International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Survey on Machine Learning: Concept, Algorithms, and Applications

Sakshini Hangloo[1], Samreen Kour[2], Sudesh Kumar[3]
[1,2]M.Tech, Department of computer science, Shri Mata Vaishno Devi University, J&K, India
[3]PhD Scholar, Department of computer science, Shri Mata Vaishno Devi University, J&K, India

ABSTRACT

In today's era machine learning concepts and algorithms are heavily used in the digital world. Machine learning algorithms can easily understand how to perform important tasks by generalizing from examples. Machine learning is often feasible and cost-effective approach where manual programming is not. From the past few decagons, Machine learning (ML) made software application more accurate to predict outputs. Also, various algorithms that are designed in machine learning are continuously used for pattern recognition, data clarification, and various other plans and have lead to a distinct research in data mining to determine underground consistencies or inconsistencies in collective data. The main objective of this paper is to discuss various concepts, approaches and procedures of machine learning used in addressing the digital world problems.
Keywords: Machine Learning, Precision, Training data, Procedures

I. INTRODUCTION

Machine learning is an area of computer science that is interrelated with designing the systems in a manner that the system itself learn and upgrade with experience. Learning means identification and interpretation of input data and decision-making based on the provided data.

The programmer sometimes has a certain motive in mind while framing a machine (a software system). Consider the case of Strike Series of Robert Galbraith and Potter Series of J.K. Rowling, two skilled persons were indulged from "The London Sunday Times" and use "Forensic Machine Learning" and they confirm that the Rowling actually wrote the books by the name "Galbraith". They both program a machine learning algorithm and "edify" it with Rowling's and other examples by writers to find out and understand the hidden patterns and "confirm" by Galbraith book. Rowling's and Galbraith's writing matched the most in various features is ended up by this algorithm.

By Machine Learning's use, a researcher trys to obtain an perspective through which the machine, i.e., the algorithm will come up with its own solution based on the example or training data set given to it initially instead of developing an algorithm to mark the problem directly.

A. MACHINE LEARNING: INTERSECTION OF STATISTICS AND COMPUTER SCIENCE

Machine Learning shows a remarkable outgrowth when Computer Science and Statistics both forces are joined together. Computer Science aims at building machines that find some answer to certain problems, and tries to identify if problems are interpretable. The main approaches on which Statistics

fundamentally focuses are data inference, modelling hypothesises and measuring quality of the outcomes.

Machine Learning idea is a slightly different but somehow dependent on both. Computer Science focuses on blue-collar programming, while ML marks the issue of getting computers to re-program themselves whenever revealed to new data rooted on some starting learning plans that are given. On the another hand, Statistics aims at probability and data inference, while Machine Learning involves the practicality and efficacy of construction and algorithms to operate those data, constituting certain learning functions and performance measures.

B. HUMAN LEARNING AND MACHINE LEARNING

Machine Learning is the learning of human and animal brain in the technical learning of the nervous system, Psyche and associated domains and a third study area that is closely concerned with it. The researchers suggested that how a machine could understand from experience is not different than how an animal or a human mind understands. However, the researches correlated with statistical - computational approach is far better than the research focusing on exploring machine learning problems using human brain's training methods which did not produce a much optimistic outcome. This might be happen due to the reality that human or animal psyche remains not completely perceivable to date. Association between human study and machine study is growing, nevertheless of these difficulties and for machine learning is used to describe certain understanding techniques present in human or animals. For example, to describe neural signals in humans study, machine study function of time-related difference was presented. It is equitably anticipated that in coming years, this association is to grow greatly.

C. DATA MINING, ARTIFICIAL INTELLIGENCE, AND MACHINE LEARNING

The three approaches Machine learning, Artificial Intelligence, and Data Mining are concerned with each other and together can produce highly effective and responsive outcomes. Data mining places the initial point for both machine learning and artificial intelligence and is primarily about explicating the data. In action, it examines and identifies patterns and correlations that occur in information which is tough to explain manually. Therefore, data mining is not a basic function to show a presumption but function for producing appropriate presumption. Artificial intelligence may be explained as machines that have the capability to resolve a given situation on their own without any human involvement. The important data and the AI illuminating the data produce an answer by itself rather than answer developed straight into the system. The illumination that goes under is nothing but a data mining algorithm. Machine learning promotes the approach to a greater level by giving the data important for a machine to train and modify accordingly when revealed to new data. This is called as "training". It centres on exploring information from large sets of data, and then discovers and determines underneath patterns using various statistical measures to increase its capability to explain new data and gives more efficient outcomes. At last, if a system lacks the ability to learn and improve from its previous exposures then that system cannot examine to be completely intelligent.

II. PRESENT RESEARCH QUESTION & RELATED WORK

The various applications declared above suggest somehow advancement in machine learning and their fundamental theory. Machine learning is a deep learning and various researchers have suggested their views in this field. In this paper, the major research question that is being taken at present are explained and it gives the references to some of recent work.

A. BY THE USE OF UNLABELED DATA IN SUPERVISED LEARNING

For supervised learning, labelled data are necessary. Most of the time, they are often obtainable in less volume, while unlabeled data may be huge [7]. Combining both unlabeled data and labelled data is of great interest, both in a theoretical and a practical sense. In recent time, for joining unlabeled and labelled data various approaches have been recommended. Supervised learning algorithms check the closeness of relation between the features and labels. The problem with this approach is that the predefined information is not consistently provided [8]. Before going for supervised classification, we process, filter and label the information using unsupervised learning, there by adding to the total cost. Supervised learning problems often have the following property: class labels have a high cost while unlabeled examples have little or no cost. This rise in cost can be reduced greatly if unlabelled data is used by supervised learning (e.g., images). In various interesting cases of learning problems with extra presumption, unlabeled data can be truly validated to upgrade the expected perfection of supervised learning [21]. For example, identifying spam emails or classifying web pages. Presently active researchers are seriously taking into consideration the new algorithms or new learning problems for making the use of unlabeled data effectively.

B. TRANSFERRING THE LEARNING EXPERIENCE

To ease the learning process to carry out a new task, transfer learning uses the facts from past similar tasks which are its main goal. The advantage of transfer leaning is mainly regulated by a minimization in the number of training examples. To minimize complexity of samples, training examples need to have a target performance on a sequence of similar learning problems, matched with number required for unfamiliar problems. In various real-life cases, only a few new concepts, training examples or process are often enough for a human learner to grab the new concept and define it. For example, learning for driving a bus becomes very much easier task if we have the knowledge of how to drive a car. For various problems of real life, the supervised algorithm may include learning a group of associated functions than just learning a single one. Even if the scientific determination functions for distinct cities (e.g., Jammu and Canada) are assumed to be comparatively different, some similarities are sure as well.

C. LINKING DIFFERENT ML ALGORITHMS

In various fields or domains, the number of machine learning algorithms have been brought into notice and tested. Among the existing ML algorithms, one experiment of research aims to detect all the possible correlations and suitable cases to use a particular algorithm [23]. Now consider two supervised classification algorithms, Naive Bayes' and Logistic Regression. They both differently tend to various data sets, but when implemented to specific types of training data, their significance can be explained. In general, the ML algorithms conceptual understanding, their combined features, and their respective efficiency and limitations to date will remain an intrinsic research matter.

D. BEST STRATEGICAL APPROACH FOR LEARNERS WHO COLLECT THEIR OWN DATA

A wider research direction centres on learning systems that energetically assembles data for its own processing and learning instead of mechanically using data assembled by some other plans. Most of the research time is given in exploring the powerful scheme to fully pass over the power to the learning algorithm. For example, to check the behaviour of a patient by considering a drug test system to learn the outcome of all possible hidden side effects and trying to reduce them.

E. PRIVACY-PRESERVING DATA MINING

For automatically and intelligently withdrawing information or knowledge from a huge number of data, which can also reveal delicate information about particular understanding the particular's right to privacy, Data mining is a popular technique. Moreover, critical information about business transactions, compromising the free competition in a business setting can be revealed by data mining techniques. Therefore, for this reason, privacy-preserving data mining (PPDM) has become a major field of study. In data mining, PPDM becomes a fresh research area, where data mining algorithms are for possible violation of privacy. PPDM research generally works on three philosophical approaches:(1) data hiding, where delicate raw data like identifiers, name, addresses, etc. are altered, blocked, or trimmed from the original database, in order for the users of the data not to compromise with another person's privacy; (2) rule hiding, where delicate knowledge explored from the data mining process keeps out for use, because private information may be extracted from the disclosed knowledge; (3) secure multiparty computation, in which distributed data is released or shared for computations, but before that it is encrypted; thus, everyone knows about its own inputs and the results but not everything. The PPDM goal is to develop effective algorithms that allow exploring relevant knowledge from a huge number of data, while preventing the delicate data and information from the broadcast.

III. MACHINE LEARNING ALGORITHMS CATEGORIZATION

Over past years a number of ML algorithms have been designed and introduced. These algorithms are broadly grouped into two categories on the basis of learning style and similarity. In this section, we will inculcate some basic idea of various types of ML algorithms.

A. GROUP BY LEARNING STYLE

1. Supervised learning

In supervised learning, the machine is provided with a given set of inputs or training data with their desired outputs or predetermined labels e.g. True/False, Positive/Negative etc. The machine needs to study those given sets of inputs and outputs, and find a general function that could predict the label of test data. The supervised learning can be of regression or classification type.

2. Unsupervised learning

Unlike in supervised learning, here the Input data or training data is not labelled which makes this type of learning harder. One of the approaches is clustering where the training data is grouped on the basis of similarity.

3. Semi-supervised learning

In semi-supervised learning, the training data contains both labelled and unlabeled data. The aim is to develop an algorithm that will predict classes of future test data better than the earlier algorithm that used only the labelled data. The way humans learn is similar to semi-supervised learning.

4. Reinforcement learning

In this type of learning, algorithm maps action to the situation and receives reward or penalty for its actions in trying to solve a problem. After several trial and error runs it learns the best policy i.e. the sequence of actions that maximize the total reward.

B. ALGORITHMS GROUPED BY SIMILARITY

1. Instance-based Algorithms

The Instance-based model simply stores instances of training data instead of developing a definition of target function. Each time when a new problem arises it is compared with the previously stored data in order to predict and determine the value of target function. This is done by assign the value of a target

function to the new instance, provided that is a better fit than the former and hence these algorithms are also known as winner-take-all method. Examples of Instance-based Algorithm are K-Nearest Neighbour (KNN), Locally Weighted Learning (LWL), Learning Vector Quantisation (LVQ), Self-Organising Map (SOM), etc.

2. Genetic algorithm

The Genetic algorithm provides a learning method that is similar to biological evolution. Instead of search from general-to-specific hypotheses, GA generates successor hypotheses by repeatedly mutating and crossover of the best currently known hypotheses to generate new genotype in the hope of finding good solutions to a given problem.

3. Decision Tree Algorithms

Decision tree algorithms, one of the most widely used methods for inductive inference. It is a type of supervised learning. A decision tree is a tree-like structure consisting of all possible solutions to a problem based on certain constraints. It begins with a single simple decision or root, which then extends to a various branches until a decision is made, forming a tree and hence named as decision tree. Some of its examples are Classification and Regression Tree (CART), Conditional Decision Trees, Chi-squared Automatic Interaction Detection (CHAID), etc.

4. Bayesian Algorithms

Bayesian algorithms use Bayes' Theorem to solve classification and regression kind of problems. Bayesian offers a possible outlook for logic estimation. It is based on the impression that the quantity of interest is governed by distribution of probability and that optimal decisions can be made by reasoning about these probabilities along with the observed data. Some of the examples of Bayesian algorithm include Naive Bayes, Bayesian Network (BN), Gaussian Naive Bayes, Bayesian Belief Network (BBN), etc.

5. Support Vector Machine (SVM)

In SVM, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. It uses a separating hyperplane among a set of data points which splits the data into two differently classified groups. SVM is a supervised classification method and can perform both linear and nonlinear classification.

6. Association Rule Learning Algorithms

Association rules aim to discover a relationship between various variables in a huge database. They are widely used in many applications areas like Market Basket analysis, intrusion detection, bioinformatics etc. Common examples are Apriori algorithm, FP Growth algorithm, Éclat algorithm etc.

7. Artificial Neural Network (ANN) Algorithms

ANN is a model based on the structure and operations of actual neural networks of the living being. ANNs are regarded as non-linear models. ANN discovers complex associations between input and output data by selecting sample from data rather than considering the entire data set and thereby reducing cost and time. Examples: Back-Propagation learning, Hop-field Network, Perceptron etc.

8. Deep Learning Algorithms

Deep learning algorithm consists of multiple hidden layers in an artificial neural network. This approach tries to work in the same way as the human brain processes light into vision and sound into hearing. They produce results comparable to human experts and in some cases the results are even better.

Some applications of deep learning are computer vision and speech recognition [1]. Examples of deep learning algorithms are Deep Boltzmann Machine (DBM), Stacked Auto-Encoders Deep Belief Networks (DBN) etc.

9. Dimensionality Reduction Algorithms

Dimensionality simply refers to the number of features or input variables in the dataset. When the number of features is very large, *certain* algorithms struggle to train models effectively, this is called the

Curse of Dimensionality. Dimensionality reduction visualises data with numerous features and helps in implementing supervised classification more efficiently [2]. Examples: Principal Component Analysis (PCA), Multidimensional Scaling (MDS) Principal Component Regression (PCR), Discriminate Analysis (LDA), Partial Least Squares Regression (PLSR), Summon Mapping, etc.

10. Clustering Algorithms

Clustering algorithm divides the population into a number of groups such that the data points in one group are more similar to the other data points in the same group than those which belong to some other groups. In simple words, clustering is concerned with using inherent patterns in the datasets to classify and label the data accordingly [2]. Some of the examples include K-Means, K-Medians, Ward hierarchical clustering, and Mean Shift, Expectation Maximisation (EM) etc.

11. Regression Algorithms

Regression analysis is subset of predictive analytics and visualises the co-relation between dependent and independent variables. Examples of regression models are: Linear Regression, Logistic Regression, and Stepwise Regression, Multivariate Adaptive Regression Spines (MARS) etc.

IV. APPLICATIONS

Machine learning has proved to be the answer to many real-world challenges. In this section, we will discuss some applications of machine learning with some examples. But still, there are a number of problems for which machine learning needs a breakthrough.

1. SPEECH RECOGNITION

In the field of speech recognition, certain methodologies are developed that enable computers to recognize and then translate the spoken language into text. All these systems use machine learning approach for better accuracy. There are various voice-controlled programs such as Apple's Siri, Google Now, Amazon's Alexa, Microsoft's Cortana etc in the market nowadays.

2. COMPUTER-AIDED DIAGNOSES

Computer-aided Diagnosis system assists doctors to interpret medical images. Various medical tests such as X-rays, MRI, ultrasounds etc are the sources of data that describes a patient's condition. Pattern recognition techniques are used to identify suspicious structures in the image to aid Computer-aided diagnosis.

3. COMPUTER VISION

The living beings use their eyes to see the world around them. Computer vision aims to give nearly same capabilities to a machine. It allows a machine to gain high-level understanding from digital images or videos and act accordingly.

Driverless cars are also one of the greatest applications of machine learning where car vision is made possible by advancement in the computer vision technology. To perform these tasks cameras are installed and they get input from these cameras. These tasks lie purely in the pattern recognition domain. A driverless robotic car named STANLEY was first to win the 2005 DARPA Grand Challenge. STANLEY is a Volkswagen Touareg that is equipped with cameras, radar, and laser rangefinders to sense the environment and the onboard software to command the steering, braking, and acceleration [30].

4. GAME PLAYING

IBM's DEEP BLUE became the first computer program to defeat the world champion Garry Kasparov by a score of 3.5 to 2.5. Then the other champions studied Kasparov's loss and were able to draw a few matches in subsequent years. But now highly efficient computer systems have been made

and most of the recent human-computer matches have been won by the computer.

5. LOGISTICS PLANNING

In logistics planning, we apply methods for cost reduction, capital reduction, and service improvement. During the Persian Gulf crisis of 1991, the U.S. forces used a Dynamic Analysis and Re-planning Tool (DART), for doing automated logistics planning and scheduling for transportation. This involved approximately 50,000 vehicles, cargo, and people at a time, and had taken into account various parameters like starting points, destinations, routes, and conflict resolution among all these parameters. The AI planning techniques generated a plan in hours that using older methods would have taken weeks.

6. TEXT MINING

It refers to the process of deriving the high-quality information from the text.

There are two different ways of mining the data i.e. Goal-oriented and Method-oriented mining. Any process that generates useful results that are not obvious is called Goal-oriented mining. And any process that involves extracting information from massive amount of data is called Method-oriented mining. Text mining is useful in a number of applications including business intelligence, automated classification of news articles, spam filter, automated placement of advertisement.

V. FUTURE SCOPE

Machine learning is a research area that has attracted the interest of many people and it has the potential to uncover many other problems.

In areas like game playing, logical inference and theorem proving, planning, and medical diagnosis, there are systems that can perform better than human experts. In other areas, such as learning, vision, robotics, and natural language understanding, there is a rapid improvement in performance through the application of better analytical methods. Continued research will give better capabilities in all of these areas.

Some of the most important future problems are discussed here.

A. EXPLAINING HUMAN LEARNING

As mentioned earlier, Machine learning is a field that provides a machine the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning theories try to imitate features of learning in humans and animals. However, the important stimuli in human or animal learning like horror, urgency, excitement, hunger are not yet taken into account in ML algorithms. This is a potential opportunity to discover a more generalised concept of learning.

B. PROGRAMMING LANGUAGES CONTAINING MACHINE LEARNING PRIMITIVES

Majority of applications in ML algorithms are incorporated with manually coded programs as part of the application software. In today's world there is an increasing need for a new programming language that is self-sufficient to support manually written code. This enables the coder to define a set of inputs-outputs for every "to be learned" program and opt for an algorithm. Some of the programming languages like Python are already making use of these concepts but in smaller scope.

C. PERCEPTION

A generalised concept of computer perception that can link ML algorithms is used highly in advanced vision, speech recognition etc. Research in machine perception solves the hard problems of understanding images, sounds, music and video. One of the main problems is the integration of different senses to

prepare a system that can induce self-supervised learning to estimate one sensory knowledge using the others.

VI. CONCLUSION

The machine learning field is concerned with the problem of how to construct a computer program that automatically improves with experience. In recent years, many successful machine learning applications have been developed. At the same time, there have been important advances in the theory and algorithms. The foremost target to design more efficient and practical general-purpose learning methods that can perform better over various domains. ML algorithms are completely data-driven and have the ability to examine a large amount of data in smaller intervals of time. Also they are often more accurate and not prone to human bias. ML algorithms have an edge over manual programming as the latter lacks the ability to adapt when exposed to a different environment.

VII. REFERENCES

[1]. Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng., Convolution Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations, Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

[2]. Kajaree Das, Rabi Narayan Behera., a Survey on Machine Learning: Concept, Algorithms and Application, International Journal of Innovative Research in Computer, and Communication Engineering, Feb 2017.

[3]. N. Cristianini and J. Shawe-Taylor, an Introduction to Support Vector Machines. Cambridge University Press, 2000.

[4]. E. Osuna, R. Freund, and F. Girosi. Support vector machines: training and applications. AI Memo 1602, MIT, May 1997.

[5]. Taiwo Oladipupo Ayodele, Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), InTech, 2010

[6]. T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K.Mazaitis, T. Mohamed, N. Nakashole, E. Platanios,A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov,M.Greaves, J. Welling, Never-Ending Learning, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2014

[7]. Wang, J. and Jebara, T. and Chang, S.-F. Semi-supervised learning using greedy max-cut.Journal of Machine Learning Research , Volume 14(1), 771-800 2013

[8]. Chapelle, O. and Sindhwani, V. and Keerthi, S. S. Optimization Techniques for Semi-Supervised Support Vector Machines, Journal of Machine Learning Research , Volume 9, 203–233, 2013

[9]. J. Baxter, A model of inductive bias learning. Journal of Artificial Intelligence Research, 12:149–198, 2000.

[10]. S. Ben-David and R. Schuller, Exploiting task relatedness for multiple task learning, Conference on Learning Theory, 2003.

[11]. W. Dai, G. Xue, Q. Yang, and Y. Yu, Transferring Naive Bayes classifiers for text classification. AAAI Conference on Artificial Intelligence, 2007.

[12]. Z. Marx, M. Rosenstein, L. Kaelbling, and T. Dietterich. Transfer learning with an ensemble of background tasks. In NIPS Workshop on Transfer Learning, 2005.

[13]. R Conway and D Strip, Selective partial access to a database, In Proceedings of ACM Annual Conference, 85 - 89, 1976

[14]. P D Stachour and B M Thuraisingham Design of LDV A multilevel secure relational database

management system, IEEE Trans. Knowledge and Data Eng., Volume 2, Issue 2, 190 - 209, 1990

[15]. R Oppliger, Internet security: Firewalls and beyond, Comm. ACM, Volume 40, Issue 5, 92 - 102, 1997

[16]. Rakesh Agrawal, Ramakrishnan Srikant, Privacy Preserving Data Mining, SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Volume 29 Issue 2,Pages 439-450, 2000

[17]. A. Carlson, J. Betteridge, B.Kisiel, B.Settles,E. R.Hruschka Jr,and T. M. Mitchell, Toward an architecture for never-ending language learning, AAAI, volume 5, 3, 2010

[18]. X. Chen, A. Shrivastava, and A. Gupta, Neil: Extracting visual knowledge from web data, In Proceedings of ICCV, 2013.

[19]. P. Donmezand J. G. Carbonell, Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In proceedings of the 17th ACM conference on information and knowledge management, 619–628. ACM, 2008

[20]. T. M.Mitchell, J. Allen, P. Chalasani, J. Cheng, O. Etzioni, M. N. Ringuetteand J. C. Schlimmer, Theo: A framework for self-improving systems, Arch. for Intelligence 323–356, 1991

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

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

[23]. Adankon, M. and Cheriet, M. Genetic algorithm-based training for semi-supervised SVM, Neural Computing and Applications , Volume 19(8), 1197-1206, 2010

[24]. T. M. Mitchell (1997), Machine Learning, McGraw-Hill International.

[25]. Stuart Russell, Peter Norvig (04-Jul-2016), Artificial Intelligence, A Modern Approach, Global Edition, Pearson Education Limited.