ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey on Machine Learning Algorithms for Building Smart Systems

Ajinkya Kunjir, Basil Shaikh

UG Student, Dept. of Computer Engineering, M.E.S College of Engineering, Pune, India

UG Student, Dept. of Computer Engineering, M.E.S College of Engineering, Pune, India

ABSTRACT: The algorithms which are implemented on the machines and which are also used to make machines intelligent are called as machine learning algorithms, also they can figure out how to perform important tasks by generalizing from examples. This approach is often tractable and cost-effective where manual programming is not. As more data becomes available, more complex problems can be tackled and solved. As a result, machine learning is widely used in computer science, artificial intelligence and other fields. However, developing successful machine learning applications requires understanding of smart systems and algorithms required to construct it. This paper aggregates and summarizes the types of machine learning types and algorithm which are required to construct a smart or an expert system. From the perspective of data and learning divide in the society, there exist few limitations in the data sets which are supposed to be investigated, the metrics we employ for evaluation, and the degree to which results are communicated back to their originating domains. The problem of learning and decision making is at the core level of discussion in biological as well as artificial aspects. So researchers and scientific practitioners introduced Machine Learning as widely used concept in Artificial Intelligence. It is the concept which teaches machines to detect different patterns and to adapt to new methods. In the 21st century, the concept of machine learning is used in many applications and is an important concept for intelligent systems which leads to the effective and innovative technology and more advanced concepts of artificial thinking.

KEYWORDS: Machine learning, Supervised learning, Unsupervised learning, Algorithms, Decision trees, Artificial intelligence, Smart systems.

I. INTRODUCTION

Machine learning systems auto-learns the programs from data and information fed to the system. This is often a very attractive alternative to manually constructing them, and in the last few years the use of machine learning has increased rapidly throughout the fields of artificial intelligence, computer science and beyond. Machine learning and algorithms are used in Web search, spam filters, decision support systems, recommender systems, ad placement, credit scoring, fraud detection, and many other applications. A recent study report from the McKinsey Global Institute asserts that machine learning will be the driver of the next big wave of innovation [16]. However, much of the "folk knowledge" that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing lessthan-ideal results. Yet much of this folk knowledge is fairly easy to communicate. This research paper emphasizes and focuseson different types of machine learning algorithms and their most efficient use to make decisions more efficient and complete the task in more optimized or trivial form. Different algorithm gives machine different learning experience and adapting other things from the environment. Based on these algorithms the machine takes the decision and performs the specialized tasks. So it is very important for the algorithms to be optimized and complexity should be reduced because more the efficient algorithm more efficient decisions will the machine makes. Machine Learning algorithms do not totally dependent on nature's bounty for both inspiration and mechanisms. Fundamentally and scientifically these algorithms depends on the data structures used as well as theories of learning cognitive and genetic structures. But still natural procedure for learning gives great exposures for understanding and good scope forvariety of different types of circumstances. Many machine learning algorithm are generally being evolved from present thinking in cognitive science and neural networks. Overall we can say that learning is defined in terms of improving performance based on some measure.

II. Related work

A. BACKGROUND AND RESEARCH

Sally Goldman et.alin their paper proposed the practical learning scenarios or situations where we have less amount of labelled data along with a large pool of unlabelled data and presented a strategy for using the unlabelled data to improve the standard supervised learning algorithms[1]. An assumption was made that there are two different supervised learning algorithms which both output a hypothesis that defines a partition of instance space. They finally concluded that two supervised learning algorithms can be used successfully for labeling data for each other. Y. Bengio in his paper gave a brief overview of unsupervised learning from the perspective of statistical modeling[2]. According to the proven facts, unsupervised learning can be cheered from information theory and Bayesian principles. The models were then further reviewed for unsupervised learning and was even concluded that statistics provides a coherent framework for learning from data and for reasoning under uncertainty and also they mentioned the types of models like Graphical model which played an important role in learning systems for variety of different kinds of data. Rich Caruana et.al [3] has studied various supervised learning methods which were introduced in last few years and provide a large-scale empirical comparison between ten supervised learning methods. These methods include: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees and boosted stumps. Various performance based criteria were used to evaluate the learning methods. Niklas lavesson et.al [4] answered the questions that how to evaluate and analyse supervised learning algorithms and classifiers. One conclusion of the analysis is that performance is often only measured in terms of accuracy, latency and response. However, some researchers have questioned the validity of using accuracy as the only performance metric. They have given a different approach for evaluation of supervised learning, i.e Measure functions, a limitation of current measure functions is that they can only handle two-dimensional instance spaces. They present the design and implementation of a generalized multi-dimensional measure function and demonstrate its use through a set of experiments. The results indicate that there are cases for which measure functions may be able to capture aspects of performance that cannot be captured by cross-validation tests. Finally, they investigate the impact of learning algorithm parameter tuning.

B. MACHINE LEARNING ALGORITHMS: A SURVEY

As described in the previous sections, we know that machine learning is a scenario in which patterns and knowledge is fed to the systems for learning. Various algorithms are designed and analysed for his purpose. To know whether an agent deployed in the environment has learned, we must define a measure of success. The measure is usually not how well the agent performs on the training experiences, but how well the agent performs for new experiences. In this survey paper we will consider the two main types of algorithms i.e. supervised & unsupervised learning. Machine Learning also meets the fields of data mining and classification.Taiwo Oladipupo Ayodele in his "Types of machine learning algorithms" described types of ML algorithms with their details, advantages, and dis-advantages[5].The ML algorithms are applied on the training datasets, so that when a new data instance comes in the play for prediction of the class label the ML algorithm acts on the new instance and also predicts its class based on previous experiences and records.

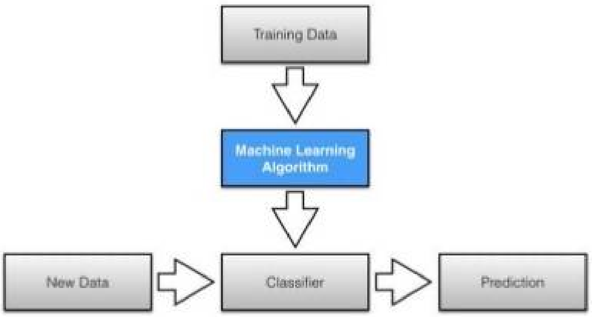

Fig 1:Use of ML Algorithms in Data Mining

- *Supervised Learning:*

Supervised learning is an learning method in which both the inputs and outputs can be perceived. Based on this training data, the algorithm has to generalize such that it is able to correctly respond to all possible inputs. This supervised algorithm is expected to produce precise output for inputs that weren't spotted during training. In supervised learning what has to be learned is specified for each example. Supervised classification occurs when a trainer provides the classification for each example. Supervised learning of actions occurs when the agent is given immediate feedback, the feedback may be either positive or negative about the value of each action. V. N. Vapnik in his springer edition of NewYork mentioned that, in order to solve a give problem using supervised learning algorithm one has to follow some certain steps[6].
1) Determine the type of training examples.
2) Gather and arrange a training set.
3)Determine the input feature representation of learned function.
4)Determine the structure of learning function & corresponding learning algorithm.
5) Complete the design and run the learning algorithm on the gather set of data.
6) Evaluate the accuracy of the learned function also the performance of the learning function should be measured and then the performance should be again measured on the set which is different from the training set.

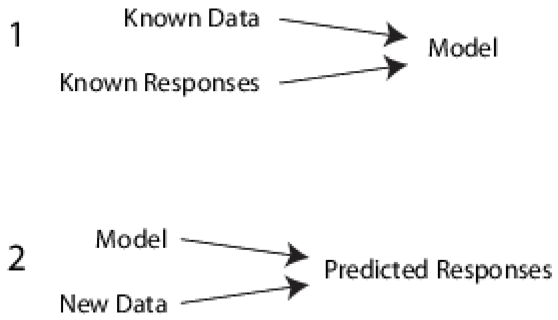

Fig 2 : Supervised Learning

Supervised Learning can be split into two classes:
1. Classification of responses that can have just two boolean values, such as 'true' or 'false'. Classification algorithm applies to nominal atttributes or values and not ordinal ones.
2. Regression for responses that are a real number, such as miles per gallon of a particular car.
Another example of supervised learning can be a baby which is learning to walk with the help or guidance of parents or gaurdians. Supervised learning can be described as a learning in which all the data examples of a dataset are labelled and the algorithm is supervised to classify the labels of the previously unknown or unseen instances.

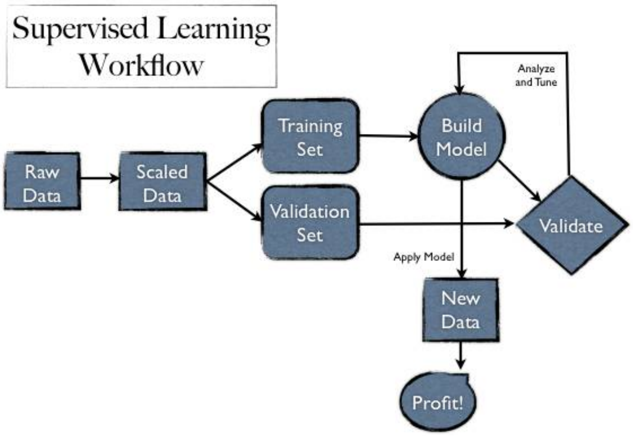

Fig 3: Supervised Learning Workflow

- *Unsupervised Learning:*

Unsupervised learning seems much harder, the goal of this technique is to have the computer learn how to do something that we don't tell it how to do! There are two approaches to unsupervised learning. The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Good decisions and actions will result in good reward and positive feedback. Note that this type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards. I. Wittenet.al in 2011 elaborated that this approach nicely generalizes to the real world, where agents might be rewarded for doing certainactions and punished for doing others[7]. Often, a form of reinforcement learning can be used for unsupervised learning, where the agent bases its actions on the previous rewards and punishments without necessarily even learning any information about the exact ways that its actions affect the world. This can be very advantageous in cases where calculating every possibility is very time consuming. On the other hand, it can be very time consuming to learn by, essentially, trial and error method. But this kind of learning can be very powerful because it assumes no pre-discovered classification of examples. In some cases, for example, our classifications may not be the best possible. One better example is that the conventional wisdom about the game of backgammon(early 90’s) was turned on its head when a series of computer programs and gammons that learned through unsupervised learning became stronger than the best human chess players merely by playing themselves over and over. A second type of unsupervised learning is called clustering. In this type of learning, the goal is not to maximize a utility function, but simply to find similarities in the training data. Clustering can also be defined as grouping items or entities which have similar properties. The assumption is often that the clusters discovered will match reasonably well with an intuitive classification. For instance, clustering individuals based on demographics might result in a clustering of the wealthy in one group and the poor in another. Although the algorithm won't have names to assign to these clusters, it can produce them and then use those clusters to assign new examples into one or the other of the clusters. This is a data-driven approach that can work well when there is sufficient data; for instance, social information filtering algorithms, such as those that Amazon.com use to recommend books, are based on the principle of finding similar groups of people and then assigning new users to groups. In some cases, such as with social information filtering, the information about other members of a clustercan be sufficient for the algorithm to produce meaningful results. In other cases, it may be the case that the clusters are merely a useful tool for a human analyst. Unfortunately, even unsupervised learning suffers from the problem of over-fitting the training data. There's no silver bullet to avoid the problem because any algorithm that can learn from its inputs needs to be quite powerful.

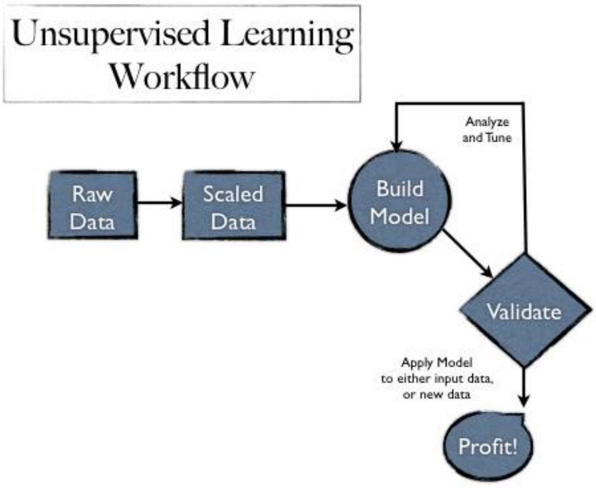
Unsupervised Learning Workflow
Raw Data
Scaled Data
Build Model
Analyze and Tune
Validate
Apply Model to either input data, or new data
Profit!

Fig 4: Unsupervised Learning Workflow

- *Analytics of Learning methods:*

As we all know, supervised learning method is a ML method in which all the data examples of a dataset are labelled and unsupervised learning method is the method in which all the data examples of the dataset are not labelled. Semi-supervised learning method is a learning method in which some data examples are labelled and some are not labelled. Apart from these, active learning method is a method in which the ML algorithm decides which data examples to label and which no to label.

DOI: 10.15680/IJIRCCE.2017. 0501057

The below given figure states the difference between supervised and unsupervised learning methods. The figure describes the observation inputs and outputs of both the learning techniques and also maps the variables from input to the output.

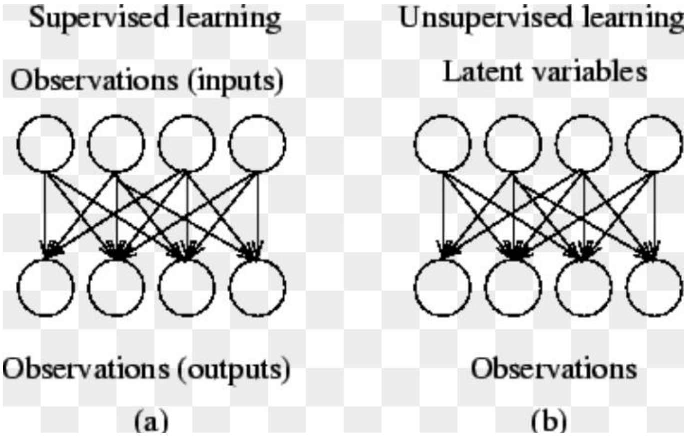

Fig 5: Comparative study of observations

III. SMART SYSTEMS

A. INTRODUCTION TO SMART SYSTEMS

A smart system or a smart device is any system that is advanced and highly equipped with technical assets such as network connectivity, distributed, connected, eco-friendly, personalized, etc. A smart system also possesses characteristics like transparency, reliability, security, authentication, etc. Smart devices can range from micro sized smart sensor to huge sized computer. Every device that acts according to you is a smart device. PDA's, tablets, laptops, smart watch, smart keychain, smart bands, mobile phones, transponders, personal computers, etc.

Fig 6: Range of Smart Devices

B. IMPLEMENTATION OF ML ON SMART SYSTEMS

ManyML studies and research cases involvedomainexpertsas helpers who help define the ML problem and label data for classification tasks. They can also provide the lost connection between an ML performance plot and its significance to the problem domain. This can help reduce thenumber of cases where a smart system perfectly solves a sub-problem of little interest to the relevant scientific community, or where the ML system's performance looks good numerically but is insufficiently reliable to ever be adopted.Many scientific and research practitioners have successfully tackled the challenges and other aspects of the context. Warrick et al. (2010) provides an example of ML work that tackles all the challenges and aspects. Working with doctors and clinicians, they developed a system to detect fetal hypoxia (oxygen deprivation) and enable emergency intervention that literally saves babies from brain injuries or death. After publishing their results, which demonstrated the ability to have detected 50% of fetal hypoxia cases early enough for intervention, with an acceptable false positive rate of 7.5%, they are currently working on clinical trials as the next step towards wide deployment. Many such examples exist[8].

IV.MACHINE LEARNING ALGORITHMS

A. DECISION TREE INDUCTION

Decision tree algorithms are supervised learning algorithms for which all the data examples of a dataset are labelled. The learning of decision tree algorithms is called as decision tree induction. The algorithms such as ID3, C4.5, CHAID and CART belong to category of decision tree algorithms. The biggest advantage of decision trees is that it can handle both categorical and numerical attributes. The drawback of decision tree however, is that it works good on small datasets but can cause a lag for big datasets. IF-THEN rules can be derived by considering the nodes and edges of the decision tree. The mapping of datasets to the graphical tree can be shown by creping the action on attributes as the nodes, and the outcome of the actions as the edges. The attribute which has the highest information gain can be used as a splitting attribute for the construction of the tree. Other methods for finding the splitting attribute in the various decision tree algorithms are gini index for ID3, and information gain for C4.5.

B. NAÏVE BAYES CLASSIFIER

Naive Bayes is a simple technique for buildingsimple classifiers: models that assigns class labels to data examples that are previously unknown, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a group of algorithms based on a mutual principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. A naive bayes classifier considers each of these features to contribute independently to the probability. For some types of probability models, naive Bayes models can be trained very efficiently in a supervised learning technique. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood and in simple words, one can coordinate with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods. Despite their naive design, drawbacks and assumptions, naive Bayes classifiers have worked pretty well in many complex real-world situations. The research and comparative study between all the algorithms in 2006 showed that Bayes classification is outperformed by otherapproaches, such as boosted trees or random forests· An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. Bayes Theorem provides a base and a set of basic principles for the naïve bayes algorithm.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad \text{Eq.1}$$

As shown in the above equation, the conditional probability of P(A|B) is given as the ratio of conditional and unconditional probabilities. Where, P(B|A) is the conditional probability of B given A, P(A) is the unconditional probability of event A, and P(B) is the unconditional probability of event B.

C. CLUSTERING ALGORITHMS

Clustering is an descriptive data mining technique in which all the data examples or instances which possess similar characteristics and properties are grouped together in a cluster. Clustering is an unsupervised learning technique in which the data examples in a dataset are not labelled. The various clustering algorithms and techniques are state below:

1. Hierarchical clustering: This a clustering technique which divides the cluster hierarchically by two methods i.e Agglomerative or Divisive.

- Agglomerative Clustering: In this method small clusters pair up to form a big cluster in a bottom up approach
- Divisive Clustering: Clustering approach in which a big cluster is broken down into small clusters in a top down manner.

2. Partitioning Clustering : Technique in which the list or the array of data instances is partitioned into equal or in-equal parts where each part is represented by a cluster mean, centroid or a data instance as a cluster representative.

- K-means: K-means is a partitioning algorithm in which the list of data examples is divided into set of small clusters where each cluster is represented by a cluster mean. The partitioning stops at the point of convergence.
- K-mediods: Algorithm in which data instances are clustered and are represented by a centroid which is a data example. This approach being robust, avoids noise interference and outliers.

V. PROPOSED WORK

We propose an efficient use of machine learning algorithms and techniques which can be implemented on smart systems for their construction and effective usage. The supervised and unsupervised machine learning techniques are discussed in this paper. The main issues in smart systems using machine learning algorithms for fast processing is task scheduling. The various issues and challenges are discussed in the later half of the survey. The paper also outlines the mechanisms and equations of ML algorithms.

A. ISSUES AND CHALLENGES

A source from the internet stated that Carbonell in 1992, proposed a list of challenges for the field, not to increase its impact but instead to “put the fun back into machine learning” they included the following points[9].

1. Invention of a new physical law leading to a published and referred scientific article.
2. Outperforming all hand-built medical diagnosis systems with an ML solution that is deployed and regularly used at at least two institutions.
3. A law passed or legal decision made that relies on the result of an ML analysis.
4. A human life saved through a diagnosis or intervention recommended by an ML system.

These challenges are meant to capture the entire process of a successful machine learning concept, including performance, infusion, and impact. They differ from currently existing challenges such as the DARPA Grand Challenge, the Netflix Prize, and the Yahoo! Learning to Rank Challenge in that they do not focus on any single problem domain, nor a particular technical capability.

VI. CONCLUSION AND FUTURE WORK

Many current ML researches suffer from a growing detachment and divide from those real problems. Many investigators and stakeholders withdraw into their private studies with a copy of the data set and work in isolation to solidify algorithmicperformance. PublishingresultstotheML community is the end of the process. The worlds of finance, politics, education, medicine, law, and more stand to benefit from systems that can adapt, analyse, and take actions. This paper identifies the problems in constructing smart systems by implementing machine learning algorithms on it. Aiming for real impact does not just increase our job satisfaction, it is the only way to get the rest of the world to notice, recognize, value, and adopt ML solutions. The paper also outlines the possibilities and advanced enhancements of machine learning and techniques in fields of pattern recognition, image processing, text processing in nearby future.

REFERENCES

1] Sally Goldman; Yan Zhou, "Enhancing Supervised Learning with Unlabeled Data", Department of Computer Science, Washington University, St.Louis, MO 63130 USA.

2]Y. Bengio. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2:1–127, 2009

3]Rich Caruana; Alexandru Niculescu- Mizil,"An Empirical Comparison of Supervised Learning Algorithms", Department of Computer Science, Cornell University, Ithaca, NY 14853 USA.

4]Niklas Lavesson,"Evaluation and Analysis of Supervised Learning Algorithms and Classifiers", Blekinge Institute of Technology Licentiate Dissertation Series No 2006:04,ISSN 1650-2140,ISBN 91-7295-083-8.

5]Types of Machine Learning Algorithms, Taiwo Oladipupo Ayodele, University of Portsmouth, United Kingdom.

6] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, NY, 1995.

7] I. Witten, E. Frank, and M. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Mateo, CA, 3rd edition, 2011

8]Warrick, P. A., Hamilton, E. F., Kearney, R. E., and Precup, D. A machine learning approach to the detection of fetal hypoxia during labor and delivery. In Proc. of the Twenty-Second Innovative Applications of Artificial Intelligence Conf., pp. 1865–1870, 2010

9]https://en.wikipedia.org/wiki/**Carbonell**