A Survey on Machine Learning: Concept, Categories and Algorithm

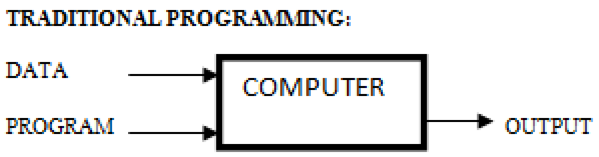
Karthiban.R[1], Jothipriya.P[2], Dharini.S[3]
[1]Assistant Professor, Dept of Computer Science and Engineering
[2, 3]Dept of Computer Science and Engineering
[1, 2, 3] Sri Shakthi Institute of Engineering and Technology, Coimbatore

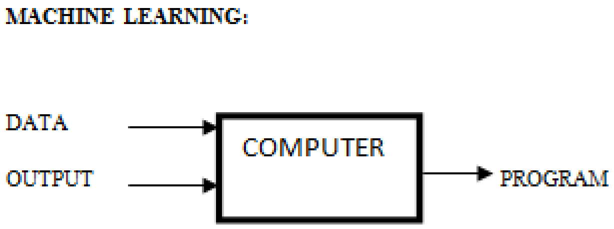
Abstract- *Machine learning is an application of artificial intelligence that improves computer systems to learn directly from examples, data, and experience. In machine learning the computer will initially performs the task of training set while studying. Then the computer will perform the same work with the data that is not available before. Through increasing the computers to perform some particular tasks intelligently, machine learning systems can carry out complex processes by learning from data sets. Here we are going to discuss about machine learning, types of machine learning and their algorithms. Machine learning contains the concept of automation. It has both training data set and testing data set. Machine learning always requires human guidance. Machine learning is a new technology that is rapidly growing in our society.*

Keywords- machine learning, supervised learning, unsupervised learning, reinforcement learning, algorithms.

I. INTRODUCTION

Machine learning is one of the fastest growing areas of computer science with far reaching applications. Learning defines the changes in the system that makes a system to do the similar task more efficiently and effectively the next time. Learning is constructing or modifying representations of what is being experienced. Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning refers to the automated detection of meaningful patterns in data. Machine learning focuses on the development of computer programs that can change when exposed to new data. Machine learning can be applied in many different fields. The goal of machine learning generally is to understand the structure of data and fitting that data into models that can be easily understood by user and the people. Even though machine learning is one of the field within computer science, it differs from traditional computing. In traditional computing the algorithms are explicitly programmed instructions to solve problems whereas in machine learning, it allows for computers to train on data inputs and uses statistical analysis for output values. The algorithm works by making data-driven decision through drawing a mathematical model from the given input data. Training data set is a set of example that are used for learning purpose. Testing data set tests how well our model has been trained. And it also identifies errors in our model. Test data mainly depends upon the size of our model.





II.TYPES OF MACHINE LEARNING

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**

Supervised learning: It is the Data mining task of inferring a function from labelled training data.A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.Training data includes inputs and desired outputs. e.g. True/False, Positive/Negative, Spam/Not Spam etc.

Algorithm Types

In the area of supervised learning which deals much with classification. These are the algorithms types:

- K-Means Clustering
- Decision Tree □ Random Forest
- Neural networks
- Bayesian Networks

1.Decision tree algorithm:

In machine learning domain the decision tree induction is currently one of the most important supervised learning algorithms. According to authors, the decision tree induction was initially designed to deal with single or multi-dimensional regression .It can handle both discrete variable and continuous variable.

ALGORITHM:

1.Check for base cases
2.For each attribute "a" calculate
i. Normalized the information gain(IG) from splitting on attribute "a"
3.Find the best "a" attribute that as highest IG
4.Create a decision node: node that splits on best of "a"
5.Recure on the sub-lists obtained by splitting on a best and add those nodes as children of node.

2.K-means clustering:

In this algorithm we have to first select the number of cluster in advance, they might converge to a local minimum. K-means can be seen as a specialization of the expectation maximization (EM) algorithm. It is more efficient (lower computational complexity) than hierarchical clustering. Continuous variable transforms symbolic attributes with the sequence of integers. Whereas discrete variable continuous attribute into symbolic values.

Algorithm:

```
K-means ((X= {d1, . . . ,dn} ⊆ Rm , k): 2R)
C: 2R /*µ a set of clusters */
d = Rm x Rm -> R /*distance function*/
µ: 2R -> R /* µ computes the mean of a cluster */
select C with k initial centers f1,….fk
while stopping criterion not true do
    for all clusters cj∈ C do
        cj ← {di|∀fld(di, fj) ≤d(di, fl)}
    done
    for all means fj do
        fj <- µ(cj)
    done
done
return C
```

3.Bayesian Network:

Bayesian Networks (BN) are graphical models that are used to illustrate relationships between events or ideas to infer probabilities or uncertainties associated with those ideas or events. Information retrieval, predictions based on limited input or recognition software is some main applications of BN. The Bayesian network structure S is a directed acyclic graph.

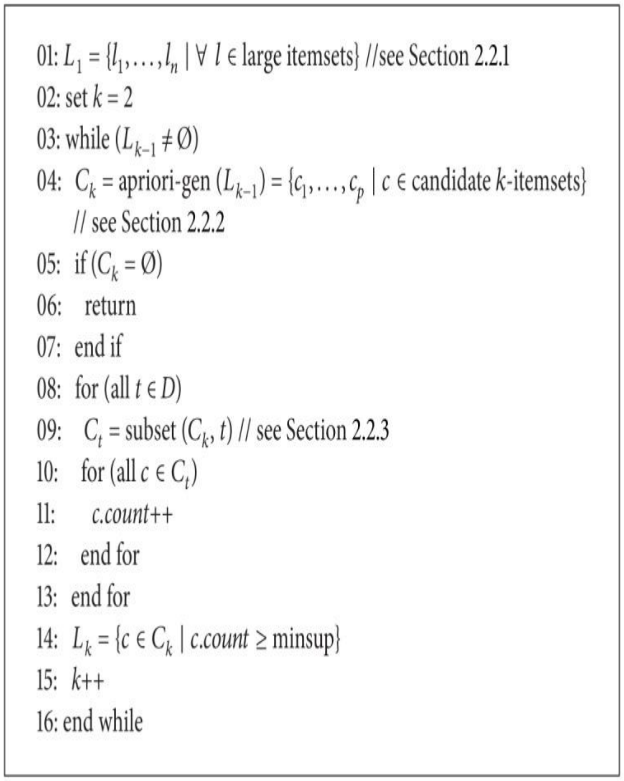
Unsupervised learning:

It is one of the type of machine learning algorithm. In this algorithm the data is unlabeled, so that the learning algorithm is left to find commodities among its input dada. Clustering analysis is one of the common method of unsupervised learning. Exploratory data analysis is used to find patterns that are hidden or data grouping. In Data mining, the issues in unsupervised learning is that of trying to find hidden structure in unlabeled. Some popular algorithm for unsupervised learning are

1.Apriori algorithm:

The Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules. Apriori algorithm is a discrete variable.

The Apriori algorithm can be divided into three steps. Algorithm 1 shows the pseudocode of the Apriori algorithm. The algorithm's first pass counts item occurrences to screen the large item sets . The second pass generates the candidate item sets Ck from large item sets Lk−1, using the apriori-gen function . Next, each transaction t checks whether the subsets of k-item sets of t belong to Ck, called subset function and described in Section 2.2.3. Finally, each c counts item occurrences in Ct, and c will be stored in L k if c. count minimum support. The algorithm terminates when L k is empty; that is, no frequent set of k or more items is present in D.



```
$L_1 = \{l_1, \ldots, l_n \mid \forall\, l \in$ large itemsets$\}$ //see Section 2.2.1
set $k = 2$
while ($L_{k-1} \neq \emptyset$)
  $C_k$ = apriori-gen ($L_{k-1}$) = $\{c_1, \ldots, c_p \mid c \in$ candidate $k$-itemsets$\}$
  // see Section 2.2.2
  if ($C_k = \emptyset$)
    return
  end if
  for (all $t \in D$)
    $C_t$ = subset ($C_k$, $t$) // see Section 2.2.3
    for (all $c \in C_t$)
      c.count++
    end for
  end for
  $L_k = \{c \in C_k \mid$ c.count $\geq$ minsup$\}$
  $k$++
end while
```

2.Anomaly detection:

Unsupervised anomaly detection techniques defect anomaies in an unlabeled test data set under the assumption that the maority of the instances in the data sets are normal by looking for instances that seem to fit least to the remainder of the data set.

3.Hierarchical clustering

Hierarchical clustering can give different partitionings depending on the level of resolution.It can be either continuous or discrete.

Divisive(top down)clustering): Starts with all data points in one cluster, the root, then Splits the root into a set of child clusters. Each child cluster is recursively divided further then stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point .

Agglomerative(bottom up) clustering: The dendrogram is built from the bottom level by merging the most similar (or nearest) pair of clusters and stopping when all the data points are merged into a single cluster (i.e., the root cluster).

Reinforcement learning:

Reinforcement Learning is a type of Machine Learning, and thereby also a branch of Artificial Intelligence. It allows machines and software agents to automatically determine the ideal behaviour within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behaviour; this is known as the reinforcement signal. Some of the reinforcement learning techniques are

1.Markov Decision Process:

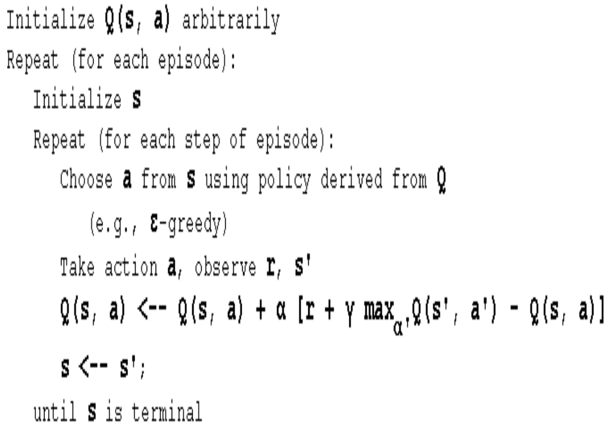
The mathematical framework for defining a solution in reinforcement learning scenario is called Markov Decision Process. This can be designed as:

- Set of states, S
- Set of actions, A
- Reward function, R
- Policy, π
- Value, V

We have to take an action (A) to transition from our start state to our end state (*S*). In return getting rewards (R) for each action we take. Our actions can lead to a positive reward or negative reward. The set of actions we took define our policy (π) and the rewards we get in return defines our value (V). Our task here is to maximize our rewards by choosing the correct policy. So we have to maximize $E(r_t \mid \pi, s_t)$ for all possible values of *S* for a time t.

2.Q-Learning:

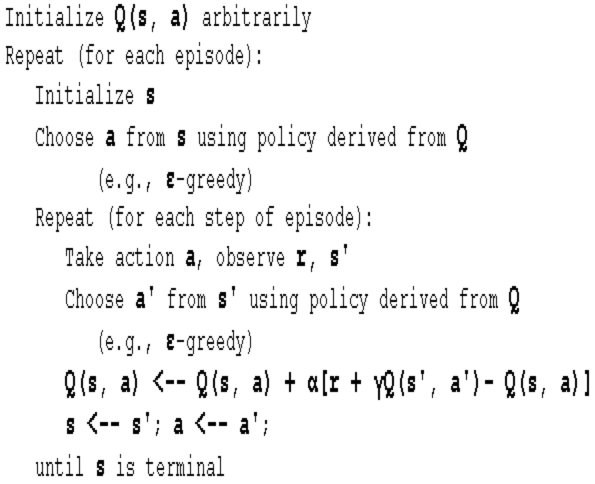
Q-Learning is an Off-Policy algorithm for Temporal Difference learning. It can be proven that given sufficient training under any ε-soft policy, the algorithm converges with probability 1 to a close approximation of the action-value function for an arbitrary target policy. Q-Learning learns the optimal policy even when actions are selected according to a more exploratory or even random policy. The procedural form of the algorithm is:



```
Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize $s$
  Repeat (for each step of episode):
    Choose $a$ from $s$ using policy derived from $Q$
      (e.g., $\varepsilon$-greedy)
    Take action $a$, observe $r$, $s'$
    $Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{\alpha'} Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$;
  until $s$ is terminal
```

3.Sarsa:

The Sarsa algorithm is an On-Policy algorithm for TD-Learning. The major difference between it and Q-Learning, is that the maximum reward for the next state is not necessarily used for updating the Q-values. Instead, a new action, and therefore reward, is selected using the same policy that determined the original action. The name Sarsa actually comes from the fact that the updates are done using the quintuple **Q(s, a, r, s', a').** Where: **s, a** are the original state and action, **r** is the reward observed in the following state and **s', a'** are the new state-action pair. The procedural form of Sarsa algorithm is comparable to that of Q-Learning:



```
Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize $s$
  Choose $a$ from $s$ using policy derived from $Q$
      (e.g., $\varepsilon$-greedy)
  Repeat (for each step of episode):
    Take action $a$, observe $r$, $s'$
    Choose $a'$ from $s'$ using policy derived from $Q$
      (e.g., $\varepsilon$-greedy)
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$; $a \leftarrow a'$;
  until $s$ is terminal
```

III. CONCLUSION

Nowadays machine learning techniques are being widely used to solve real world problems by storing, manipulating, extracting and retrieving of data from large sources. Machine learning approaches are steadily increasing search outputs. Machine learning plays a vital role in the fields like quality improvement and in the research fields. Machine learning is one of the popular field which is rapidly growing in our society. It has applications in nearly every other field of study and is already being implemented commercially because machine learning can solve problems that are too difficult or time consuming for humans to solve. Machine learning is a general technology that used in Google, Netflix and in many different fields. This technology is quite new and not very mature, so it contains many ethical problems. On the other hand, this technology ca helps us in dealing with security problem. This technology will influence social and economical structure. In future deep learning will become a main area in machine learning study.

REFERENCES

[1] Alpaydin, E. (2004). *Introduction to Machine Learning.* Massachusetts, USA: MIT Press.
Anderson, J. R. (1995). *Learning.*

[2] Cour, T. and Sapp, B. and Taskar, B. Learning from partial labels, Journal of Machine Learning Research, Volume 12, 1501-1536 2012.

[3] Ghahramani, Z. (2008). Unsupervised learning algorithms are designed to extract structure
from data. *178*, pp. 1-8. IOS Press.

[4] Gregory, P. A. and Gail, A. C. Self-supervised ARTMAP Neural Networks, Volume 23, 265-282, 2010.

[5] Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell and Weiru Liu, “Learning Bayesian networks from data: An information-Theory based approach”, *The Artificial Intelligence Journal*, Vol. 137, pp. 43-90, 2002.

[6] Kotsiantis.S.B, “Supervised Machine Learning: A Review of Classification Techniques”, *Informatica*, Vol. 31, No. 3, pp. 249-268, 2007.

[7] Marc G Bellemare, Will Dabney, and R´emiMunos. A Distributional

[8] MehryarMohri, AfshinRostamizadeh and Ameet Talwalkar, “*Foundations of Machine Learning*”, One Rogers Street Cambridge MA: The MIT Press, 2012.

[9] Mitchell, T. M. (2006). *The Discipline of Machine Learning.* Machine Learning Departmenttechnical report CMU-ML-06-108, Carnegie Mellon University.

[10] Mitchell.T.M, Machine Learning, McGraw-Hill International, 1997.

[11] Mitchel.T.M, The Discipline of Machine Learning, CMU-ML-06-108, 2006.

[12] Pierre Geurts, Alexandre Irrthum, Louis Wehenkel, “Supervised learning with decision tree-based methods in computational and systems biology”, *Molecular BioSystems*, Vol. 5, No. 12, pp. 1593-1605, 2009.

[13] Pieter Abbeel and John Schulman. Deep Reinforcement Learningthrough Policy Optimization, 2016. Tutorial at NIPS 2016.

[14] TaiwoOladipupoAyodele, Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), InTech, 2010

[15] Tom M. Mitchell, "*Machine Learning: A Guide to Current Research*", The Springer International Series in Engineering and Computer Science Series, McGraw Hill, 1997.

[16] Tom, M. (1997). *Machibe Learning.* Machine Learning, Tom Mitchell, McGraw Hill, 1997:McGraw Hill.

.