Review on Studies of Machine Learning Algorithms

Peiyuan Xu
Chengdu Foreign Languages School, 611731 Chengdu, China

Abstract. This paper mainly introduces machine learning algorithms in the field of artificial intelligence. First, it describes the classification of such algorithms and their main application scenarios. Then the paper introduces the principles behind those algorithms and presents the author’s views. Finally, the development trend of machine learning algorithms is envisioned.

1. Introduction

1.1 Current Situation of Machine Learning Development
In recent years, machine learning has become one of popular research fields in the study of artificial intelligence. Currently, well-known Internet companies are researching and applying technologies related to artificial intelligence far and wide, among which machine learning is also a hot subject including image and speech recognition. In particular, against the backdrop of big data developing rapidly, machine learning blended with big data can effectively combine systems and algorithms, so that machine learning algorithms can work concurrently under multiple cores and process mass data, which is also the current research direction in the study of artificial intelligence. The rapid development of the Internet industry not only provides a large number of samples to train deep learning, continuously undergoing breakthroughs, improvements, and innovation by the joint effort from various fields.

2. Development of Machine Learning

2.1 Developing Process of Machine Learning [1]
Machine learning is not an emerging concept, whose prototype appeared as early as the mid-1950s to the mid-1960s instead. The main research goal at the time was to realize various self-organizing and adaptive systems. In other words, it could also be, to some extent, considered as an unsupervised mode of machine learning, the idea of which was sought after by many scholars.

By the middle of the 1970s, research on machine learning was centered on the process of concept learning. The tools used at that time were semantic networks and predicate logic, and scholars’ enthusiasm also slackened off.

After another 10 years or so, concept learning remained as the research goal. However, studies were extended from learning a single concept to multiple concepts, exploring various learning strategies and learning methods, namely, progress was made both in breadth and in depth. Then the development of machine learning entered a new phase in the 1980s, which could be said to be a period witnessing vigorous development in this field. By far, the recovery of neural networks has driven the research into various machine learning methods and systems, accompanied by the shift from theoretical studies to practical applications in the field of machine learning. In the future, it is believed
that the development of machine learning will be of greater significance. As big data further develops, machine learning will find wider application.

3. Machine Learning Algorithms

3.1 The Classification of Machine Learning Algorithms [2]

3.1.1 Supervised Learning
Supervised learning is relatively classical among machine learning algorithms, which includes logistic regression and reverse neural networks. Typical methods are BN, SVM, KNN, CBR, and so on. As for supervised learning, the input data is called training dataset, the basis function models include algebraic function or probability function, the way of training the model is iterative calculation, and the result of the training dataset is known to the functions. This algorithm works in this way: first computational predictions are made based on the training dataset, and then prediction training is continuously iterated until the results match the known ones. Therefore, supervised learning is often used for classification and regression.

3.1.2 Unsupervised Learning
The characteristic of unsupervised learning lies in that the result value of input data is not pre-set, that is, results produced by the algorithm are uncertain. Unsupervised learning can be roughly classified into clustering, anomaly detection, and competitive learning. For example, clustering refers to summarizing the structure and numerical values of the data, and then classifying the results. Typical unsupervised learning algorithms include K-means algorithm, Apriori algorithm, and SOM algorithm.

3.1.3 Semi-supervised Learning
Semi-supervised learning has wider application in practical cases because it is more efficient than other machine learning algorithms. Its principle lie in mixing data that has been labeled or has definite results with data that has no pre-set results, the goal of which is to learn the relationship among various attributes and output the classification model for predictive analysis. The algorithm is applied in both classification and regression problems.

3.1.4 Reinforcement Learning
Reinforcement learning, which is a learning method based on statistical and dynamic programming, uses environmental feedback as input data, not reinforcing the learning process. It is mainly used in problems related to control precision of robots, and the mainstream algorithms include Q-Learning and the temporal difference learning algorithm.

3.2 Introduction to Mainstream Supervised Learning Algorithms

3.2.1 Decision Tree
Here the decision tree in the field of supervised learning algorithms is specifically introduced. The decision tree is a tree-like prediction model in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Its structure can help restore and understand the decision-making process of a given problem. Classical decision tree algorithms include ID3 and C4.5, which can solve optimal problems and multi-stage decision problems.

3.2.2 Random Forests
Based on Bagging, random forest is an important ensemble learning method for classification, regression and other tasks, which is operate by inputting unlabeled samples and outputting the results of classification voted by individual trees. The random forest adds the bagging technique to the decision tree, which enables it to solve the performance bottleneck faced by the decision tree. Its
introduction of randomness optimizes the anti-noise ability and reduces the risk of over-fitting, proving good scalability and parallelism in the classification of high-dimensional data. Random forests can handle discrete and regression problems well, whose input dataset does not need to be normalized.

3.2.3 Naive Bayes
In the early period from 1950s to 1960s, Bayesian School came into form. The Bayesian classification method, one of the statistical classification methods, can be used to predict the probability featuring a membership relationship and predict the probability of a given classification. Bayesian algorithm is a general term for a class of algorithms, among which naive Bayes is a simpler and more typical one. It is simply embodied in the requirement that the naive Bayes algorithm satisfy the conditional independence assumption, namely, the effect of each attribute on its target variable of the given classification is independent.

The naive Bayes algorithm utilizes the Bayes Theorem:

\[ P(H|X) = \frac{P(H|X)P(H)}{P(X)} \]

3.2.4 Support Vector Machines (SVM)
The SVM algorithm was quite popular for a time, which maps the input space to a high-dimensional feature space by nonlinear transformation and finds the optimal linear boundary hyperplane in the high-dimensional special space. When the given dataset is linearly separable, the working principle of the support vector machine classification algorithm requires that the empirical risk be minimized. Therefore, in order to find an optimal boundary plane that can not only correctly separate the data of the two categories but also maximize the classification interval or classification gap between them. The purpose of maximizing the classification interval is to make the support vector machine have better generalization capabilities.

3.3 Main Unsupervised Learning Algorithms

3.3.1 The K-means Algorithm
MacQueen proposed the K-means algorithm in the 1960s, which is popular for cluster analysis. The idea of MacQueen’s K-Means algorithm is as follows: Given n data points \( \{x_1, x_2, ..., x_n\} \), K cluster centers \( \{a_1, a_2, ..., a_k\} \) is found so that the square sum of distance between each data point and its closest cluster center is the smallest, and the square sum of the distance is called the objective function \( W_n \) whose mathematical expression is:

\[ W_n = \sum_{i=1}^{n} \sum_{j<k} m_{i<k} |x_i - a_j|^2 \]

The processing of its algorithm is[3]:

a. K sample points are selected in the sample dataset D, and values of these sample points are assigned to the initial clustering center \( \bar{m}_i (i = 1, ... k) \);

b. The distance between each sample point from \( p_j (i = 1, ... n) \) and its clustering center \( \bar{m}_i \) is calculated:

\[ d(i, j) = \sqrt{|p_j - \bar{m}_i|^2} \]

c. The minimum distance \( \min (d(i, j)) \) between \( p_j \) and \( \bar{m}_i \) is found, and place \( p_j \) in the cluster that is closest to \( \bar{m}_i \).

d. The clustering center of each cluster again is calculated:

\[ \bar{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} p_{j} \quad (i = 1, ... k) \]

e. The squared difference \( E(t) \) of all points in dataset D is calculated according to step one and
comparing with the previous error $E(t-1)$.

If $E(t)-E(t-1)$ is observed if below zero, otherwise the algorithm ends.

Because the K-means algorithm is easy to describe, time-efficient and suitable for processing large-scale data, it has found wide application both at home and abroad in natural language processing, soil, archaeology, and many other fields since the 1970s.

3.4 Introduction to Main Semi-supervised Learning Algorithms

3.4.1 The Self-training Algorithm Based on the K-nearest Neighbor

Essentially, the principle of the so-called K-nearest neighbor algorithm is relatively simple, which uses the training set to divide the feature space into different regions and each sample occupies a certain region. When the test sample falls in the region of a training sample, it is considered to belong to the very training sample category. The above is all about the K-nearest neighbor algorithm for supervised learning. For the self-training K-nearest neighbor algorithm, there is no so-called training set. Instead, the feature space is used to divide different regions, and then gradually predict and classify the data category. Based on this, the prediction classification is gradually spread until all the samples are classified. By achieving self-training in the way mentioned above, we can obtain semi-supervised learning model of K-nearest neighbor.

3.4.2 The Semi-supervised Learning Algorithm Based on Divergence

The semi-supervised learning method based on divergence actually begins with a method. The main process of cooperative training is to first train a classifier by use of the training set, and then use the classifier to classify and label samples selected from unlabeled test samples. Then those classified test samples are added to the training set of the classifier, and the classification is continuously performed in this way until all the labeling is completed. This method has the advantage of receiving less interference from other factors, for instance, not being subject to model assumptions, non-convex loss function and data scale. Besides, this simple and effective method has a relatively rigorous theoretical basis, finding a wide range of applications. [4]

3.4.3 Semi-supervised Cluster

The clustering methods can be roughly classified into seven types, namely, partition clustering, hierarchical clustering, density-based clustering, network-based clustering, model-based clustering, clustering high-dimensional data, outlier analysis, and constrained clustering, which are a relatively important type of learning method in semi-supervised learning. The so-called clustering is the process of classifying the sample dataset into types featuring similarity. Semi-supervised clustering has achieved great progress in gene analysis, text mining, intrusion detection and other fields. Meanwhile, semi-supervised clustering has also accomplished much in image segmentation, and road detection and edge detection in GPS data. Therefore, semi-supervised clustering has become one of the machine learning directions worth researching and exploring in the future [5].

4. The Significance of Machine Learning Research

4.1 The Significance of Machine Learning

The research of machine learning system not only has important scientific significance, but also has a wide range of application value. The question of how biology and humans form skills in the environment and how people acquire concepts and knowledge in their learning has long been one of the key issues studied in the fields of physiology, psychology, cybernetics, education, and philosophy. Therefore, studying the process of machine learning is also the way for further exploring human learning process, which will in turn facilitate the development of those subjects mentioned above. In addition, there are some machine learning examples or tools that have already been commercialized, proving that people are increasingly aware of the potential and significance of machine learning in the
future. What’s more, as for the computer field, the demand from various industries for automation or intelligence is getting higher and higher, which means that the design of software programs becomes challenging and will face much higher requirement. However, the development of machine learning can be quickly applied to the design of programs and alleviate burdens on human beings through automation and intelligent programming, which is also one of the benefits that machine learning brings to programmers.

At present, base class systems such as the expert system and the semantic understanding system in artificial intelligence systems need to constantly learn from humans or existing training sets to master language skills and improve the function of judging semantics. The process of machine learning is similar to that of human learning, as if the “apprentice” gradually grows into an experienced “expert” through continuous absorption of new knowledge. Therefore, Samuel, known as the “pioneer of machine learning”, believed that machine learning was the theoretical basis and at the heart of artificial intelligence. Because in both speech and image recognition, a better artificial intelligence model could only be obtained through machine learning, and a better “artificial intelligence” machine could only be created through continuous training and knowledge accumulation. Meanwhile, Samuel also believed that a man-machine system with complete intelligent interface would not appear or become possible until significant progress was made in machine learning.

To sum up, the significance of machine learning lies in that it is the theoretical basis and core part of artificial intelligence. Therefore, there is still plenty of space in the field of machine learning for study and exploration.

4.2 Applications of Machine learning

4.2.1 Computer Vision
Traditional machine vision methods rely primarily on custom features which cannot capture high-level boundary information. In order to make up for the insufficiency of small-scale samples to effectively express complex features, computer vision was begun to turn to deep learning. For instance, in 2012, Krizhevsky adopted DNN to classify the dataset of Image Net LSVRC 2010. The error rates on top1 and top5 were 37.5% and 17.0%, far more than those of the traditional method. Besides, deep learning has also performed well in face recognition. For example, in 2014 Sun Yi used deep-hidden identity features to represent facial features, whose test accuracy on LFW reached 97.45%.

4.2.2 Speech Recognition
Speech recognition has developed for decades, of which statistical method is the traditional method, mainly based on the combination of hidden Markov model and Gauss mixture model (HMM-GMM). The characteristics of the traditional method cannot cover the original structural features of the speech data, so the tolerance for data correlation is low, which can be made up by replacing GMM with DNN. For example, in 2012, Microsoft’s speech video retrieval system reduced the word error rate from 27.4% to 18.5% through deep learning. DNN demonstrates improvement by about 10% compared with HMM-GMM, and CNN has stronger adaptability to data correlation than DNN.

4.2.3 Information Retrieval
Traditional information retrieval uses the TF-IDF system which is not only inefficient in dealing with large vocabulary problems, but fails to consider semantic similarity. Deep learning has been applied to information retrieval since 2009. DNN can well represent the word count feature of documents, and store semantically similar documents at similar addresses through deep self-encoders, thus improving retrieval efficiency. For example, in 2014, Shen Yelong proposed the convolution of deep semantic model which could project the same semantic words in the context through the convolution structure to the space vectors of the context features, thus improving the accuracy.
4.2.4 Natural Language Processing

Traditional natural language processing separately processes problems, such as language models and semantically related words, and there is no overall processing. Traditional systems have some shortcomings including shallow structure, linear and separable classification, and requiring preprocessing of artificial features. In 2008, Collobert began to apply DNN in the field of natural language processing, generating an error rate of 14.3%.

4.3 Future Trends

Over the past two decades featuring rapid Internet development, machine learning has gradually become a foundation and has been applied to various disciplines and fields. The development and application of machine learning not only promotes the improvement of its own form, but also facilitates the maturity of machine learning technology. From the current development and research trends, the future development trend of machine learning features the following aspects:

a. Unlabeled data learning;
b. Machine learning is applied to the mobile terminal or other mobile devices.
c. Study of machine learning is extended to the level of deep learning.
d. Deep learning is extended from static tasks to dynamic decision-making tasks.
e. Practical applications of natural language are further extended.

Natural language processing is mainly used to realize the technology of man-machine interaction and communication, whose main purpose is to make the machine intelligent enough to communicate with human beings, conforming to the idea and original intention of artificial intelligence. The examples that have been put into daily use include products with voice assistants or similar, because voice assistants can penetrate a wide range of fields and establish interactions between people and people, or people and information data. Many mobile devices and Internet applications involve voice products. For example, Apple’s Siri informs people about the weather condition, the temperature, and smog index of the day, and also gives suggestions for dressing and tripping. With the voice assistant, manual search for relevant information through the search engine is no longer necessary, and all it takes is one sentence requiring specific information.

f. Machine vision will keep penetrating into the production and living areas.

Human vision is limited and the range and precision that can be seen with the naked eye are very limited. Therefore, human beings are constantly creating various tools to break through such restrictions. For instance, telescopes and microscopes were invented to extend the visible range for human beings to better understand the world. Since the birth of computer technology, various tools have been created to assist humans in visual expansion. At present, high-precision instruments and impurity detection are typical machine learning tools mainly used in production and life, compared with which it is difficult for human eyes to reach such precision. This is exactly what computer vision needs to solve. Meanwhile, Baidu’s driverless cars, intelligent sensing systems, and medical imaging technologies are all realized through computer intelligent image analysis, pattern recognition or other technologies. These visual technologies, by utilizing intelligent sensors, video surveillance equipment, and other means, can sense changes in the environment so as to provide the most important and correct decision support by conducting analysis through a deep learning platform in the background. Computer vision is also a direction of artificial intelligence which will have great application in human production and life.

g. Blockchain accelerates platform establishment and integrated development.

The most important feature distinguishing a blockchain from a traditional database is its decentralization. The blockchain is based on distributed access when storing data and needs to reach consensus with other nodes. Therefore, if someone wants to illegally tamper with the data, he has to get confirmed by all the distributed nodes at the same time and change the previous records. One node does not confirm and the data cannot be altered, which is the biggest benefit of the blockchain technology. It guarantees data security, because all data is agreed upon by all nodes and each node is equally important, demonstrating high credibility. If the blockchain technology were to be used in the
banking sector, the clearing center would be no more necessary because data can be stored through the
nodes, which greatly saves the costs of time and storage. For example, the security of IoT (Internet of
Things) devices is confronted with great challenges, but the blockchain technology can ensure the
security of IoT data storage. Therefore, combining the two enables the IoT technology to better
develop. However, the blockchain technology has limited applications for now, which requires further
exploration.

5. Conclusions and Prospects
Machine learning, which is at the core of artificial intelligence, is an interdisciplinary subject itself
involving many fields such as the data structure of the computer itself, probability theory, statistics and
other related subjects. The ultimate goal of machine learning is to maximize the processing of the data
itself through algorithms or knowledge in different fields, so that computers can achieve the process of
self-learning through corresponding algorithms. Therefore, it can be seen that the research on and
selection of the algorithms are the most important part of machine learning on the whole.

The main advantage of machine learning lies in that the computer can achieve the purpose of
self-learning and even predict the trend through operating algorithms, which is of great value to
artificial intelligence as well as other fields. Because of this feature, the computer can be continuously
trained, the training dataset can be increased, and over time more accurate results can be obtained
through data accumulation. It thus can conduct analysis and prediction for many fields, greatly
enhancing the performance of artificial intelligence and making an unusual difference to the
development of human society.

References
[1] Liu Qin, Machine Learning [J], Journal of Wuhan Engineering Institute, 2001, 06:41-44.
[2] Li Zhiwei, The Development of Machine Learning and Several Learning Methods[J], Industrial &
Science Tribune, 2016, volume 15, issue 10.
[3] Han Xiaohong & Hu Yu, Research on K-means Clustering [J], Journal of Taiyuan University of
Technology, 2009, 05.
[4] Zhou Zhihua, Semi-supervised Learning Based on Divergence [J], Acta Automatica Sinica, 2013,
11.
[5] He Gaili, Research on Outlier Clustering Method Based on High-dimensional Sparse Data[D],
Journal of Chongqing University of Posts and Telecommunications, 2001.