Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018)
IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN:978-1-5386-2842-3

A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification

R. Saravanan
Department of Computer Science
Pondicherry University
Puducherry, India
r.saravanan26@gmail.com

Pothula Sujatha
Department of Computer Science
Pondicherry University
Puducherry, India
spothula@gmail.com

Abstract— **Machine Learning (ML) is a kind of Artificial Intelligence (AI) technique which allows the system to obtain knowledge with no explicit programming. The main intention of ML technique is to enable the computers to learn with no human assistance. ML is mainly divided into three categories namely supervised, unsupervised and semi-supervised learning approaches. Supervised algorithms need humans to give input and required output, in addition to providing feedback about the prediction accuracy in the training process. Unsupervised learning approaches are contrast to supervised learning approaches where it does not require any training process. But, supervised learning approaches are simpler than unsupervised learning approaches. This paper reviews the supervised learning approaches which are widely used in data classification process. The techniques are reviewed on the basis of aim, methodology, advantages and disadvantages. Finally, the readers can get an overview of supervised ML approaches in terms of data classification.**

Keywords— Classification Problem, ML, Reinforcement Learning, Supervised Learning, Training Process.

I. INTRODUCTION

ML is a type of AI technique which makes the system to automatically obtain knowledge with no explicit programming [1]. It mainly concentrates on the designing of computer programs for accessing data and allowing it to learn by itself. This procedure starts from the inspection of data and search for patterns to achieve better decisions [2]. The main goal is to enable systems to learn automatically with no human intervention. It is not necessary to gain knowledge by consciousness but it identifies the patters present in the data. Using this way, ML algorithms eliminates the need of humans in the learning process. On the other hand, ML algorithms found to be quite difficult in diverse environments. The validation and computational investigation of ML method is an area of statistics called as computational learning theory. The process of ML is almost same as data mining and predictive modeling [3],[4]. These two processes recognize patterns in the data and modify the program actions. Numerous researches have been done that to recommend advertisements to the users based on their purchased products in e-shopping. It is due to the recommendation engines which employs ML to customize online ad delivery in real time scenario. Apart from customized marketing, some ML algorithms are used to detect frauds, spam filtering, detecting threats, predictive maintenance and developing newsfeeds [5],[6].

For instance, newsfeed in Facebook is customized to each individual user, when the particular user constantly stops at a post to read or share a friend post, then the newsfeed will begin to add more friend posts than the number of previous feeds. At the back end, the software program uses statistical and analytical tools to find patterns in the user data and utilizes the patterns to increase the newsfeed. When the user is not reading or liking or sharing the newsfeed, then the recent data is added in the data set and the newsfeed will be changed consequently.

On the basis of algorithm procedure, ML techniques can be classified to four types which include supervised ML techniques, unsupervised ML techniques, semi-supervised ML techniques and reinforcement ML techniques [7], [8].

- Supervised ML techniques utilizes from what it is has gained knowledge from the previous and present data with the help of labels to forecast events. This approach initiates from the training process of dataset, ML develop an inferred function to foresee the output values. The system is capable of providing results to an input data with adequate training process. ML algorithm compares the obtained results with the actual and expected results to identify errors to change the model based on results.
- Unsupervised ML techniques are employed when the training data is non-classified and not labeled. It analyses how the system can deduce a function to explain the hidden patterns from the unlabeled data. The system does not identify the proper output, but it discovers the data and writes observations from dataset to find hidden patterns from unlabeled data.
- Semi-supervised ML techniques lie between supervised and unsupervised ML techniques where it uses labeled and unlabeled data for training process.

Generally, it considers a smaller quantity of labeled data and a larger quantity of unlabeled data. These kinds of techniques can adjust itself to attain higher accuracy. It is preferable in case where the acquired labeled data need skillful and appropriate resources to train or learn from it. In rest of the cases, the obtained unlabeled data does not need extra resources.

- Reinforcement ML techniques interact with the environment by actions and locates errors or rewards. Trial and error search and delayed rewards are some of the common features of reinforcement method. It enables the systems and software programs to identify the ideal behavior in a specific context to increase the performance.

ML involves the investigation of large amount of data. It needs to provide precise results to find the profitable chances or hazardous risks and it is essential to take more time for proper training. The integration of ML with AI leads to efficiently process large amount of data. Though several types of ML techniques are available, supervised ML approaches are the most popular and commonly used technique. In this paper, the supervised learning techniques are mainly concentrated. The existing works of supervised ML classifiers are reviewed in this study. The techniques are reviewed on the basis of aim, methodology, advantages and disadvantages. Finally, the reviewed approaches are compared to each other another.

The residue of the paper is arranged as follows. The existing survey works on supervised ML classifiers are reviewed in Section 2. The paper is concluded in the last section.

II. Review of Supervised ML Approaches

In general, ML approaches are mainly divided into supervised and unsupervised ML techniques. The classification hierarchy of supervised ML techniques is given in Fig.1. The supervised ML techniques are classified into probabilistic classifiers and linear classifiers.

A. Probablistic classifiers

This classifier uses mixture models to classify data [9]. It considers every class is an element of the mixture. Every mixture element is a generic model which gives the possibility to sample a specific term of the element. This kind of classifier is termed as generative classifiers. The popular kinds of probabilistic classifiers are explained below.

1) Naive Bayes Classifier (NB): NB classifier is a simpler and widely used classifier for data classification. It calculates the posterior possibility of a class using the distribution of words in a document [10]. It basically follows the bag of words (BOW) feature extraction to eliminate the word location in the document. It employs Bayes Theorem to identify the possibility of a feature set whether it fits into a specific label.

$$P(l/f) = \frac{P(l) * P(f/l)}{P(f)} \quad (1)$$

where P (l) is the previous likelihood of a label, P(f/l) is the previous likelihood prior possibility that a given feature set is classified as label. P(f) refers to previous possibility that a given feature set appeared. NB stated that every individual features are not dependent and is represented in Eq. (2)

$$P(l/f) = \frac{P(l) * P(f1/l) * \ldots * P(fn/l)}{P(f)} \quad (2)$$

2) Bayesian Network (BN): The important characteristic of BN classifier is the non-dependency of the features [11], [12]. Another consideration is the dependency of all features. It results the BN model to a directed acyclic graph where the nodes and edges indicates random variables and conditional dependencies. It is assumed as an entire model for the variable and the relation between them. As a result, a total joint probability distribution (JPD) for all the variables is defined for a model. As the BN classifier is computationally complex, it is not recommended to mine text [13].

3) Maximum Entropy Classifier (ME): ME classifier translates the labeled feature sets to vectors with the help of encoding. These vectors are employed to determine weights of every feature which can be integrated to calculate the most probable label for a feature set. ME classifier is parameterized by a collection of X {weights}, it combines the joint features which are created from a feature-set by an X{encoding}. Particularly, the encoding matches every C {(feature set, label)} pair to a vector. The possibility of every label is calculated as

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)} \quad (3)$$

It is employed in Kaufmann [14] to find parallel statements among any language pairs with smaller quantity to train data. Some other techniques are also proposed to filter parallel statements in an automatic manner from non-parallel corpora which utilizes larger amount of training data. The performance analysis depicted that ME classifier attains better results for all language pairs. It allows creating parallel corpora for various new languages.

B. Linear classifiers

Linear classifier is aimed to cluster the items into groups which have same feature values. Timothy et al. [15] defined that the linear classifier attained the classification decision using the value of linear combination of features. For an input real vector, the output value will be

$$y = f(\vec{w} \cdot \vec{x}) = f\left(\sum_j w_j x_j\right) \quad (4)$$

where $\vec{w}$ represents real vector, f is a function which transforms the (.) of two vectors to the required output.

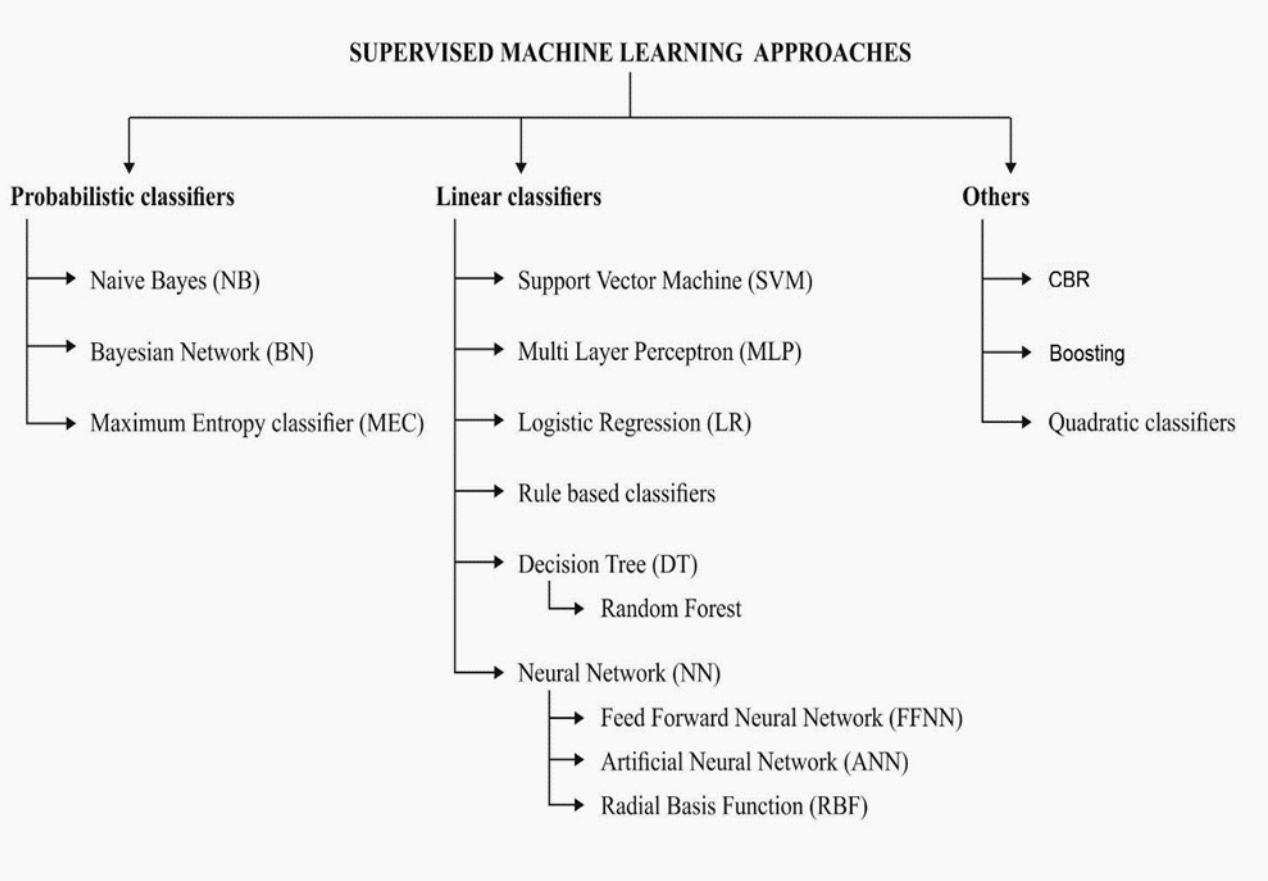

Fig.1. Classification of supervised ML approaches

Here, f is a simple function which assigns every value above a particular threshold to the first class and the rest of the values to second class. A more complex f may have the possibility of an item to a certain class.

1) Support Vector Machine (SVM): SVM is introduced in [16], which classifies data by building an n dimension hyper plane which finely partitions to two types. SVM is highly correlated to NN. Generally, a SVM model with a sigmoid kernel function is same as to a 2layer, Perceptron NN. These models are closer to traditional MLP. With an additional kernel function, SVM can acts as a substitute to train polynomial RBF and MLP. The weights of the network are determined when the quadratic programming problem is solved under linear limitations instead to solve a non-convex, unconstrained minimization problem in default NN training process.

2) Multilayer Perceptron (MLP): MLP [17] is one of the commonly used classifier which can be activated by passing the input layer with input vector and transmitting the actions in a feed forward manner via weighted connections in the whole network. For an inputw_1, the state of i^{th} neuron (s_i) is equated as

$$s_i = f\left(w_{i,0} + \sum_{j \in P_i} w_{i,j} \times s_j\right) \quad (5)$$

where $\boldsymbol{f}$ denotes the activation function, $\boldsymbol{P}_{\boldsymbol{i}}$is the collection of nodes reaches node $\boldsymbol{i}$, $\boldsymbol{w_{i,j}}$ is the weight of the link from node $\boldsymbol{i}$ and $\boldsymbol{j}$. MLP employs a repeated process to learn data which starts with random weights. A training method is also utilized to manage the weights to a specified target values. The training will have ended only the error slope attains a value of zero.

3) Logistic Regression (LR): LR [18] is a commonly used statistical technique employed for application scoring and credit risk modelling. Some research works proved that these methods are inefficient and inaccurate in comparison with advanced ML techniques. Logistic distribution gives the basis of the logit model with its distribution function as represented in Eq. (6)

$$F(X_i\beta) = \frac{\exp(X_i\beta)}{1+exp(X,\beta)} \quad (6)$$

$$F(X_i\beta) = \frac{\exp(X_i\beta)}{[1+exp(X,\beta)]^2} \quad (7)$$

*4) Rule-based classifier:*This kind of classifiers models the data space to a collection of rules. The left side indicates the

state of the feature set represented in disjunctive normal form and the right side indicates the class label. The constraints are on the term presence. Term absence is sparsely utilized due to the non-informative nature in sparse data. Various regulations are available to create rules; the training phase builds the rules based on the regulations. The two regulations are support and confidence [19]. The support indicates the actual number of instances in the training data set which are appropriate to the rule. The confidence denotes the conditional possibility that the right hand side of the rule is fulfilled when the left-hand side is fulfilled. Few combined rule algorithms are developed in [20].

5) Decision tree (DT) classifiers: DT classifier provides a hierarchical partition of training data space where the constraint on the attribute value is utilized to split data [21]. The division of the data space is performed iteratively till the leaf nodes hold a particular number of fewer records which can be utilized for classification purposes.

6) Random Forest (RF): RF [22] method is highly applicable to various classification problems. It is a ML approach incorporates the processes to aggregate data, bagging and DT models using subsets. It selects a subset of features from individual node of the tree with the avoidance of correlation in the bootstrapped set. For a task to classify n companies $X_j\,(j = 1, n)$ and p ratios, RF is a forest of k trees and is computed as

$$RF = \{DT_j\}, j = 1, k \qquad (8)$$

The classification capacity of RF is computed by out of bag (OOB) classification error.

7) Neural Network (NN): NN contains numerous neurons where each neuron is a fundamental unit. The input to the neuron is indicated by a vector overline X_i, represents the frequency of words in the i^{th} document. A group of weights 'A' are integrated to every neuron for the purpose of computing a function of its inputs f(.) [23].

Radial Basis Function (RBF)

RBF [24] is also a popular ML method which has the advantage of quicker learning capacity. It holds a hidden layer and MLP has many hidden layers. The training process of RBF is very simple. It uses Gaussian function as activation function and least-square criteria as the objective function. It comprises of an input, hidden andoutput layer. The input layer measures the norm of the input from the neuron. The output of the n^{th} output neuron is equated as

$$y_n = \sum_{k=1}^{K} c g_k(x) = \sum_{k=1}^{K} w_{kn} \exp\left(\frac{-\|x-\mu_k\|^2}{2{\sigma_k}^2}\right) \qquad (9)$$

where K indicates the total number of Gaussian neurons, w_{kj}denotes the weight fromk^{th}Gaussian neuron toj^{th} output neuron, μ_k centre location, $\sigma_k$width k^{th} neuron and x is the input vector.

C) Other classifiers

1) Case Based Reasoning (CBR): CBR [25] provides solution to some of the earlier problems. For instance, a car mechanic repairs an engine by remembering another car which shows same symptoms using CBR. It can be used for computer solving problems using a 4 step process namely Retrieving, Reusing, Revising and Retaining. It can be observed in rule induction approaches of ML. Similar to rule-induction methods, CBR begins from a group of cases or training examples by finding the similarities among the remembered case and target problem.

2) Boosting: Boosting is a ML based meta-heuristic algorithm used to reduce bias and variance in supervised learning [26], and a family of ML approaches, transforms the weak learner to stronger learner. It basically depends on the question posed by [27-28], whether it is possible to generate a strong learner from a group of weak learners or not. A weak learner is less correlated than the strong learner with the true classification. On the contrary, strong learner is a classifier which is well-correlated with the true classification. Many boosting methods comprises of repeatedly weak classifiers with respect to a distribution and appending them to a complete strong classifier. The upcoming weak learners can concentrate more on the illustrations that the past weak learners which fail to classify it correctly.

3) Quadratic classifier: A quadratic classifier is employed in ML and statistical analysis classification technique to split measurements of two or many classes of objects or events by a quadric surface [29]. It is almost same as a linear classifier.

III. Conclusions

Supervised algorithms need humans to give input and required output, in addition to providing feedback about the prediction accuracy in the training process. Unsupervised learning approaches are contrast to supervised learning approaches where it does not require any training. But, supervised learning approaches are simpler than unsupervised learning approaches. This paper reviews the supervised learning approaches which are widely used in data classification process. The techniques are reviewed on the basis of aim, methodology, advantages and disadvantages. In future, supervised ML approaches can be utilized to design an efficient classification model for various real time applications.

References

[1] A. Holzinger, H. Plass, K. Holzinger, G.C. Crisan, C.M. Pintea, C. M., and V. Palade, "A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop," arXiv preprint arXiv:1708.01104.

[2] I.H.Witten, E. Frank, M.A. Hall, and C.J. Pal, "Data Mining: Practical machine learning tools and techniques," Morgan Kaufmann. 2016.

[3] R. Kohavi, "Glossary of terms," Machine Learning, vol. 30, pp. 271-274, 1998.

[4] H. Mannila, "Data mining: machine learning, statistics, and databases,"Scientific and Statistical Database Systems,Proceedings., Eighth International Conference on. IEEE, 1996.

[5] P. Ongsulee and Pariwat, "Artificial intelligence, machine learning and deep learning," ICT and Knowledge Engineering (ICT&KE), 2017 15th International Conference on. IEEE, 2017.

[6] J. H. Friedman, "Data mining and statistics: What's the connection?," Computing Science and Statistics, vol. 29, pp. 3-9, 1998.

[7] A.K. Jain, M.N. Murty and P.Flynn, "Data clustering: a review," ACM Comput Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[8] M. Dutton and G.V. Conroy, "A review of machine learning," The Knowledge Engineering Review, vol. 12, pp. 341-367, 1997.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "The elements of statistical learning," New York: Springer series in statistics, vol. 1, pp. 337-387, 2001.

[10] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001.

[11] N. Friedman, D. Geiger, and M. Goldszmidt. "Bayesian network classifiers," Machine learning, vol. 29, pp. 131-163, 1997.

[12] F. Jensen, "An introduction to Bayesian networks". Vol. 210. London: UCL press, 1996.

[13] C.C. Aggarwal, and C. Zhai, "A survey of text clustering algorithms," In Mining Text Data, pp. 77-128, 2012.

[14] M. Kaufmann, "JMaxAlign: A maximum entropy parallel sentence alignment tool," Proceedings of COLING 2012: Demonstration Papers, pp. 277-288, 2012.

[15] Shepard, P. J. "Decision Fusion Using a Multi-Linear Classifier," Proceedings of the International Conference on Multisource-Multisensor Information Fusion, 1998.

[16] L. Gonzalez, C. Angulo, F. Velasco, and A. Catala, "Unified dual for bi-class SVM approaches," Pattern Recognition, vol. 38, no. 10, pp.1772-4, 2005.

[17] B. Back, T. Laitinen, and K. Sere, "Neural networks and genetic algorithms for bankruptcy prediction," Expert Systems with Applications, vol. 1, pp. 407-413, 1996.

[18] S.S.Haykin, "Neural networks and learning machines," Vol. 3. Upper Saddle River, NJ, USA:: Pearson, 2009.

[19] B.L. Ma and B. Liu, "Integrating classification and association rule mining," Proceedings of the fourth international conference on knowledge discovery and data mining, 1998.

[20] W. Medhat, A. H. Yousef, and H. K. Mohamed, "Combined algorithm for data mining using association rules," arXiv preprint arXiv:1410.1343 , 2014.

[21] J. Quinlan, "Induction of decision trees," Machine learning, vol. 1.1, pp. 81-106, 1986.

[22] L. Breiman, "Random Forests Machine Learning," View Article PubMed/NCBI Google Scholar, pp. 45-32, 2001.

[23] M.E. Ruiz and P Srinivasan, "Hierarchical neural networks for text categorization," Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

[24] J. Moody and C.J. Darken, ""Fast learning in networks of locally-tuned processing units," Neural computation, vol. 1, pp. 281-294, 1989.

[25] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," AI communications, vol. 7, pp. 39-59, 1994.

[26] Z.H. Zhou, "Ensemble Methods: Foundations and Algorithms," Chapman Hall/CRC Data Mining and Knowledge Discovery Series, 2012.

[27] M. Kearns, "Thoughts on hypothesis boosting," Unpublished manuscript, vol. 45, 1988.

[28] M. Kearns and L. Valiant, "Cryptographic limitations on learning Boolean formulae and finite automata," Journal of the ACM (JACM), vol. 41, pp. 67-95, 1994.

[29] T.M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,"IEEE transactions on electronic computers, vol. 3, pp. 326-334, 1965.