International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 395 - 399

A review on Machine Learning Techniques

Sandhya N. dhage
Assistant Professor,Information technology
Bapurao deshmukh college of Engineering,
Sevagram,wardha,India
sandhyadhage@yahoo.co.in

Charanjeet Kaur Raina
Assistant Professor,Information technology
Bapurao deshmukh college of Engineering,
Sevagram,wardha,India
ckraina23@gmail.com

***Abstract*—** Machine learning is the essence of artificial intelligence. Machine Learning learns from past experiences to improve the performances of intelligent programs. Machine learning system builds the learning model that effectively "learns" how to estimate from training data of given example. IT refers to a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction, based on models derived from existing data. In this new era, Machine learning is mostly in use to demonstrate the promise of producing consistently accurate estimates. The main goal and contribution of this review paper is to present the overview of machine learning and provides machine-learning techniques. Also paper reviews the merits and demerits of various machine learning algorithms in different approaches.

Keywords-*machine learn;supervised learnig;unsupervised learning*

*****

I. INTRODUCTION

Machine learning is multidisplinary field in artificial intelligence, probability, statistics, information theory, philosophy, psychology, and neurobiology. Machine learning solves the real world problems by building a model that is good and useful approximation to the data. The study of Machine learning has grown from the efforts of exploring whether computers could learn to mimic the human brain, and a field of statistics to a broad discipline that has produced fundamental statistical computational theories of learning processes. In 1946 the first computer system ENIAC was developed. The idea at that time was that human thinking and learning could be rendered logically in such a machine. In 1950 Alan Turing proposed a test to measure its performance. The Turing test is based on the idea that we can only determine if a machine can actually learn if we communicate with it and cannot distinguish it from another human. Around 1952 Arthur Samuel (IBM) wrote the first game-playing program, for checkers, to achieve sufficient skill to challenge a world champion. In 1957 Frank Rosenblatt invented the perceptron which connects a web of points where simple decisions are made that come together in the larger program to solve more complex problems. In 1967, pattern recognition is developed when first program able to recognize patterns were designed based on the type of algorithm called the nearest neighbor. In 1981, Gerold Dejong introduced explanation based learning where prior knowledge of the world is provided by training examples which makes the use of supervised learning. In the early 90's machine learning became very popular again due to the intersection of Compter Science and Statistics.

Advances continued in machine learning algorithm within the general areas of supervised and unsupervised learning. In the present era, adaptive programming is in explored which makes use of machine learning where programs are capable of recognizing patterns, learning from experience, abstracting new information from data and optimizing the efficiency and accuracy of its processing and output. In the discovery of knowledge from the multidimensional data available in a diverse amount of application areas, machine learning techniques are used.

Because of new computing technologies, machine learning today is not like machine learning of the past. Though many machine learning algorithms have been developed from long time, recent development in machine learning is the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster.

More interest is developed in machine learning today is because of growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results even on a very large scale. Machine learning model produces high-value predictions that can guide better decisions and smart actions in real time without human intervention.

Machine learning do not just respond to current demand, but to be able to predict demand in real time. As computation gets cheaper, machine learning makes the impossible things possible and people tend to start doing them and end up with making intelligent infrastructure. It is the need to develop newer algorithms to advance the science of machine learning and huge amount of work that needs to be done to replace existing algorithms from new algorithms. To make the existing algorithms more robust and consumable it is not important to develop a perfect model algorithm, because a perfect model will not be the final products data often has only a temporal value.

The paper is organized as follows: Section II describes the machine learning model. Section III provides the overview of the machine learning methods. Section IV discusses the various learning algorithms used to perform learning process. Section V describes the application based

on type of learning examples and different machine learning tools. Section VI concludes the manuscript.

II. MACHINE LEARNING MODEL

Learning process in machine learning model is divided into two steps as

- Training
- Testing

In training process, samples in training data are taken as input in which features are learned by learning algorithm or learner and build the learning model. In the testing process, learning model uses the execution engine to make the prediction for the test or production data. Tagged data is the output of learning model which gives the final prediction or classified data.

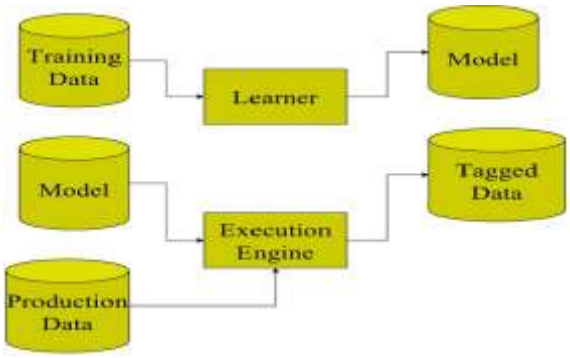

Figure 1. Operational model of machine learning

III. MACHINE LEARNING TECNIQUES

Machine learning techniques are classified into three broad categories, depending on the nature of the learning signal or feedback available to a learning system as follows:

A. Supervised learning

Supervised learning is trained using labeled examples, such as an input where the desired output is known. Supervised learning provides dataset consisting of both features and labels. For example, a piece of equipment could have training data points labeled either as F (failed) or as R(runs). The task of supervised learning is to construct an estimator which is able to predict the label of an object given the set of features. The learning algorithm receives a set of features as inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. This model is not needed as long as the inputs are available, but if some of the input values are missing, it is not possible to infer anything about the outputs.

Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim. Another application is predicting the species of iris given a set of measurements of its flower. Other more complicated examples includes recognition system as given a multicolor image of an object through a telescope, determine whether that object is a star, a quasar, or a galaxy, or given a list of movies a person has watched and their personal rating of the movie, recommend a list of movies they would like.

Supervised learning tasks are divided into two categories as classification and regression. In classification, the label is discrete, while in regression, the label is continuous. For example, in astronomy, the task of determining whether an object is a star, a galaxy, or a quasar is a classification problem where the label is from three distinct categories. On the other hand, in regression problem, the label (age) is a continuous quantity, for example, finding the age of an object based on observations.

Supervised learning model is given in figure 2 which shows that algorithm makes the distinction between the raw observed data X that is training data which may be text, documemt or image and some label given to the model during training. In the process of training, supervised learning algorithm builds the predictive model. After training, the fitted model will try to predict the most likely labels for new a set of samples X in test data. Depending on the nature of the target y, supervised learning can be classified as follows:

- If y has values in a fixed set of categorical outcomes (represented by integers) the task to predict y is called classification.
- If y has floating point values (e.g. to represent a price, a temperature, a size...), the task to predict y is called regression.

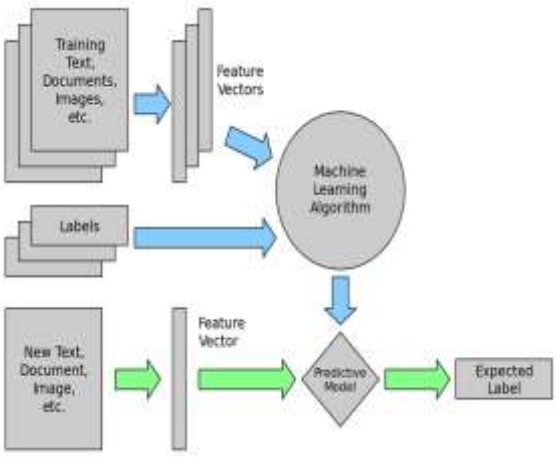

Figure 2. Supervised learning model

B. Unsupervised Learning

Unsupervised learning used data that has no historical labels and the goal is to explore the data and find similarities between the objects. It is the technique of discovering labels from the data itself. Unsupervised learning works well on transactional data such as identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find

the main attributes that separate customer segments from each other.

Other unsupervised learning problems are:

- given detailed observations of distant galaxies, determine which features or combinations of features are most important in distinguishing between galaxies.
- given a mixture of two sound sources for example, a person talking over some music, separate the two which is called the blind source separation problem.
- given a video, isolate a moving object and categorize in relation to other moving objects which have been seen.

Typical unsupervised task is clustering where a set of inputs is divided into groups, unlike in classification, the groups are not known before. Popular unsupervised techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition. These algorithms are also used to segment text topics, recommend items and identify data outliers.

The unsupervised learning model is given in figure 3 which shows that unsupervised learning algorithm only uses a single set of observations X with n samples and n features and does not use any kind of labels. In the training process, unsupervised learning algorithm builds the predictive model which will try to fit its parameters so as to best summarize regularities found in the data.

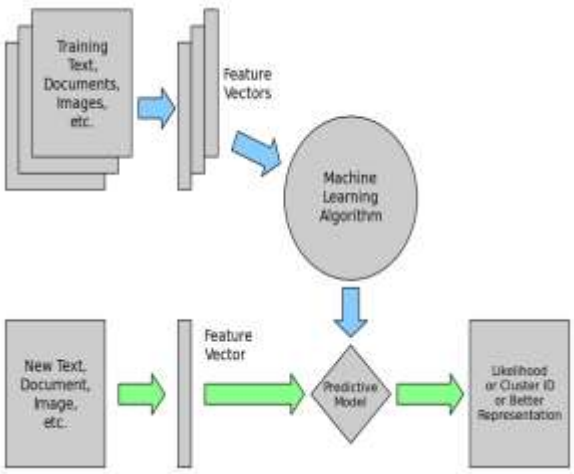

Figure 3. Unsupervised learning model

C. Semi-supervised Learning

In many practical learning domain such as text processing, video indexing, bioinformatics, there is large supply of unlabeled data but limited labeled data which can be expensive to generate .So semi supervised learning is used for the same applications as supervised learning but it uses both labeled and unlabeled data for training. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions. Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. This type of learning can be used with methods such as classification, regression and prediction. Early examples of this include identifying a person's face on a web cam. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabelled data.

D. Reinforcement Learning

It is often used for robotics, gaming and navigation. It is the learning technique which interacts with a dynamic environment in which it must perform a certain goal without a teacher explicitly telling it whether it has come close to its goal. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. So in the chess playing, reinforcement learning learns to play a game by playing against an opponent which performs trial and error actions to win.

This type of learning has three primary components: the learner, the environment and actions. The objective is for the learner to choose actions that maximize the expected reward over a given amount of time. The learner will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

IV. MACHINE LEARNING ALGORITHM

A large set of machine learning algorithms are developed to build machine learning models and implement an iterative machine learning process. These algorithms can be classified on the basis of learning style as follows:

A. Regression algorithms

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression is the task of predicting the value of a continuously varying variable such as a price, a temperature if given some input variables like features and regressor. The most popular regression algorithms are:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

B. Instance-based Algorithms

Instance based learning model a decision problem with instances or examples of training data that are deemed important or required to the model.Such methods typically build up a database of training data and compare test data to the database using a similarity measure in order to find the best match and make a prediction. Instance-based methods are also called lazy learner. Lazy learning simply stores training data and waits until it is given a test data then performs the learning. So lazy learner takes less time in training but more time in predicting. The most popular instance-based algorithms are:

- k-Nearest Neighbour (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

C. Decision Tree Algorithms

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning. The most popular decision tree algorithms are:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

D. Bayesian Algorithms

Machine Learning is a hybrid of Statistics and algorithmic Computer Science. Statistics is about managing and quantifying uncertainty. To represent all forms of uncertainty, bayesian algorithms are used which are based on probability theory. Bayesian methods are those that are explicitly apply Bayes' Theorem for problems such as classification and regression. The most popular Bayesian algorithms are:

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

E. Clustering Algorithms

Clustering is the method of classification of objects into different groups. It partitions the data set into subsets or clusters, so that the data in each subset share some common trait often according to some defined distance measure. Clustering is the type of unsupervised learning. Clustering like regression describes the class of problem and the class of methods. Clustering methods are classified as hierarchical clustering and partitional clustering. K-means is partional clustering algorithms which uses centroid-based approach. The most popular clustering algorithms are:

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

F. Association Rule Learning Algorithms

Association rule learning are methods that extract rules that best explain observed relationships between variables in data. These rules can discover important and commercially useful associations in large multidimensional datasets that can be exploited by the organization. The most popular association rule learning algorithms are:

- Apriori algorithm
- Eclat algorithm

G. Artificial Neural Network Algorithms

Artificial neural networks are models which uses supervised learning that are constructed based on the structure of biological neural networks. It has artificial neurons which has highly weighted interconnections among units and learns by tuning the connection weights to perform parallel distributed processing. Hence artificial neural networks are also called parallel distributed processing networks. The most popular artificial neural network algorithms are:

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

H. Deep Learning Algorithms

Deep Learning methods are a modern update to artificial neural networks that exploit abundant cheap computation. They are concerned with building much larger and more complex neural networks, as many methods are concerned with semi-supervised learning problems where large datasets contain very little labeled data. The most popular deep learning algorithms are:

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

I. Dimensionality Reduction Algorithms

Dimensionality reduction is an effective solution to the problem of curse of dimensionality. When the number of dimensions increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Dimensionality reduction is the study of methods for reducing the number of dimensions describing the object. Its general objectives are to remove irrelevant and redundant data to reduce the computational cost and to improve the

quality of data for efficient data organization strategies. Like clustering methods, dimensionality reduction seek and exploit the inherent structure in the data in an unsupervised manner. Many of these methods can be adapted for use in classification and regression. The dimensionality reduction algorithms are:

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)

J. Ensemble Algorithms

Ensemble methods are models based on unsupervised learning composed of multiple weak learner models that are independently trained and whose predictions are combined in some way to make the overall prediction. It divides the training data into number of subsets of data for which independent learning models are constructed..All learning models are combined to make correct hypothesis. This is a very powerful class of techniques and as such is very popular. The popular ensemble algorithms are

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

V. APPLICATION AND TOOLS

Machine learning applications are broadly classified based on learning techniques: supervised and unsupervised learning. Classification applications make use of supervised learning that are pattern recognition, face recognition, character recognition, medical diagnosis,web advertizing. Unsupervised learning applications are clustering, summarization, association analysis, customer segmentationin CRM, image ompression, bioinformatics. Robot control and game playing are the example application of reinforcement learning.

Tools are a big part of machine learning and it is needed to choose the right tool which is as important as working with the best algorithms. Machine learning tools make applied machine learning faster, easier and more fun. Machine learning learning tools provide capabilities to deliver results in a machine learning project. Also it is used as a filter to decide whether or not to learn a new tool or new feature. Machine learning tools provide an intuitive interface onto the sub-tasks of the applied machine learning process. There's a good mapping and suitability in the interface for the task. Great machine learning tools embody best practices for process, configuration and implementation. Examples include automatic configuration of machine learning algorithms and good process built into the structure of the tool. Machine learning tools are separated into platforms and libraries**.** A platform provides all capabilities needed to run a project, whereas a library only provides discrete capabilities or parts needed to complete a project. Examples of machine learning platforms are:

- WEKA Machine Learning Workbench.
- R Platform.
- Subset of the Python SciPy like Pandas and scikit-learn.

Library may provide a collection of modeling algorithms. Examples of machine learning libraries are:

- scikit-learn in Python.
- JSAT in Java.
- Accord Framework in .NET

VI. CONCLUSION

This paper provides a review of machine learning techniques. Machine learning model is given which describes the overview of machine learning process. Also describes the various machine learning algorithm based on types of machine learning styles. Examples of machine learning applications and need of tools are provided.

REFERENCES

[1] Pedro Domingos, Department of Computer Science and Engineering University of Washington Seattle, "A Few useeful Things to Know about Machin Learning".2012.
[2] Yogesh Singh, Pradeep Kumar Bhatia & Omprakash Sangwan "A review of studies in machine learning technique". International Journal of Computer Science and Security, Volume (1) : Issue (1) 70 – 84, June 2007
[3] Petersp,"The Need for Machine Learning is Everywhere" March 10, 2015.
[4] Jason Brownlee,"A Tour of Machine Learning algorithms" November 25, 2013
[5] Jason Brownlee,"Machine Learning Tools", December 28, 2015
[6] Taiwo Ayodele,"Types of Machine Learning algorithms", FEBRUARY 2010
[7] http://www.mlplatform.nl/what-is-machine-learning/