A SURVEY ON MACHINE LEARNING APPROACHES AND ITS TECHNIQUES:

2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS) 978-1-7281-4862-5/20/$31.00 ©2020 IEEE 10.1109/SCEECS48394.2020.190

Thomas. Rincy. N
Ph.D. (Scholar)
University Institute of Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal (M.P), India
rinc_thomas@rediffmail.com

Dr. Roopam Gupta
Professor
University Institute of Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal (M.P), India
roopamgupta@rgtu.net.

Abstract: With the data and information is available at a tremendous rate, there is a need for machine learning approaches. Machine learning, it analyses the study and constructs the algorithms by making prediction on data. It builds model from the inputs to make the decisions or predictions. Machine learning algorithms it assists in bridging the gap of understanding. In this literature we investigate different machine learning approaches and its techniques.

Keywords- Machine learning, Supervised learning, Un-supervised learning, Semi-supervised learning, Reinforcement Learning.

I. Introduction

Allan et.al in his research paper "Computing Machinery and Intelligence" [1] asked an important question "WHETHER THE MACHINE THINKS" can it may be replaced with "WHETHER MACHINES DO WHAT WE CAN DO". This proposal leads to notable definition of machine learning. Without programmed explicitly, machine learning is the area of study that helps the computer to learn automatically [2]. Machine learning is developed from the area of pattern recognition and artificial intelligence notably, machine learning it is the subfield of computer science [3]. Machine learning is associated with computational statistics and specialized in prediction making. The current research on machine learning focuses on natural language processing, computer vision, pattern recognition, cognitive computing and knowledge representation. Machine learning methods may be referred as predictive modeling when employed in industrial contexts [4]. The reminder of the literature is as follows: Section II, it introduces the machine learning approaches. Section III, concentrates on the supervised learning and its approaches, Section IV, depicts unsupervised learning and its techniques, Section V, focuses on semi-supervised learning and its concepts, Section VI, reviews reinforcement learning and its methods and we conclude with Section VII.

II. Machine Learning Approaches

It is the field through which the various computer algorithms are studied, that improves incrementally through the experience. Machine learning is classified in to supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [5]. Supervised learning is a machine learning task that assumes a function from the labeled training data. In un-supervised learning the data is not labeled, more precisely we have an unlabelled data. Semi-Supervised learning is a merger of labeled and unlabeled data. In reinforcement learning the software agent gathers from the interaction with the environment to take actions that would maximize the reward. The fig. 1 it depicts the classification of machine learning system.

III. Supervised Learning

Supervised learning is a machine learning task that assumes a function from the labeled training data. In supervised learning, there is an input variable (P) and output variable (Q). From the input variable, the function of the algorithm is to study the mapping function to the output variable Q= f(P). The goal of supervised learning is to analyze the training data that produces a complete function that can be utilized to map the new instances. The learning algorithm will be able to analyze and generalize the labels in the class correctly from the unobserved instances. This section introduces the various algorithms used in supervised learning.

III.I. Decision Trees. Decision tree [6] is termed as a directed tree structure, in which there are no incoming edges in the root node, while the remaining node consists of incoming edges. Each leaf node is available with a label; non leaf node is having a feature which is called feature set. Decision tree splits the data, which falls inside the non leaf node, according to the distinct values in the feature set. The testing of feature is operated from leaf node and its outcome is achieved until the leaf node is arrived. The optimal decision tree algorithm is beneficial to the limited problem. There is a need for heuristics methods for solving these problems. The heuristic methods can be solved either by bottom–up and top–down approaches. Examples of top–down decision trees includes ID3 [7], C4.5 [8], CART [9].

III.II. Rule Based Classifiers. Quinlan [1993] stated that, by transforming the decision tree in to distinct set of rules, a different path can be created for the set of rules. From root of the tree to the leaf of the tree, a distinct rule for each path is created; the decision tree is altered in to set of rules. Directly, from the training data the rules can be inducted by different algorithms that apply these rules. The idea is to construct the

978-1-7281-4862-5/20/$31.00 ©2020 IEEE

nominal set of rules that is familiar with the training data. The main goal is to construct the smallest set of rules that is similar with the training data.

independence among the features. Bayes theorem can be stated in mathematical terms:

$$P(X/Y) = P(X)(Y/X) / P(Y)$$

Where X and Y are events.

P(X) and P(Y) are events.

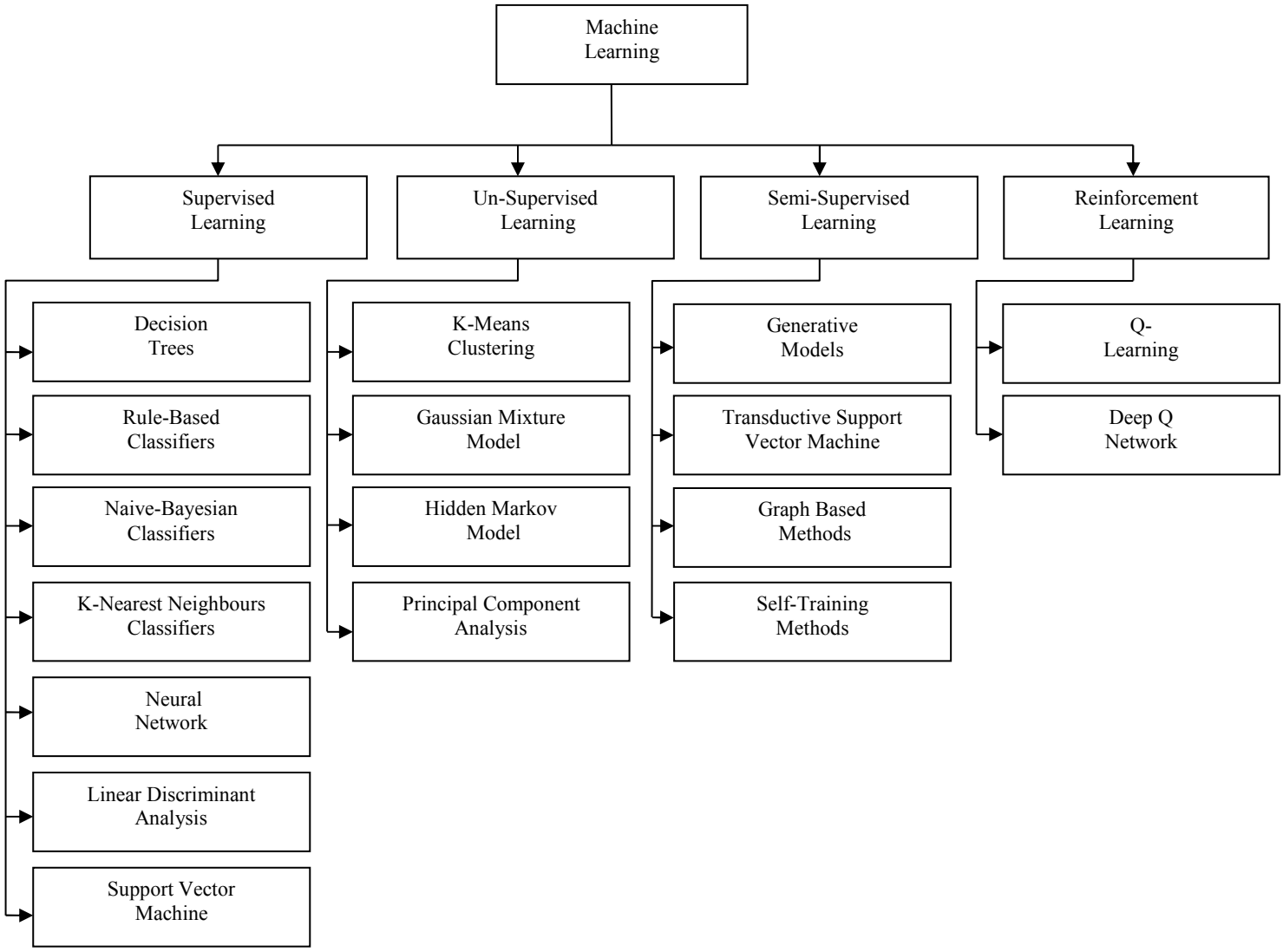

Fig: 1 Classification of Machine Learning Algorithms

RIPPER [10] is an algorithm that is based on rules. Through the process of imitated growing and pruning, it generates rules. For learning the set of rules the Genetic algorithms (GAs) [11] are also applied. Finding the quality chromosomes is the ultimate aim of Genetic Algorithm. The finesses of a chromosome are described in the Genetic Algorithm by the function known as fitness function [12].

III.III. Naïve-Bayesian classifier. Naive Bayesian classifiers [13] are probabilistic classifiers with their relation related to Bayes theorem having strong assumption of naive

P(X) and P(Y) are the prior probabilities of X and Y. P (X/Y) is a posterior probability, of the probability of observing the event X, given that Y is true. P (Y/X) is known as likelihood, the probability of observing the event Y, given that X is true. The advantage of the Naive Bayes classifier is the least computational time required for training the data.

III.IV. k- Nearest Neighbor classifiers. k-NN [14] is a nonparametric technique used for regression and classification. In the feature space the input of k-NN contains the k closest training examples. Then the output, it will depend

whether k-NN is applied for regression or classification purposes. In k-NN classification, the class membership is an output. With object allocated to its class with trivial among its k-NN (k being a positive integer, and small) the object is classified by its majority vote of its neighbors. When k = 1, then the object is allocated to that class having single nearest neighbor. The output is the property value for the object in k-NN regression.

III.V. Neural Network. The neural network conceptual model was proposed in 1943 by [15]. It consists of different cells. The cell receives data from other cells, processes the inputs, and passes the outputs to other cells. Since then, there was an intensive research to develop the ANNs. A perceptron [16] is a neural network that consists of a single neuron that can receive more than one input to produce a single output. To classify linearly separable classes, by finding m-dimensional hyper plane in the feature space that separates the instances of the two classes, perceptron are used. In Radial Basis Function RBF [17] a radial activation function is performed by every hidden unit, while the weighted sum of hidden output unit is performed by each output unit. It is commonly known as tri-layer feedback network.

III.VI. Linear Discriminant Analysis. A linear classifier [18] contains the vector, having weight w and bias having b. Given an instance p, the predicted class label q, is obtained according to:

$$Q = \text{sign}\,(w^T p + b)$$

With the help of weight vector w, the instance space is mapped onto a one-dimensional space, afterwards; to isolate the positive instances from negative instances, a point on the line is identified. A linear learning algorithm, which finds the best w and b for separating different classes, is Fisher's linear discriminant analysis [19]. Fisher's linear discriminant analysis it allows the instances of the same class to be adjacent, by keeping the variance of each class smaller, on the other hand it allows those instances having distinct class to be far, by accomplishing the distance between centers of distinct classes larger.

III.VII. Support Vector Machines. SVMs [20] revolve around the margin on either side of a hyperplane that separates two data classes. To reduce an upper bound on the generalization error, the main idea is to generate the largest available distance between its instance on either side and separating hyperplane. Finding an optimum hyperplane is the main idea of linearly separable data. The data points that lie on the margins of optimum hyperplane are termed as Support Vector Points, and it is characterized as the linear combination of these points. An alternative data point is neglected. The different features available on the training data do not affect the complexity of SVM. This is the primary reason the SVM are employed with learning tasks having substantial amount of features with respect to the number of training data.

Solving the n^{th} dimensional quadratic programming (QP) the training is performed on Support Vector Machines, where n represents the number of samples in the training data. Large problems of the SVM cannot not be solved as it may contain the large quadratic operations also there is numerical computation which makes the algorithm slow in terms of processing time. There is variation of SVM called Sequential Minimal Optimization (SMO). SMO can solve the SVM quadratic problem without employing additional matrix storage and without applying the optimization steps on the numerical quadratic programming [21].

IV. Un-Supervised Learning

In un-supervised learning the data is not labeled, more precisely we have an unlabelled data. In un-supervised learning we have the input variable (P), but there is no output variable. The representation is seen as a model of data. The aim of un-supervised learning is to discover the hidden structures from unlabelled data or to infer a model having the probability density of input data. This section investigates the basic algorithms used in un-supervised learning.

IV.I. K- Means Clustering. K-means algorithm [22]: The main idea of the algorithms is to partition the N observation in space in to K clusters. The information and nearest mean belongs to this cluster and works as model of the cluster. As a result the data space it splits in to Voronoi cells. K-means algorithm is an iterative method, which starts with a random selection of the k-means v_1, v_2… v_k. .With each number of iteration the data points are grouped in k-clusters, in keeping with the closest mean to each of the points, mean is then updated according to the points within the cluster. The grouping of data with regards to data points in accordance with the cluster means and updating the cluster means in accordance to set of points will continue until there is no change in the cluster means or points. The variant of K-means is termed as K-medoids. In K-medoids, instead of taking the mean the larger part of the cluster, having the centrally located data point is investigated as a reference point of the corresponding cluster [23].

IV.II. Gaussian Mixture Model. The Gaussian mixtures were popularized by Duda and Hart in their seminal text, Pattern Classification and Scene Analysis in 1973 [24]. A Gaussian Mixture is a function that consists of several Gaussians, each identified by $k \in \{1,\ldots, K\}$, where K represents the number of clusters of dataset. In Gaussian mixture model (GMM), the each Gaussian is characterized by the sequence of mean and variance that consists of mixture of M Gaussian distributions. Then the weight of each Gaussian will ultimately be the third parameter that is associated to each Gaussian distribution in a Gaussian mixture model. When clustering is performed using Gaussian Mixture Model, the goal is to find the criterion such

as mean and covariance of each distribution and the weights, so that the resulting model fits optimally in the data. In Gaussian Mixture Model, the likelihood of the data should be expanded so that the data can be optimally fitted. It can be obtained by applying iterative expectation maximization (EM) algorithm [25].

IV.III. Hidden Markov model. Hidden Markov Model (HMM) [26] is a parameterized distribution for sequences of observations. Basically, (HMM) is a Markov process that is divided in two components called observable components and unobservable or hidden components. That is, a hidden Markov model is a Markov process (Y_k, Z_k) $k \geq 0$ on the state space C $\times$ D, where we presume that we have a means of observing Y_k, but not Z_k as the signal process and C as the signal state space, while the observed component Y_k is called the observation process and D is the observation state space.

HMM, is sometimes called as a doubly stochastic process. Markovian stochastic process can be logically modeled by an HMM-based approach in which the actual states are not visited, these states are presumed to be unobserved or hidden; instead, the state can be observed that is stochastically dependent on the unobserved state.

IV.IV. Principal Component Analysis. PCA [27] is an analytical procedure that converts the correlated variables into linearly uncorrelated variables, with the help of an orthogonal transformation. This is named as principal components. The PCA is a multivariate dimensionality reduction tool that extracts the features representing most of the features in the given data and thus removing the least features having less information without losing the crucial information in data. When real data is collected, the random variables representing the data attributes are presumed to be highly correlated. The correlation between random variables can be found in the covariance matrix. The aggregate of the variances will give the overall variability.

V. Semi-Supervised Learning

Semi-Supervised learning is the sequence of labeled and unlabeled data. The labeled data is very sparse while there is an enormous amount of unlabelled data. The data is used to create an appropriate model of the data classification. The goal of semi-supervised learning is to classify the unlabelled data from the labeled data. This section explores some of the most familiar algorithms used in the Semi-Supervised learning.

V.I. Generative models. Generative model [28] considers a model p (u, v) = p (u/v) p (v) where p (u/v) is known as mixture distribution. The mixture components can be analyzed when there are large numbers of unlabelled data is available. The generative model is model of condition probability of the observable value X, given a value Y. Consider $\{P_\theta\}$ be a distribution family and is denoted by parameter vector θ. θ may be identified only if $\theta_1 \neq \theta_2 \Rightarrow y_{\theta 1} \neq y_{\theta 2}$ to the mixture components transformation. The expectation-maximization (EM) algorithm is applied on the multinomial mixture for the job of text classification [29].

V.II. Transductive Support Vector Machine: TSVM [30], it extends the Support Vector Machine (SVM) having the unlabelled data. The idea is to have the maximal margin among the labeled and unlabeled data on its linear boundary by labeling the unlabeled data. Unlabeled data has the least generalization error on a decision boundary. The linear boundary is put away from the dense region by the unlabeled data. With all the available estimation solutions to TSVM, it is curious to understand just how valuable TSVM will be a global optimum solution. Global optimal solution on small datasets is found in [31]. Overall an excellent accuracy is obtained on small dataset.

V.III. Graph based approaches. Graph structure, it defines the set of vertices V and set of edges E. More intuitionally, the structure can be defined as G = (V, E). Graph is created by the nodes and edges, the nodes it defines labeled and unlabeled patterns or samples, the edges it determines the affinity between labeled and unlabeled data. Labeling information of each pattern is proliferated to its adjoining pattern till the global optimum state is attained. The labeled data pattern is progressed to its adjoining points. The graph based techniques are focus of interests among researchers due to its better performance. Graph mincut problem is proposed by Blum et.al [33] in semi-supervised learning. A step Markov random walk is achieved on the graph by Szummer et.al [34].

V.IV. Self Training Methods. Self-training is a methodology applied in semi-supervised learning. On a small amount of data, the classifier is trained and then classifier is applied to classify the unlabeled data. The highest promising unlabeled points with its labels predicted are appended to training dataset. The classifier is again trained with the training dataset. This procedure goes on repeating itself. For teaching itself, classifier had its own predictions. This methodology is called bootstrapping or self-teaching [35]. Various natural language processing tasks applies the methodology of self teaching.

VI. Reinforcement Learning

In reinforcement learning the software agent gathers from the interaction with the environment to take actions that would maximize the reward. The environment is formulated as markov decision process. In reinforcement learning there is no availability of input/output variables. The software agent it receives the input i, the present state of environment s, then the software agent it determines an action a, to achieve the output. The values of state transition and the state of environment, which is changed by the action of the software agent is communicated through scalar reinforcement signal. After the action is chosen, reinforcement learning tells its

software agent to reward its subsequent state. The software agent is not told which action would be best in terms of long term interest. The software agent needs to gather information about the states, actions, transition, rewards for optimal working. This section reviews algorithms used in reinforcement learning.

VI.I. Q-Learning. Q-learning [36] is a type of model free reinforcement learning. It can also be known as an approach of asynchronous dynamic programming (DP). Q-learning allows the agents having the ability of learning to perform exemplary in markovian field by recognizing the effects of its actions, which is no longer required by them to build domain maps. Q-learning finds an optimal policy and it boosts the predicted value of the total reward, from beginning of the current state to any and all successive steps, for a finite markov decision process, given the infinite search time and a partially random policy. An optimal action-selection policy can be associated with Q-learning.

VI.II. Deep Q-Networks. (DQNs) [37] combines reinforcement learning with a deep network. Through a sequence of observations, actions and rewards, the DQNs consider a task in which the agent interacts with an environment. The main aim of the agent is to select actions in a manner that it augments the cumulative future reward. It applies the replay experiences that randomizes on top of the data, by eliminating correlations in the observation sequence and smoothing over changes in the data distribution. To reduce the correlations within the target, iterative update techniques are applied, so that the target values are periodically updated.

VII. Conclusion

In this study, various machine learning techniques and its approaches were analyzed. The classification of machine learning approaches such as supervised learning, un-supervised learning, semi-supervised learning and reinforcement learning and its various algorithms are the important contributions of this study. In future we intend to develop a model based on machine learning techniques.

References

[1] A.M Turing, “Computing Machinery and Intelligence”. Mind 49 pp: 433-460, 1950.

[2] Phil Simon, “Too Big to Ignore”: The Business Case for Big Data. Wiley. ISBN 978-1-118-63817-0, 2013.

[3] Mohssen Mohammed, Muhammad Badruddin Khan, “Machine Learning Algorithms and Applications”. CRC press Taylor and Francis Group, 2017.

[4] Chih-Fong Tsai, Yu-Feng Hsu, “Intrusion detection by machine learning: A review”. Expert Systems with applications. pp: 11994-12000. Elsevier, 2009.

[5] Myeongsu Kang, Noel Jordan Jameson, “Machine learning Fundamentals”. Prognostics and health management in electronics: Fundamentals, Machine Learning, and Internet of Things. Willey Online Library, 2018.

[6] Ross Quinlan, “Machine learning”. Vol.1 no.1, 1986.

[7] Quinlan, J.R, “Induction of Decision trees” Machine Learning. Vol. 1, Issue 1. pp: 81-106, Springer, 1986.

[8] Quinlan, J. R, “C4.5: Programs for Machine Learning”. Morgan Kaufmann Publishers, 1993.

[9] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J, “Classification and Regression Trees”, Wadsworth, Belmont, CA. Republished by CRC Press, 1984.

[10] William Cohen, “RIPPER Fast Effective Rule Induction”, Proceedings of the 12th International Conference on Machine Learning, 1995.

[11] Eiben A.E, "Genetic algorithms with multi-parent recombination". PPSN III: In Proc. International Conference on Evolutionary Computation. The Third Conference on Parallel Problem Solving from Nature: pp: 78–87, 1994.

[12] Colin R. Reeves, Jonathan E. Rowe, “Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory”. Kluwer Academic Publishers Norwell, MA, USA, 2002.

[13] Rish, Irina, “An empirical study of the naive Bayes classifier”. IJCAI Workshop on Empirical Methods in AI, 2001.

[14] Altman, N. S, "An introduction to kernel and nearest-neighbor nonparametric regression”. The American Statistician. 46 (3): 175–185, 1992.

[15] McCulloch, Warren; Walter Pitts, "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4): 115–133, 1943.

[16] Freund, Y.; Schapire, R. E, "Large margin classification using the perceptron algorithm". Machine Learning. 37 (3): 277–296, 1999.

[17] Buhmann, Martin Dietrich, “Radial basis functions: theory and implementations”. Cambridge University Press, 2003.

[18] Guo-Xun Yuan, Chia-Hua Ho, “Recent Advances of Large-Scale Linear Classification”. pp: 2584-2603, Proceedings of the IEEE, 2012.

[19] Fisher, R. A, "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics. 7 (2): 179–188, 1936.

[20] Cortes, Corinna, Vapnik, Vladimir N, "Support-vector networks". Machine Learning. 20 (3): 273–297, 1995.

[21] John C. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. pp: 61-74, Advances in large margin classifiers MIT press, 1999.

[22] MacQueen J. B, “Some Methods for classification and Analysis of Multivariate Observations”. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. pp: 281–297, University of California Press, 1967.

[23] Kaufman, L. and Rousseeuw, P.J, “Clustering by means of Medoids”, in Statistical Data Analysis Based on the Norm and Related Methods, pp: 405–416, North-Holland, 1987.

[24] Duda, R. O. and Hart, P. E, “Pattern Classification and Scene Analysis”. John Wiley and Sons, Inc, 1973.

[25] Dempster, A.P, Laird N.M, Rubin, D.B, "Maximum Likelihood from Incomplete Data via the EM Algorithm". pp: 1–38, Journal of the Royal Statistical Society, 1977.

[26] Baum, L. E.; Petrie, T, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". pp: 1554–1563, The Annals of Mathematical Statistics, 1966.

[27] Pearson, K., ″On Lines and Planes of Closest Fit to Systems of Points in Space". pp: 559–572, Philosophical Magazine, 1901.

[28] Ng, Andrew and Jordan, Michael, “On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes” Advances in Neural Info. Process system, 2002.

[29] Kamal Nigam, Andrew K Mccallum, "Text classification from labeled and unlabelled documents using EM". Vol. 39, pp: 103-134, Machine learning, Springer, 2000.
[30] Vapnik, V, Chervonenkis, A, "Theory of Pattern Recognition". Nauka, Moscow, 1974.
[31] Chapelle, O., Sindhwani, V., & Keerthi, S. S, "Branch and bound for semisupervised support vector machines". Advances in Neural Information Processing Systems (NIPS), 2006b.
[32] Xiaojin, In the "Semi-Supervised Learning Literature Survey". University of Wisconsin, Madison, 2005.
[33] Blum, A., & Chawla, S, "Learning from labeled and unlabeled data using graph mincuts". Proc. 18th International Conf. on Machine Learning, 2001.
[34] Szummer, M., & Jaakkola, T, "Partially labeled classification with Markov random walks". Advances in Neural Information Processing Systems, 2001.
[35] Chapelle Olivier, Schölkopf Bernhard, Zien, Alexander, "Semi-supervised learning", MIT Press, 2006.
[36] Christopher J. C. H. Watkins, Peter Dayan, "Q-Learning". Machine learning, Vol.8, pp: 279-292, Springer, 1989.
[37] Volodymyr Mnih Koray Kavukcuoglu, "Playing Atari with Deep Reinforcement Learning", Deep Mind Technologies. pp: 1-9, Toronto, 2013.