Sapinderjit Kaur, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 12, Issue 3, (Series-I) March 2022, pp. 33-39

RESEARCH ARTICLE OPEN ACCESS

An Overview and Comparison of Machine Learning Techniques

Sapinderjit Kaur[1], Navpreet Rupal[2], Jagpreet Singh[3]
[1,2,3]*Department of Computer Science and Engineering, Global Group of Institutes, Amritsar*

ABSTRACT:
Machine learning is a field of computer engineering that gives ability to PCs to learn without being unequivocally programmed. Machine learning is used in an arrangement of computational tasks where arranging and programming unequivocal calculations with high performance isn't straightforward. One of the principle goals of machine learning is to get ready PCs to utilize data to deal with a predefined issue. This paper provides an extensive review of studies related to expert estimation of software development using Machine-Learning Techniques (MLT). Machine learning system effectively "learns" how to estimate from training set of completed projects. This paper presents the most commonly used machine learning techniques such as neural networks, case based reasoning, classification and regression trees, rule induction, genetic algorithm & genetic programming for expert estimation in the field of software development.

Keywords: *Machine Learning Techniques (MLT), Neural Networks (NN), Case Based Reasoning (CBR), Classification and Regression Trees (CART), Rule Induction, Genetic Algorithms and Genetic Programming.*

Date of Submission: 02-03-2022 Date of Acceptance: 16-03-2022

I. INTRODUCTION

Machine Learning is the study of computational techniques for improving performance by mechanizing the acquisition of knowledge from experience.Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the exact information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data[2].

Expert performance requires much domain specific knowledge, and knowledge engineering has produced several AI expert systems which are now used regularly in industry. The aim of machine learning is to provide increasing levels of automation in the knowledge engineering process, and to replace time- consuming human activities with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data. The ultimate test of machine learning is its ability to produce systems that are used regularly in industry, education, and elsewhere. Most evaluation in machine learning is experimental in nature, aimed at showing that the learning method leads to performance on a separate test set, in one or more realistic domains, that is better than performance on that test set without learning.

Basically, there are two types of machine learning: inductive, and deductive. **Deductive learning** works on existing facts and knowledge and deduces new knowledge from the already existing knowledge. Inductive machine learning aims at creating computer programs by extracting rules and patterns out of massive data sets. **Inductive learning** takes examples and thengeneralizes rather than starting with already existing knowledge. Concept learning is a subclass of inductive learning which takes examples of a concept and tries to build a general description of the concept. Very often, the examples are described using attribute-value pairs.

In our study we concentrate on the various approaches, which are used in machine learning. This review paper also examines the comparative study of machine learning technique with suitable application area.

This paper is organized as follows: Section 2 discusses about the use of Neural Network in machine learning. CBR with application area is presented in section 3. CART is another efficient learning method described in section 4. Another machine learning approach rule induction is discussed in section 5. In section 6 the impact of genetic algorithm and programming are discussed. Section 7 presents the discussion on various machine-learning tech-

niques and conclusions and future direction are presented in section 8.

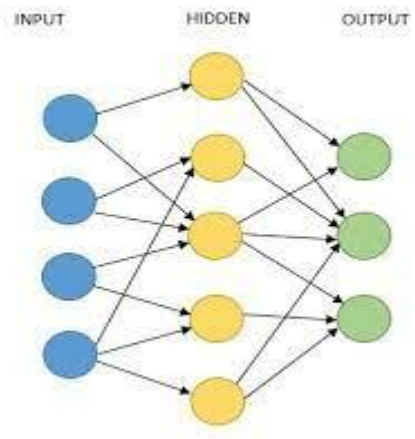
Supervised Learning Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples. The supervised machine learning algorithms are those algorithms which needs external assistance. The input dataset is divided into train and test dataset. The train dataset has output variable which needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification.

1. **Decision Tree** Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. Each tree consists of nodes and branches. Each node represents attributes in a group that is to be classified and each branch represents a value that the node can take.

II. NEURAL NETWORKS

Neural networks have been established to be an efficient tool for pattern classification and clustering [8, 15]. There are broadly two approaches of neural learning algorithms named supervised and unsupervised. Unsupervised neural algorithms are best suited for clustering patterns on the basis of their inherent characteristics [8,14]. There are three major approaches for unsupervised learning: -

(a) Competitive Learning
(b) Self Organising feature Maps
(c) ART Networks



The other approach of neural learning is the supervised learning paradigm. These networks have been established to be universal approximators of continuous/discontinuous functions and therefore they are suitable for usage where we have some information about the input-output map to be approximated. A set of data (Input-Output information) is used for training the network. Once the network has been trained it can be given any input (from the input space of the map to be approximated) and it will produce an output, which will correspond to the expected output from the approximated mapping. The quality of this output has been established to correspond arbitrarily close to the actual output desired owing to the generalization capabilities of these networks.

III. CASE BASED REASONING(CBR)

Case Based Reasoning is a technique withwhich we solve new problems by adapting the solutions from similarly solved problems. We take the instances of solutions from problems that occurred in the past and try to solve new problems by using these cases. Each such solution available to us can be termed as a case.

3.1 CBR Process

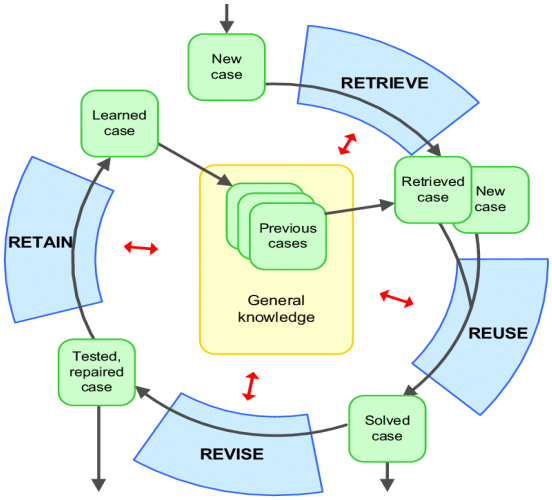
A general CBR process includes the following four processes.

A new case is defined by the initial description of any problem. This new case is *retrieved* from a collection of previous cases and this retrieved case is then combined with the new case through reuse into a *solved case*. This solved case is nothing but a proposed solution to the defined problem. Once this solution is identified, applying it practically to the real world tests it. This process of testing is termed as *revision* of the problem. Then comes the process of *retain* where useful experience is retained for future reuse and the case base is updated by a new

learned case or by modification of some existing cases.

Thus we can say that CBR is a four-step process:

1. RETRIEVE
2. REUSE
3. REVISE
4. RETAIN



3.2 Fundamentals of Case Based Reasoning

3.2.1 Case Retrieval

The process of retrieval in CBR cycle begins with the problem description and ends when the best possible case from the set of previous cases has been obtained. The subtasks involved in this particular step include identifying features, matching, searching and selecting the appropriate ones executed in that order. The identification task finds a set of relevant problem descriptors, then the matching task returns those cases that are similar to the new case and finally the selection task chooses the best possible match. Among well-known methods for case retrieval are: nearest neighbour, induction, knowledge guided induction and template retrieval.

3.2.2 Case Reuse

This involves obtaining the solved case from a retrieved case. It analyses the differences between the new case and the past cases and then determines what part of the retrieved case can be transferred to the new case. CBR is essentially based on the concept of analogy wherein by analyzing the previous cases we formulate a solution for the new cases [5].

3.2.3 Copy

In the trivial cases of reuse we generally copy the solution of the previous cases and make it the solution for the new cases. But many systems take into consideration the differences between the two cases and use the adaptation process to formulate a new solution based on these differences.

3.2.4 Adaptation

The adaptation process is of two kinds: *Structural adaptation*- Adaptation rules are applied directly to the solution stored in cases i.e. reuse past case solution.

Derivational adaptation- Reuse the method that constructed the solution to a past problem. In structural adaptation we do not use the past solution directly but apply some transformation parameters to construct the solution for the new case. Thus this kind of adaptation is also referred to as transformational adaptation. In derivational adaptation we use the method or algorithm applied previously to solve the new problem [17].

3.2.5 Case Revision

After reusing the past cases to obtain a solution for the new case we need to test that solution. We must check or test to see if the solution is correct. If the testing is successful then we retain the solution, otherwise we must revise the case solution using domain specific knowledge.

3.2.6 Case Retainment- Learning (CRL)

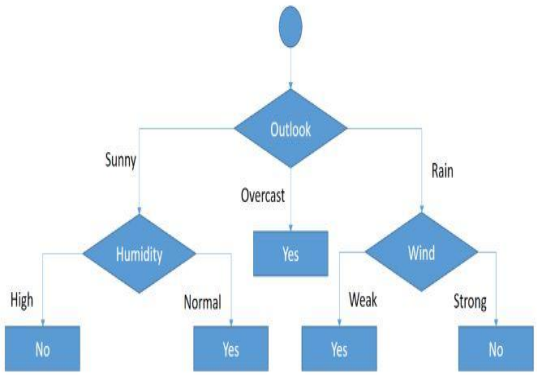
The solution of the new problem after being tested and repaired may be retained into the existing domain specific knowledge. This process is called Case Retainment Learning or CRL. Retaining information involves selecting what information to retain, in what form to retain it, how to index the case for later retrieval from similar problems, and how to integrate the new case in the memory structure.

3.2.7 Case Based Learning

An important feature of CBR is its coupling to learning [2]. Case-based reasoning is also regarded a sub-field of machine learning. Thus, the notion of case-based reasoning does not only denote a particular reasoning method, irrespective of how the cases are acquired, it also denotes a machine learning paradigm that enables sustained learning by updating the case base after a problem has been solved. Learning in CBR occurs as a natural by-product of problem solving. When a problem is successfully solved, the experience is retained in order to solve similar problems in the future. When an attempt to solve a problem fails, the reason for the failure is identified and remembered in order to avoid the same mistake in the future. CBR can be applied to solve real world problems for instance handling of multiple disorders [16] or for engineering sales support [23].

IV. CLASSIFICATION AND REGRESSION TREES (CART)

CART is a very efficient machine learning technique. The difference between this technique and other machine learning technique is that CART requires very little input from the analyst. This is in contrast to other technique where extensive input from the analyst, the analysis of interim results and modification of method used is needed. Before going into the details of CART we identify the three classes and two kinds of variables, which are important while defining classification and regression problems.



There are 2 main kinds of variables: 1) **Continuous variables** -- A continuous variable has numeric values such as 1, 2, 3.14, -5, etc. The relative magnitude of the values is significant (e.g., a value of 2 indicates twice the magnitude of 1). Examples of continuous variables are blood pressure, height, weight, income, age, and probability of illness. Some programs call continuous variables "ordered" or "monotonic" variables.

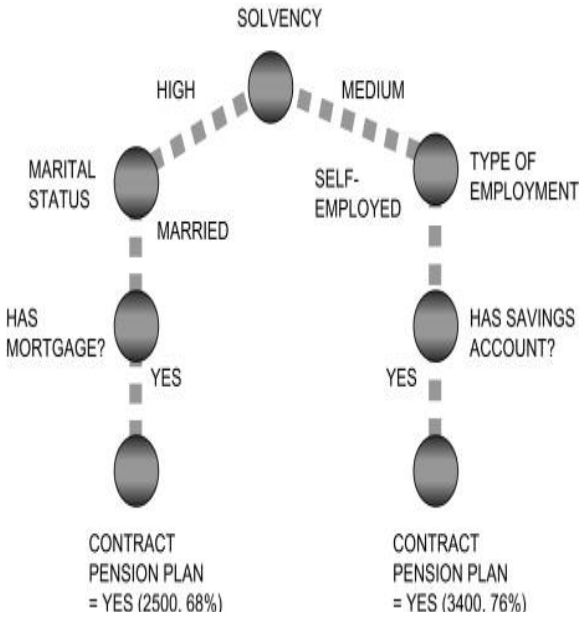
2) **Categorical variables** -- A categorical variable has values that function as labels rather than as numbers. Some programs call categorical variables "nominal" variables. For example, a categorical variable for gender might use the value 1 for male and 2 for female. The actual magnitude of the value is not significant; coding male as 7 and female as 3 would work just as well
CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification).
Regression-type problems: These are generally those where one attempts to predict the values of a continuous variable from one or more continuous and/or categorical predictor variables. Classification-type problems: These are generally those where one attempts to predict values of a categorical dependent variable from one or more continuous and/or categorical predictor variables.

CART is a non-parametric statistical methodology developed for analyzing classification issues either from categorical or continuous dependent variables [24, 25]. If the dependent variable is categorical, CART produces a classification tree. When the dependent variable is continuous, it produces a regression tree.

CART is basically used to avoid the disadvantage of the regression techniques. CART analysis is a form of binary recursive partitioning [20]. The term "binary" implies that each node in a decision tree can only be split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term "partitioning" refers to the fact that the dataset is split into sections or partitioned.

V. RULE INDUCTION

Rule Induction is another very important machine learning method and it is easier because the rules in rule induction are transparent and easy to interpret than a regression model or a trained neural network. This paradigm employs condition-action rules, decision trees, or similar knowledge structures. Here the performance element sorts instances down the branches of the decision tree or finds the first rule whose conditions match the instance, typically using an all-or- none match process [19]. Information about classes or predictions is stored in the action sides of the rules or the leaves of the tree. Learning algorithms in the rule induction framework usually carry out a greedy search through the space of decision trees or rule sets, typically using a statistical evaluation function to select attributes for incorporation into the knowledge structure. Most methods partition the training data recursively into disjoint sets, attempting to summarize each set as a conjunction of logical conditions.

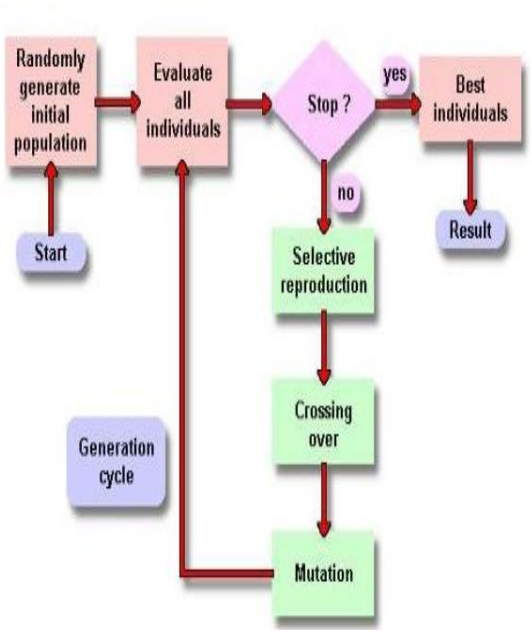


VI. GENETIC ALGORITHMS AND GENETIC PROGRAMMING

The genetic approach to machine learning is a relatively new concept. Both genetic algorithms and Genetic Programming (GP) are a form of evolutionary computing which is a collective name for problem solving techniques based on the principles of biological evolution like natural selection. Genetic algorithms use a vocabulary borrowed from natural genetics in that they talk about *genes* (or bits), chromosomes (individuals or bit strings), and population (of individuals) [10]. Genetic algorithm approach is centered around three main processes- crossovers, mutation and selection of individuals. Initially many individual solutions are gathered together to make a randomly generated population. Genetic algorithms are based upon the Darwin theory of " The survival of the Fittest" depending upon the fitness function the best possible solutions are selected from the pool of individuals. The fitter individuals have greater chances of its selection and higher the probability that its genetic information will be passed over to future generations. Once selection is over new individuals have to be formed. These new individuals are formed either through crossover or mutation. In the process of crossover, combining the genetic make up of two solution candidates (producing a child out of two parents) creates new individuals. Whereas in mutation, we alter some individuals, which means that some randomly chosen parts of genetic information is changed to obtain a new individual. The process of generation doesn't stop until one of the conditions like minimum criteria is met or the desired fitness level is attained or a specified number of generations are reached or any combination of the above [21].

John Koza popularized GP, an offset of Genetic Algorithm in 1992. It aims at optimizing computer programs rather than function parameters. GP is a supervised machine learning technique where algorithms are modeled after natural selection. These algorithms are represented as function trees where these trees are intended toperform a given task [6]. In GP the fitter individuals are retained and allowed to develop whereas others are discarded [4].

GP works in a manner similar to genetic algorithm. It also follows the principles of natural evolution to generate a solution that maximizes (or minimizes) some fitness function [3]. GP differs from GA in the sense that GP tends to find the solution of a given problem by representing it as a array of integers while the goal of a GP process is to produce a computer program to solve the optimization problem at hand. GP cycle works as any evolutionary process. New individuals are created; tested and fitter ones succeed in creating their own children. The unfit individuals are removed from the population. The figure:6 illustrates how GP cycle works.



Comparison of Various Machine Learning Techniques			
Technique	**Area of Application**	**Advantages**	**Limitations**
Neural Networks (NN)	Testing Effort Estimation Function Point Analysis Risk Management Reliability Metrics Sales Forecasting	Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time.	Minimizing over fitting requires a great deal of computational effort. The individual relations between the input variables and the output variables are not developed by engineering judgment so that the model tends to be a black box or input/output table without analytical basis.
Case Based Reasoning (CBR)	Help-Desk Systems Software Effort Estimation Classification and Prediction	No Expert is Required CBR can handle failed cases (i.e. those cases for which accurate prediction cannot be made)	Case data can be hard to gather. Predictions are limited to the cases that have been observed.
Classification and Regression Trees (CART)	Financial applications like Customer Relationship Management(CRM)	It is inherently non-parametric in other words no assumptions are made regarding the underlying distribution of values of the predictor variables. CART identifies splitting variables based on an exhaustive search of all possibilities. It has methods for dealing with missing variables.	Scientific research (Biological Evolution) Computer Games purposes CART splits only by one variable.
Rule Induction	Making Credit Decisions (in various loan companies) Diagnosis of Mechanical Devices Classification of Celestial Objects Preventing breakdowns in transformers	Simplicity of input variables. The representation in rule-based technique is easier to depict and understand.	No sufficient background knowledge is available. It is deduced from examples. The representation in rule-based technique is easier to depict and understand.
Genetic Algorithms (GA) and Genetic Programming (GP)	Optimization Simulation of economic processes Scientific research (Biological Evolution) Computer Games purposes	GA and GP techniques can be applied to a variety of problems. GP is based on the 'Survival of the Fittest Scheme' allowing fitter individuals to develop and discarding unfit ones.	Resource requirements are large. It can be a time consuming process. GA practitioners often run many copies of the same code with the same inputs to get statistically reliable results.

VII. CONCLUSIONS AND FUTURE DIRECTIONS

The main contribution of this review is to discuss the various Machine-Learning Techniques employed in effort estimation, cost estimation, size estimation and other field of Software Engineering. The paper also gives a relative comparison of all the techniques based on their applications, advantages and limitations. After analysis of all the techniques, we cannot state as any one technique being the best. Each technique has different application areas and is useful in different domains based on its advantages. Thus, keeping in mind the limitations of each of the techniques and also the prime focus being the improvement in performance and efficiency we should use that technique, which best suits a particular application. For instance GA and GP prove to be useful in the area of scientific research involving biological evolution whereas rule based techniques and CART analysis may be useful in many financial applications. Similarly CBR is being developed for use in Help- Desk Systems, a relatively new application and NN may be employed for Risk Management or Sales Forecasting.

Our study also encourages that no one technique can be classified as being the perfect machine learning technique. For this reason there is a strong need for better insight into the validity and generality of many of the discussed techniques. In particular we plan to continue with research on: -

When to use machine-learning techniques and estmationmodels.
How to select and combine a set of test cases for effective estimation technique & to get better results?

REFERENCES:

[1]. Aggarwal, K. K., et al. "A neural net based approach to test oracle." *ACM SIGSOFT Software Engineering Notes* 29.3 (2004): 1-6.

[2]. Mahesh, Batta. "Machine Learning Algorithms-A Review." *International Journal of Science and Research (IJSR).[Internet]* 9 (2020): 381-386.

[3]. Ray, Susmita. "A quick review of machine learning algorithms." *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, 2019.

[4]. Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma. "A review of supervised machine learning algorithms." *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee, 2016.

[5]. Singh, Yogesh, Pradeep Kumar Bhatia, and OmprakashSangwan. "A review of studies on machine learning techniques." *International Journal of Computer Science and Security* 1.1 (2007): 70-84.

[6]. Khan, Aurangzeb, et al. "A review of machine learning algorithms for text-documents classification." *Journal of advances in information technology* 1.1 (2010): 4-20.

[7]. Cho, Gyeongcheol, et al. "Review of machine learning algorithms for diagnosing mental illness." *Psychiatry investigation* 16.4 (2019): 262.

[8]. Shinde, Pramila P., and Seema Shah. "A review of machine learning and deep learning applications." *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018.

[9]. Angra, Sheena, and Sachin Ahuja. "Machine learning and its applications: A review." *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, 2017.

[10]. Saranya, T., et al. "Performance analysis of machine learning algorithms in intrusion detection system: A review." *Procedia Computer Science* 171 (2020): 1251-1260.

[11]. Praveena, M., and V. Jaiganesh. "A literature review on supervised machine learning algorithms and boosting process." *International Journal of Computer Applications* 169.8 (2017): 32-35.

[12]. Gamal, Donia, et al. "Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains." *Machine Learning and Knowledge Extraction* 1.1 (2019): 224-234.