Machine Learning Review and Future Perspectives

Naina Handa

Assistant Professor

DAV College, Amritsar

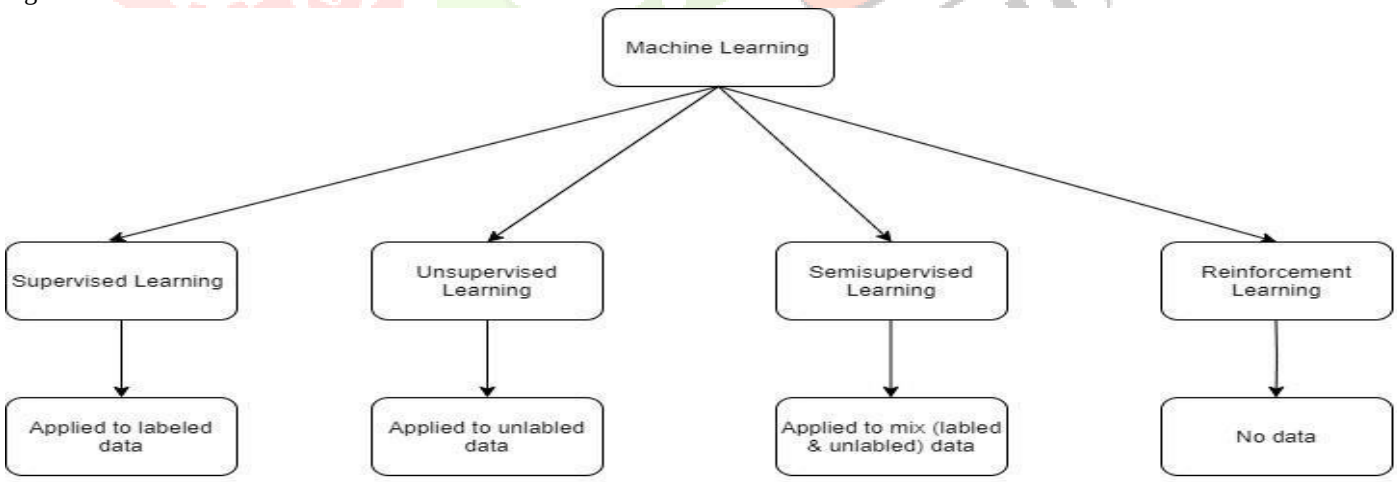
Abstract

Machine learning has gained much prominence in today's world, and its techniques are used in all fields such as pattern recognition, object detection, text analysis, software engineering, natural language processing and various research areas. Machine learning, a part of AI (artificial intelligence), is used in algorithm design, based on recent data trends. This paper aims to introduce the machine learning algorithms, its principles and to highlight the advantages and disadvantages in this area. So, this paper generally produces the research performed by the authors in the area of machine learning and its applications and draws attention to the scholars working in this field. Machine learning algorithms have helped to solve complex domain problems in different engineering fields in recent years.

Introduction

Machine learning (ML) is an important field of today's world of science that carries out a task without explicitly programmed. Machine learning creates a function, model or algorithm; to learn and make extractions on the basis of existing datasets known as training datasets [1]. Training dataset acts as an input to an already known output. Training datasets are studied under supervised learning to extraction future events when historical data is available [2]. The most popular and efficient supervised learning algorithms are support vector machine, J48, and random forest to name a few [31]. Although unsupervised learning utilizes unlabeled input data and considers input data the secret patterns. The relevant unsupervised learning techniques are k mean clustering and artificial neural networking etc. The various techniques for machine learning are as follows:

Fig 1



Different Machine Learning techniques and type of data handled by them

J48

J48 is a decision tree in which a number of principles characteristics are taken into account in the training dataset. IT is the implementation of an algorithm for classification and data determination based on the attribution values of the data provided. The possible values of the samples observed are calculated by varied

interior nodes. J48 categorizes data on a Hierarchical basis. The structured information is split sequentially to create more leaf nodes.it effectively decreases the waiting time for the sorted elements and with minimum tree approach the value is created more. J48's weakness is its requirement that data be sorted as a preprocess [3].

SVM (Support Vector Machine)

SVM also falls within the classification category to be used for binary classification. This approach takes advantage of the geometric characteristics of the given training dataset rather than input space specifications. The efficacy of the SVM lies in choosing and using different kernels [4] . It also effectively recognizes soft margin parameters. This approach to machine learning can pretty reliably model complex non-linear decisions. Yet algorithmic complexity is becoming one of their significant limitations. Another weakness is kernel size [5]. SVM is a type of supervised learning system used to analyze data and to analyze classification and regression. SVM is best used to classify data. In SVM the key concept is to construct a hyper plane in an infinite and high-dimensional space. Classification trees are used to classify what class the data belongs to. Regression trees are used to estimate destination variable value.

Random Forest

From arrangement of decision trees, Random forest arbitrarily chooses a subset of a preparation set .summarizing the votes from various choice trees, it chooses the last class of the test object.it is the aftereffect of a developing gathering of trees.it can be said to be a mix of tree extractionors. Where in each tree relies upon the estimations of arbitrary vector which is freely inspected alongside the comparable circulation for all trees in the backwoods [6]. The quality of this model is relying upon the individual tree in the forest and their relationship with one another. An outcome additionally gives precision and significance of variable data. Irregular forest is less inclined to commotion.

Decision Tree

From arrangement of decision trees, Random forest arbitrarily chooses a subset of a preparation set .Summarizing the votes from various choice trees, it chooses the last class of the test object.it is the aftereffect of a developing gathering of trees.it can be said to be a mix of tree extractionors. Where in each tree relies upon the estimations of arbitrary vector which is freely inspected alongside the comparable circulation for all trees in the forest. The quality of this model is relying upon the individual tree in the woodland and their relationship with one another. An outcome additionally gives precision and significance of variable data.

Artificial neural network (ANN)

ANN is an approach to self-learning which imitates the behaviors of biological nervous systems. This is a process that advances its parameters in the light of external or internal data moving in the learning stage through the program[14].

Ensembling

Ensembling involves the use of several clustering algorithm system.it is very successful and flexible in solving all domain problems. Ensemble approach uses any machine learning algorithm as its base model by breaking the same training dataset and algorithm differently or using algorithms on the same dataset.it is a mix that combines several models to construct an extraction model. Ensemble approach seeks to use multiple methods and models to achieve better performance [7]. A plethora of terminology such as grouping, mixture, group, mixture and aggregation are used to describe the term. The types of Ensembling are given below.

Voting

Voting is being used to determine average and its classification using regression / identification models on other datasets.it could be used to track majority voting, weighted average or simple average by using different splits of the same dataset with the same algorithm or with different algorithms [8].

Stacking

Different types of machine learning combine to construct multilevel classification / regression types. First classifier output will serve as the input training that is assigned to the next classifier. It's used to explore various possibilities for the same question. Stack seeks to boost overall efficiency and is going to be the best approach than the intermediate model[9].

Bagging

The random subsamples of the datasets are created with the aid of bootstrap aggregation. This technique of machine learning ensemble not only aims at finding statistical categorization and regression, but is also used to improve performance and precision. Through this vote, it is possible to do generation and training through parallel.

Boosting

Weak models are transformed to strong models using a family of boosting algorithms. Boosting creates an ensemble that is in essence sequential and so can not be used for concurrent operations. Project is also known as stage wise additive modeling with the same dataset. With this model.it, binary classification as well as multi-class issues are solved, is better than bagging in producing noise-free results. Boosting can be implemented by either weighing or sampling.it does not require the poor learners 'prior knowledge[10].

Feature selection

The process of selecting a specific subset with the necessary variables or extractors to create a model is known as the feature selection. It aims at improving accuracy efficiently, enabling machine learning algorithms to train faster, reduce complexity and reduce overfitting, in addition to extracting data sets [10]. Feature filtering technique is used to eliminate redundant or obsolete functions, and it is used in complex domains with comparatively few data samples in order to minimize information.it loss. The different classes of feature selection algorithm are process, wrapper, and embedded process of filtering.

Filter method

A pre-processing phase filter approach selects features independent of any algorithm for the machine learning. In each function a statistical factor is assigned to the scoring. Feature selection is not affected by the classifier character used. The score ranking is performed according to the appropriate approach [11].

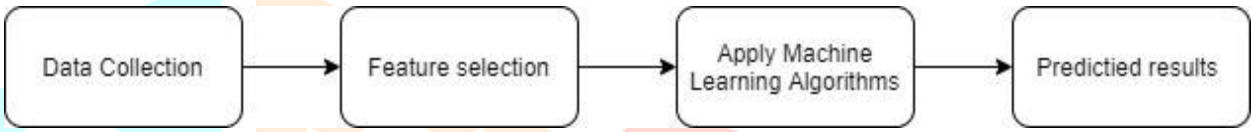
Wrapper method

In comparison to the filter method, the wrapper method uses a classifier to pick features.it takes higher computing resources and is therefore costly. Wrapper approaches are important examples of recursive elimination of features, forward selection of features and backward elimination of features [12].

Embedded method

The filter and wrapper method combined is known as embedded method. The most popular forms of embedded approach are methods of regularisation or penalization. In feature selection methods this extraction method has its own built-in. In embedded system the validity and relative capacity of the methods in the given scenario cannot be expressed. The LASSO, Elastic Net, and Ridger Regression are regularized algorithms [13].

Machine learning investigates the analysis and design of algorithms from which data can be learned and projected. The phases of apply Machine Learning are as follow:



Machine Learning Phases

The Machine Learning process can be divided into different phases like Data collection, Feature Selection, Apply Machine Learning algorithms and finally get the predicted results.

Machine learning focuses on prediction, based on the knowledge of the properties learned from the training data as:

Comparison of Machine Learning algorithms

Machine Learning Method	Advantages	Disadvantages
Support Vector Machine	Good performance, repeatable training process	Supervised instruction, depending upon parameter.
Artificial Neural Network	Robustness and broad applicability	Initial value, Long Overtraining time and Contingent training.
Random Forest	Rebellion to over-training, Predictability increases	Basically discreet, large numbers of trees will slow the algorithm to predict in real time
Decision Tree	Small complexity	Precision depends on the nature of the traits and the tree

Merugu and Akepogu addressed the importance of NFRs for the creation of critical applications. The scientists expected a four-layer research system for detecting NFRs and suggested some guidance for individual layers as well. The authors have done case studies on the network of ATM and Online Library. The checklist has been used for process elicitation [14].
Iqbal and Babar addressed the significance of SOA Service Oriented Architecture. It also focused as a main issue on price. The authors recommended the creation of a specific quality model for service-oriented architecture, based on the latest ISO / IEC 25010[15].

Ameller et al. discussed how difficult the NFRs have been for software engineers for several years, though for a long time different methods and techniques have been proposed to enhance elicitation, tracking, and testing. Authors suggested that learning more about these issues in their everyday routine practice, i.e. clinicians and researchers, would be of interest to both parties. The empirical analysis was performed in paper [16].

Ezami has developed a classification model for derived NFRs that could be used in the source code and source code remarks. The databases were derived from the Electronic Health Records (EHR) database. The maximum execution was achieved using the SVM classifier, with an F1 ratio of 0.86. The findings show that the supervised approach outperforms the unsupervised technique by which NFRs are extracted from statements. It's also found that bag-of-words is better than doc2vec [17].

Martino et al. have discussed that the Requirement Design in cloud computing is a very complicated job, particularly when stakeholders need to advance. An automated framework is designed for modeling and classifies specifications set out in Natural Language to facilitate the development of cloud applications. The goal datasets were taken from PROMISE repository [18].

Kiran and Ali investigated that for Open Source Systems the Requirement elicitation procedure is extremely unpredictable and critical. The various ways for extracting and simplifying the elicitation process have been explored. The tools, techniques, and methods available regarding Open Source Software have been detailed in the paper [19].

Conclusion

Through the findings of this study, the use of machine learning techniques for multiple domains is apparent. Given the rising pattern, it is an evolving area of study.The study and comparison of different Machine Learning methods has discussed in paper along with their benefits and shortcoming. The adoption of machine learning methods may aid traditional and advance data analysis. Machine Learning is always an advance automatic technique for solving different problems in different domains.

References

1. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
2. Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2020). Machine learning for fluid mechanics. Annual Review of Fluid Mechanics, 52, 477-508.
3. Aljawarneh, S., Yassein, M. B., & Aljundi, M. (2019). An enhanced J48 classification algorithm for the anomaly intrusion detection systems. Cluster Computing, 22(5), 10549-10565.
4. Karimi, F., Sultana, S., Babakan, A. S., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. Computers, Environment and Urban Systems, 75, 61-75.
5. Zhang, D. (2019). Support vector machine. In Fundamentals of Image Data Mining (pp. 179-205). Springer, Cham.
6. Reis, I., & Baron, D. (2019). PRF: Probabilistic Random Forest. Astrophysics Source Code Library.
7. Desai, S., Sinno, B., Rosenfeld, A., & Li, J. J. (2019). Adaptive Ensembling: Unsupervised Domain Adaptation for Political Document Analysis. arXiv preprint arXiv:1910.12698.
8. Venosa, P., García, S., & Díaz, F. J. (2019). Ensembling to improve infected hosts detection. In XXV Congreso Argentino de Ciencias de la Computación (CACIC)(Universidad Nacional de Río Cuarto, Córdoba, 14 al 18 de octubre de 2019).
9. Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. Information Fusion, 52, 1-12.
10. Khulenjani, N. B., & Abadeh, M. S. (2020). A Hybrid Feature Selection and Deep Learning Algorithm for Cancer Disease Classification. International Journal of Computer and Information Engineering, 14(2), 55-59.
11. Chen, Z., Pang, M., Zhao, Z., Li, S., Miao, R., Zhang, Y., ... & Huang, L. (2020). Feature selection may improve deep neural networks for the bioinformatics problems. Bioinformatics, 36(5), 1542-1552.
12. Chen, H., Li, T., Fan, X., & Luo, C. (2019). Feature selection for imbalanced data based on neighborhood rough sets. Information Sciences, 483, 1-20.
13. Sen, R., Mandal, A. K., Goswami, S., & Chakraborty, B. (2019, October). A Comparative Study of the Stability of Filter based Feature Selection Algorithms. In 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) (pp. 1-6). IEEE.
14. Merugu, G., & Akepogu, A. (2017). Four Layered Approach to Non-Functional Requirements Analysis. IJCSI International Journal of Computer Science Issues, Retrieved, 4.
15. Iqbal, H., & Babar, M. (2016). An approach for analyzing ISO/IEC 25010 product quality requirements based on fuzzy logic and Likert scale for decision support systems. International Journal of Advanced Computer Science and Applications, 7(12), 245-260.
16. Ameller, D., Franch, X., Gómez, C., Martínez-Fernández, S., Araújo, J., Biffl, S., ... & Muccini, H. (2019). Dealing with non-functional requirements in model-driven development: A survey. IEEE Transactions on Software Engineering.
17. Ezami, S. (2018). Extracting non-functional requirements from unstructured text (Master's thesis, University of Waterloo).
18. Di Martino, B., Pascarella, J., Nacchia, S., Maisto, S. A., Iannucci, P., & Cerri, F. (2018, May). Cloud Services Categories Identification from Requirements Specifications. In 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA) (pp. 436441). IEEE
19. Kiran, H. M., & Ali, Z. (2018). Requirement Elicitation Techniques for Open Source Systems: A Review. International Journal of Advanced Computer Science and Applications, Pakistan, 330-334.