A Review on Various Machine Learning Algorithms

Ambarish Ravindra Bhuyar [1]

Sumaiyya Z Khan [2]

Prof. Hemant N. Watane [3]
Department of Information Technology
Sipna College of Engineering and Technology, Amravati, Maharashtra, India

Abstract: Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. Learning algorithms in many applications that's we make use of daily. Every time a web search engine like Google is used to search the internet, one of the reasons that work so well is because a learning algorithm that has learned how to rank web pages. These algorithms are used for various purposes like data mining, image processing, predictive analytics, etc. to name a few. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically. In this paper, a brief review and future prospect of the vast applications of machine learning algorithms has been made.

Keywords: Algorithm, Machine Learning, Supervised learning, Unsupervised learning, Reinforcement learning.

I. Introduction

Since their evolution, humans have been using many types of tools to accomplish various tasks in a simpler way. The creativity of the human brain led to the invention of different machines. These machines made the human life easy by enabling people to meet various life needs, including travelling, industries and computing and Machine learning is the one among them [1]. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves [2] [3]. Many mathematicians and programmers apply several approaches to find the solution of this problem. Some of them are demonstrated in Figure 1.

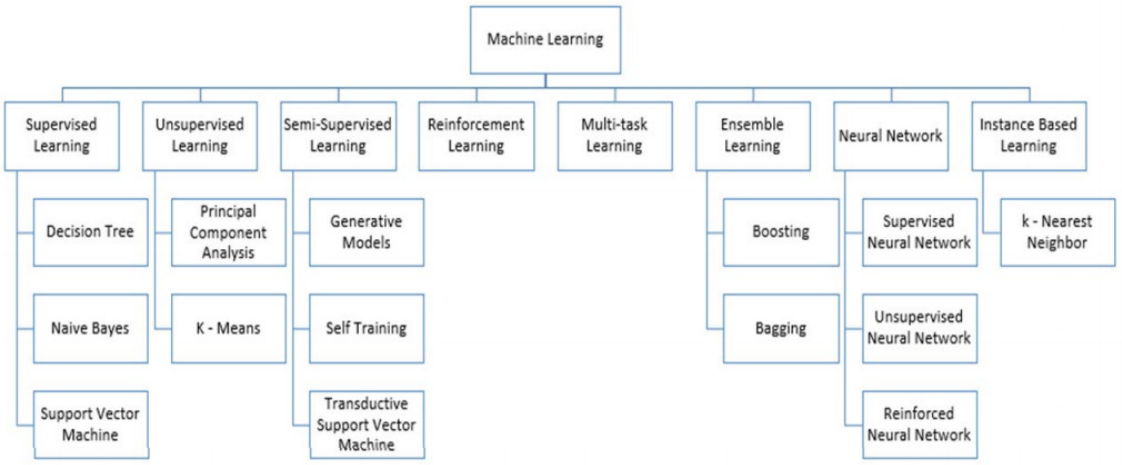

Figure 1: Types of Machine Learning [2][3]

According to Arthur Samuel Machine learning is defined as the field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel was famous for his checkers playing program. Machine learning (ML) is used to teach machines how to handle the data more efficiently [4]. Sometimes after viewing the data, we cannot interpret the extract 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. Many

studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets [5].

II. Types of Real World Data.

Machine learning algorithms typically consume and process data to learn the related patterns about individuals, business processes, transactions, events, and so on. In the following we discuss various types of real-world dataas well as categoriesof machine learning algorithms.Usually, the availability of data is considered as the key to construct a machine learning model or data-driven real world systems [6, 7]. Data can be of various forms, such as structured, semi-structured, or unstructured [8, 9].Besides, the "metadata" is another type that typically represents data about the data. In the following, we briefly discuss these types of data.

- *Structured:* It has a well-defined structure, conforms toa data model following a standard order, which is highlyorganized and easily accessed, and used by an entity or acomputer program. In well-defined schemes, such asrelationaldatabases, structured data are typically stored, i.e.,in a tabular format. For instance, names, dates, addresses, credit card numbers, stock information, geo location, etc. are examples of structured data.
- *Unstructured:* On the other hand, there is no pre-definedformat or organization for unstructured data, makingitmuch more difficult to capture, process, and analyze, mostly containing text and multimedia material. Forexample, sensor data, emails, blog entries, wikis, andword processing documents, PDF files, audio files,videos, images, presentations, web pages, and many. Other types of business documents can be considered asunstructured data.
- *Semi-structured:* Semi-structured data are not stored ina relational database like the structured datamentionedabove, but it does have certain organizational propertiesthat make it easier to analyze. HTML, XML, JSON documents, No SQL databases, etc., are some examplesof semi-structured data.
- *Metadata:* It is not the normal form of data, but "dataabout data". The primary difference between "data"and"metadata" is that data are simply the material that canclassify, measure, or even document somethingrelativeto an organization's data properties. On the other hand,metadata describes the relevant datainformation, givingit more significance for data users. A basic example of adocument's metadata might bethe author, file size, dategenerated by the document, keywords to define the document, etc.

III. Types of Machine Learning Techniques

Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning [10]. In the following, we briefly discuss each type of learning technique with the scope of their applicability to solve real-world problems.

3.1 *Supervised:* Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs [8]. It uses labelled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identified to be accomplished from a certain set of inputs [7], i.e., a task driven approach. The most common supervised tasks are "classification" that separates the data, and "regression" that fits the data. For instance, predicting the class label or sentiment of a piece of text, like a tweet or a product review, i.e., text classification is an example of supervised learning.

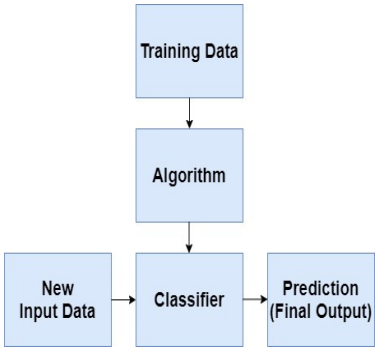

Figure 2: Supervised Learning Model.

1) Decision Tree: Decision trees are those type of trees which groups attributes by sorting them based on their values. Decision tree is used mainly for classification purpose. 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

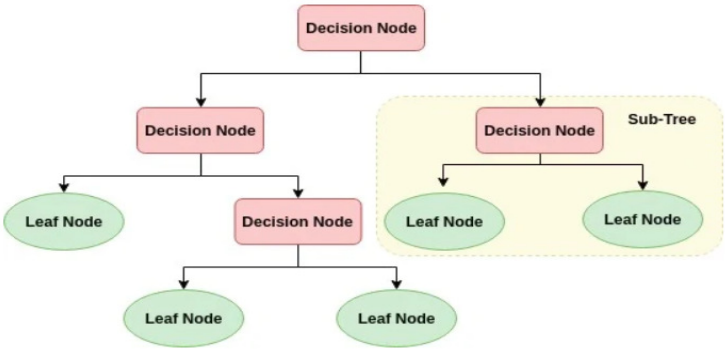

Figure 3: Decision Tree

2) Naïve Bayes: Naïve Bayes mainly targets the text classification industry. It is mainly used for clustering and classification purpose [6]. The underlying architecture of Naïve Bayes depends on the conditional probability. It creates trees based on their probability of happening. These trees are also known as Bayesian Network.

3) Support Vector Machine: Another most widely used state-of-the-art machine learning technique is Support Vector Machine (SVM). It is mainly used for classification. SVM works on the principle of margin calculation. It basically, draw margins between the classes. The margins are drawn in such a fashion that the distance between the margin and the classes is maximum and hence, minimizing the classification error. An example of working of SVM is given in Figure 4.

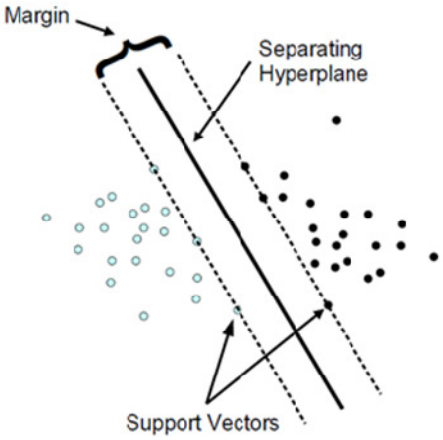

Figure 4: Support Vector Machine

3.2 *Unsupervised:* These are called unsupervised learning because unlike supervised learning above there is no correct answer and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data. The unsupervised learning algorithms learn few features from the data. When new data is introduced, it uses the previously learned features to recognize the class of the data. It is mainly used for clustering and feature reduction.

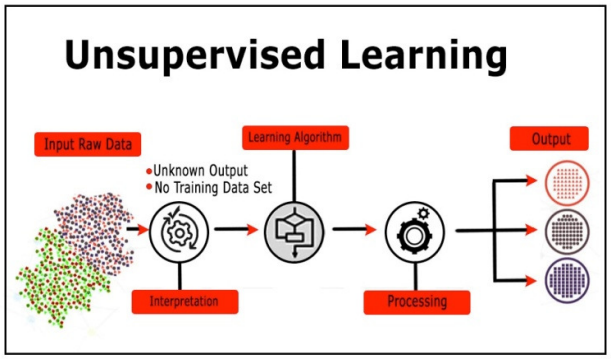

Figure 5: Unsupervised Learning

The two main algorithms for clustering and dimensionality reduction techniques are discussed below.

1) K-Means Clustering: K-means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. As shown in Figure. 6.

The way k-means algorithm works is as follows:

- Specify number of clusters K.

-Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

-Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

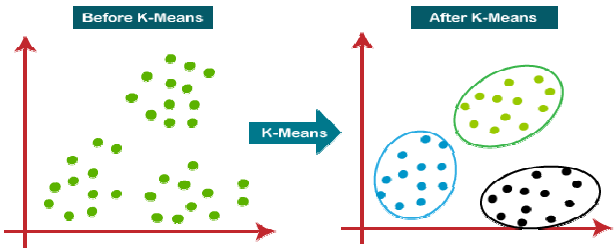

Figure 6: K-Mean Clustering

3.3 *Semi – Supervised:* Semi Supervised learning is a technique which combines the power of both supervised and unsupervised learning. It can be fruit-full in those areas of machine learning and data mining where the unlabeled data is already present and getting the labeled data is a tedious process [11]. There are many categories of semi-supervised learning [12]. The ultimate goal of a semi-supervised learning model is to provide a better outcome for prediction than that produced using the labeled data alone from the model. Some application areas where semi-supervised learning is used include machine translation, fraud detection, labelling data and text classification.

1) Generative Models: Generative models are one of the oldest semi-supervised learning method assumes a structure like p(x,y) = p(y)p(x|y) where p(x|y) is a mixed distribution e.g. Gaussian mixture models. Within the unlabeled data, the mixed components can be identifiable. One labelled example per component is enough to confirm the mixture distribution.

2) Self-Training: In self-training, a classifier is trained with a portion of labeled data. The classifier is then fed with unlabeled data. The unlabeled points and the predicted labels are added together in the training set. This procedure is then repeated further. Since the classifier is learning itself, hence the name self-training.

3) Transductive SVM: Transductive support vector machine or TSVM is an extension of SVM. In TSVM, the labelled and unlabeled data both are considered. It is used to label the unlabeled data in such a way that the margin is maximum between the labelled and unlabeled data. Finding an exact solution by TSVM is a NP-hard problem.

3.4 *Reinforcement learning:* is a type of learning which makes decisions based on which actions to take such that the outcome is more positive. The learner has no knowledge which actions to take until it's been given a situation. The action which is taken by the learner may affect situations and their actions in the future. Reinforcement learning solely depends on two criteria: trial and error search and delayed outcome [13]. It is a powerful tool for training AI models that can help increase automation or optimize the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing and supply chain logistics, however, not preferable to use it for solving the basic or straightforward problems.

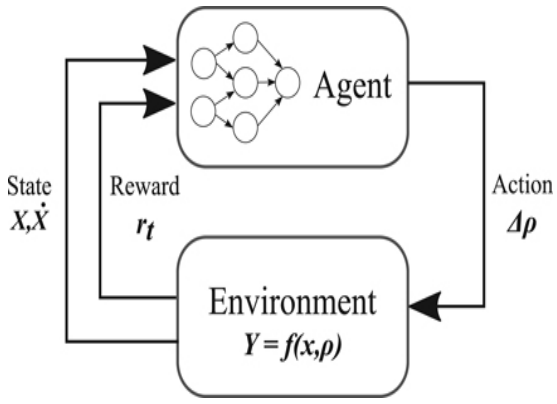

Figure 7: The Reinforcement Learning Model

3.5 *Multi-Task learning* is a sub-field of Machine Learning that aims to solve multiple different tasks at the same time, by taking advantage of the similarities between different tasks. This can improve the learning efficiency and also act as a regularizer. Formally, if there are n tasks (conventional deep learning approaches aim to solve just 1 task using 1 particular model), where these n tasks or a subset of them are related to each other but not exactly identical, Multi-Task Learning (MTL) will help in improving the learning of a particular model by using the knowledge contained in all the n tasks. We usually aim to learn a good representation of the features or attributes of the input data to predict a specific value. Formally, we aim to optimize for a particular function by training a model and fine-tuning the hyper parameters till the performance can't be increased further. By using MTL, it might be possible to increase performance even further by forcing the model to learn a more generalized representation as it learns (updates its weights) not just for one specific task but a bunch of tasks. Biologically, humans learn in the same way. We learn better if we learn multiple related tasks instead of focusing on one specific task for a long time.[14]

3.6 *Ensemble Learning:* When various individual learners are combined to form only one learner then that particular type of learning is called ensemble learning. The individual learner may be Naïve Bayes, decision tree, neural network, etc. Ensemble learning is a hot topic since 1990s. It has been observed that, a collection of learners is almost always better at doing a particular job rather than individual learners [15]. Two popular Ensemble learning techniques are given below [17]:

1) Boosting: Boosting is a technique in ensemble learning which is used to decrease bias and variance. Boosting creates a collection of weak learners and converts them to one strong learner. A weak learner is a classifier which is barely correlated with true classification. On the other hand, a strong learner is a type of classifier which is strongly correlated with true classification [17].
2) Bagging: Bagging or bootstrap aggregating is applied where the accuracy and stability of a machine learning algorithm needs to be increased. It is applicable in classification and regression. Bagging also decreases variance and helps in handling over fitting [18].

3.7 *Neural Network Learning:* Neural Network Learning The neural network (or artificial neural network or ANN) is derived from the biological concept of neurons. A neuron is a cell like structure in a brain. To understand neural network, one must understand how a neuron works. A neuron has mainly four parts (see Fig. 8). They are dendrites, nucleus, soma and axon.

The dendrites receive electrical signals. Soma processes the electrical signal. The output of the process is carried by the axon to the dendrite terminals where the output is sent to next neuron. The nucleus is the heart of the neuron. The inter-connection of neuron is called neural network where electrical impulses travel around the brain.

An artificial neural network behaves the same way. It works on three layers. The input layer takes input (much like dendrites). The hidden layer processes the input (like soma and axon). Finally, the output layer sends the calculated output (like dendrite terminals) [17]. There are basically three types of artificial neural network: supervised, unsupervised and reinforcement [18].

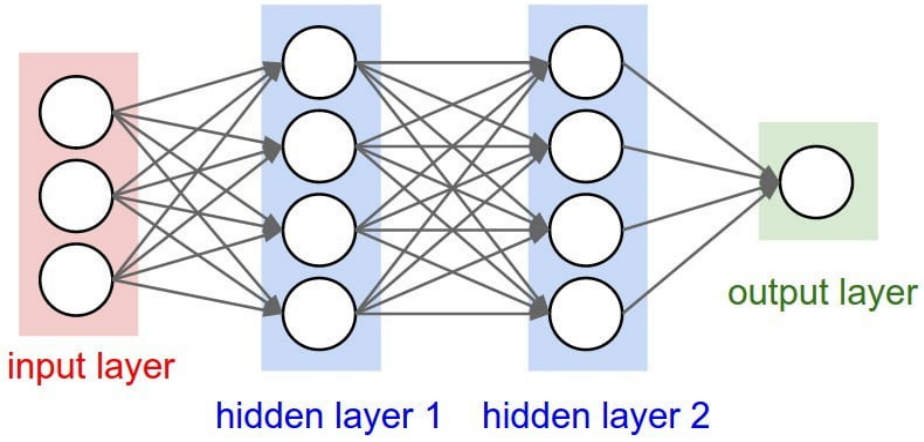

Figure 8: Structure of Artificial Neural Network

1) Supervised Neural Network: In the supervised neural network, the output of the input is already known. The predicted output of the neural network is compared with the actual output. Based on the error, the parameters are changed, and then fed into the neural network again. Figure 9 will summarize the process. Supervised neural network is used in feed forward neural network.

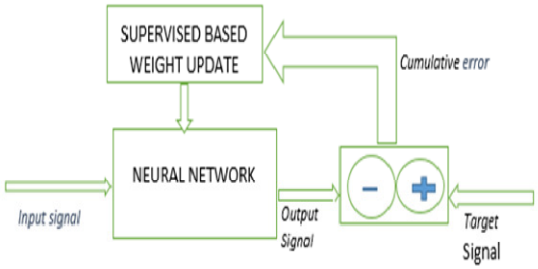

Figure 9: Supervised Neural Network

2) Unsupervised Neural Network: Here, the neural network has no prior clue about the output the input. The main job of the network is to categorize the data according to some similarities. The neural network checks the correlation between various inputs and groups them. The schematic diagram is shown in Figure 10.

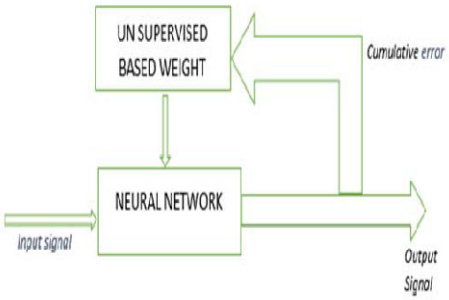

Figure 10: Unsupervised Neural Network

3) Reinforced Neural Network: In reinforced neural network, the network behaves as if a human communicates with the environment. From the environment, a feedback has been provided to the network acknowledging the fact that whether the decision taken by the network is right or wrong. If the decision is right, the connections which points to that particular output is strengthened. The connections are weakened otherwise. The network has no previous information about the output. Reinforced neural network is represented in Figure 11.

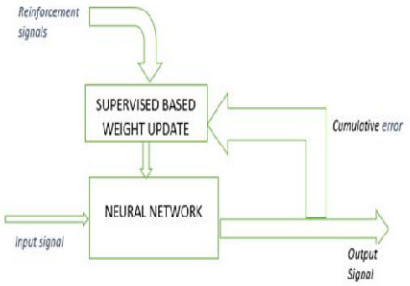

Figure 11: Reinforced Neural Network

3.8 Instance-Based Learning: The Machine Learning systems which are categorized as instance-based learning are the systems that learn the training examples by heart and then generalizes to new instances based on some similarity measure. It is called instance-based because it builds the hypotheses from the training instances. It is also known as memory-based learning or lazy-learning. The time complexity of this algorithm depends upon the

size of training data. The worst-case time complexity of this algorithm is O (n), where n is the number of training instances.

For example, If we were to create a spam filter with an instance-based learning algorithm, instead of just flagging emails that are already marked as spam emails, our spam filter would be programmed to also flag emails that are very similar to them. This requires a measure of resemblance between two emails. A similarity measure between two emails could be the same sender or the repetitive use of the same keywords or something else.

Some of the instance-based learning algorithms are :

- *K Nearest Neighbor (KNN):* 1) K-Nearest Neighbor: In k-nearest neighbor (or KNN), the training data (which is well-labeled) is fed into the learner. When the test data is introduced to the learner, it compares both the data. K most correlated data is taken from training set. The majority of k is taken which serves as the new class for the test data [27].
- *Self-Organizing Map (SOM):* The Self-Organizing Map (SOM) is an unsupervised learning algorithm introduced by Kohonen. In the area of artificial neural networks, the SOM is an excellent data-exploring tool as well. It can project high-dimensional patterns onto a low-dimensional topology map. The SOM map consists of a one or two dimensional (2-D) grid of nodes. These nodes are also called neurons. Each neuron's weight vector has the same dimension as the input vector. The SOM obtains a statistical feature of the input data and is applied to a wide field of data classification. SOM is based on competitive learning. In competitive learning , neuron activation is a function of distance between neuron weight and input data. An activated neuron learns the most and its weights are thus modified. If a similar pattern is found again, then the same neuron may be activated again. This means that a particular neuron wins repeatedly [19].
- *Learning Vector Quantization (LVQ)*: Learning Vector Quantization (or LVQ) is a type of Artificial Neural Network which also inspired by biological models of neural systems. It is based on prototype supervised learning classification algorithm and trained its network through a competitive learning algorithm similar to Self Organizing Map. It can also deal with the multiclass classification problem. LVQ has two layers, one is the Input layer and the other one is the Output layer. The architecture of the Learning Vector Quantization with the number of classes in an input data and n number of input features

IV. Conclusion:

Today each and every person is using machine learning knowingly or unknowingly. From getting a recommended product in online shopping to updating photos in social networking sites, this paper surveys various machine learning algorithms. We have also introduced most of the popular machine learning algorithms. Also conducted a comprehensive overview of machine learning algorithms for intelligent data analysis and applications. According to our goal, we have briefly discussed how various types of machine learning methods can be used for making solutions to various real-world issues.

References

[1] Batta Mahesh, "Machine Learning Algorithm – A Review", International Journal of Science and Research(IJSR), ISSN: 2319-7064.

[2] M. Welling, "A First Encounter with Machine Learning"

[3] M. Bowles, "Machine Learning in Python: Essential Techniques for Predictive Analytics", John Wiley & Sons Inc., ISBN: 978-1-118-96174-2

[4] W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0

[5] J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-15, No. 4, August 1985

[6] Sarker IH, Hoque MM, MdKUddin, Tawfeeq A. Mobile datascience and intelligent apps: concepts, aibasedmodeling andresearch directions. Mob NetwAppl, pages 1–19, 2020.

[7] Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, Ng A.Cybersecurity data science: an overviewfrom machine learningperspective. J Big Data. 2020;7(1):1–29.

[8] Han J, Pei J, Kamber M. Data mining: concepts and techniques.Amsterdam: Elsevier; 2011.

[9] McCallum A. Information extraction: distilling structured datafrom unstructured text. Queue. 2005;3(9):48–57.

[10] Mohammed M, Khan MB, Bashier Mohammed BE. Machine learning: algorithms and applications. CRC Press; 2016.

[11] X. Zhu, A. B. Goldberg, "Introduction to Semi – Supervised Learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, Vol. 3, No. 1, Pages 1-130

[12] X. Zhu, "Semi-Supervised Learning Literature Survey", Computer Sciences, University of Wisconsin-Madison, No. 1530, 2005

[13] R. S. Sutton, "Introduction: The Challenge of Reinforcement Learning", Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992

[14] R. Caruana, "Multitask Learning", Machine Learning, 28, 41-75, Kluwer Academic Publishers, 1997

[15] D. Opitz, R. Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of Artificial Intelligence Research, 11, Pages 169-198, 1999

[16] Z. H. Zhou, "Ensemble Learning", National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[17] Z. H. Zhou, "Ensemble Learning", National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China.

[18] https://en.wikipedia.org/wiki/Bootstrap_aggregating.

[19] Vikas Chaudhary, R.S. Bhatia, Anil K. Ahlawat, A novel Self-Organizing Map (SOM) learning algorithm with nearest and farthest neurons, Alexandria Engineering Journal, Volume 53, Issue 4, 2014, Pages 827-831, ISSN 1110-0168.