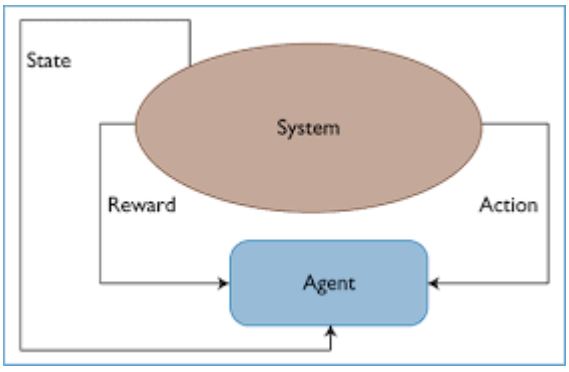

Figure 3: Reinforcement learning workflow

Every **action** performed by the Agent yields a reward from the environment. The decision of which **action** to choose is made by the policy. Agent: The **learning** and acting part of a **Reinforcement Learning** problem, which tries to maximize the **rewards** it is given by the Environment as per figure 3.

Some of the existing reinforcement learning algorithms are listed below [7]

- Q-Learning.
- State Action Reward State Action (SARSA).
- Deep Q Network.
- Deep Deterministic Policy Gradient (DDPG).

Though there are advantages and disadvantages of all the above mentioned algorithms in both supervised and unsupervised learning it is not in the scope of this paper to discuss about them since it is very elaborate. So this paper only concentrates how to choose a correct machine learning algorithm and get a desired or predicted output.

II. How Machine Learning Works

Machine learning algorithms working can be classified into the following steps.

- Gathering data from various sources
- Cleaning the data
- Model Building.
- Gaining knowledge from models
- Visualization of data

Here each step is broadly classified in to a number of sub steps and detailed explanation is mentioned below.

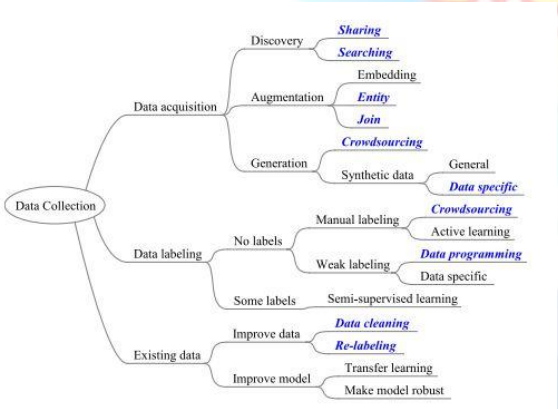

Figure 4: Data collection for machine learning [3]. Here some of topics mentioned (blue color italics) are least concerned.

2.1 Gathering of data from various sources

Data collection can be classified in to three broad categories namely 1) data acquisition 2) data labeling and 3) existing data [3].

Data acquisition is a process to find datasets that can be used to train learning models and there are three different approaches for this process [3].

Data Generalization can be used when there no external data set but when there is a possibility to generate crowd sourcing or synthetic data [3].

Data Labeling is an interlinked process with data acquisition since it starts only after data gathering is completed.

Data labeling can be done by two different methods

- Using existing labels.
- Using crowd based labels.

Both these methods follow the process of machine learning algorithm classification and regression using semi supervised learning [3].

Existing data can be used to train models in machine learning by two ways. First by improving the data using data cleaning techniques described in next section and data labeling techniques, Second by improving the model by using various learning models [3].

2.2 Cleaning data to have homogeneity

It is a process of dealing with missing values and deciding what can be done with outliers [2].

Missing values leads to various problems related to wrong prediction and in some cases the missing data might even be very sensitive data there leading to produce a wrong model which cannot be used properly. So choosing a correct algorithm to deal with missing data also plays a very important role in performance of machine learning. So to produce effective prediction missing data can be replaced by the following set of procedures [4].

- Deleting rows that have missing data.
- Replacing missing data with mean or median or mode.
- Assigning a unique category.
- Predicting the missing values.
- Using algorithms which support missing values.

Outliers are a part of any multidimensional data which lies a distance away from any consideration for the model and the reason for this category will probably be poor collection of data [4].

2.3 Model Building - Selecting the Right machine learning algorithm.

This step can be broadly classified in to the following steps [5]:

- Know your data.
- Categorize the problem.
  - Categorize by input.

- Categorize by output.
- Find the available algorithms.

For **knowing our data** we need to look at the summary statistics and visualizations (eg: Percentages, average, median etc).

- Visualize the data by using different types of plotting (box plotting, density plotting etc)

For categorizing the problem there are two different steps:

- Categorization by input.
- Categorization by output.

In case of **categorization by input** it means the following is applicable.

- Supervised learning algorithm is followed if we have a labeled data.
- Unsupervised learning algorithm is followed if we have aunlabelled data and we want to find its structure.
- Reinforcement learning problem is followed if the objective function is to be minimized.

In case of **categorization by output** the following is applicable.

- If the output is a model it is identified as regression problem.
- If the output of the model is set into input groups then it is referred as clustering problem.
- If the output of the model is a class then it will be a classification problem.
- If any anomaly has to be detected then it is a problem of anomaly deduction.

Constraints to be considered are:

- Data storage capacity.
- Speed of prediction.
- Does the learning have to be fast?

After understanding the problem and categorizing the problem, the next step is **finding the available algorithms** that suit the current problem. Some of the existing

algorithms are prescribed in the previous section under supervised and unsupervised learning is followed.

2.4 Gaining knowledge from model result

For gaining knowledge from the model the we can use the help any of machine learning tools.

These tools helps the user to gain knowledge from the output. Since they have following main advantages [8].

- No programming required to work with the tool. Since they are automatic.
- Better management of work.
- Produces result faster than human processing.
- Produces better quality of result.
- It has better uniformity.

2.5 Visualization – Transforming results in visual representation

For visualizing the result in an understandable form there are lot free open source tools available in market for machine learning tools. A few of them include

- Weka.
- R Programming.
- Python.
- Apache spark.
- Sas.
- Rapid miner.
- KNIME.
- Splunk.
- QLinkView.
- Orange.

Sample visualization is given using weka tools below

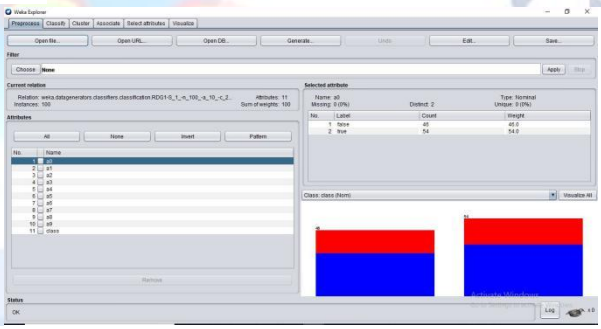

Figure 5: A Sample Screen from WEKA Tool.

This tools helps the user to select the corresponding data set and generate whatever is the need for the end user. Thus helps the end user to understand the system easily with the help of representation in graphs. This tool mainly follows supervised and unsupervised learning models.

III. Conclusion

In this paper we have discussed more about introduction to machine learning and different types of learning algorithms that are in existence and how is the work flow organized according to the type of algorithm.

Future Work

In future work we have to focus more on the working of the tools and the correct algorithm for producing the effective output for prediction of machine learning.

REFERENCES

[1] https://searchenterpriseai.techtarget.com/definition/machine-learning-ML

[2] https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1

[3] Yuji Roh, Geon Heo, Steven Euijong Whang, Member, IEEE , "A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective" arxiv.org, November 2018\

[4] https://www.analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/

[5] https://www.goodworklabs.com/machine-learning-algorithm/

[6] https://www.cuelogic.com/blog/choosing-right-machine-learning-model-combinations-for-business-problem-at-hand

[7] https://towardsdatascience.com/introduction-to-various-reinforcement-learning-algorithms-i-q-learning-sarsa-dqn-ddpg-72a5e0cb6287

https://dimensionless.in/top-10-data-science-tools/