A Review of Machine Learning

Ruqing Zhang[1, a], Wei Wei[2, b]

[1]School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450000, China;

[2]School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450000, China.

[a]zhangruq_haut@163.com, [b]nsyncw@163.com

Abstract

This paper briefly summarizes the development of machine learning in recent years, introduces the classification of machine learning, summarizes the research progress and related application fields of machine learning, and finally makes a summary of this paper.

Keywords

Machine learning, deep learning, decision tree.

1. Introduction

Machine learning is a multidisciplinary subject involving probability theory, statistics, algorithmic complexity theory and other fields. After preprocessing the data, a prediction model is established by the algorithm to predict the data. Since machine learning can predict after training data, it has attracted the attention of researchers and scholars.

2. Classification of Machine Learning

Machine learning is mainly divided into supervised learning and unsupervised learning based on learning methods. Supervised learning is to learn the data set with the features and labels, learn the classification rules from the training data set according to the features and labels, verify the prediction results on the test data set according to the division rules, and the accuracy of the prediction model is verified by comparing the prediction results with the label results. Supervised learning is mainly used for classification and prediction. Common supervised learning algorithms include regression analysis and statistical classification.

Unsupervised learning is to find good features in the sample data set without classification marks and does not need to be trained beforehand, it is mainly used for the data sets lacking of sufficient prior knowledge, the goal is to distinguish between the samples according to the principle of similarity in the process of learning, all the samples are automatically divided into different categories, and these categories are marked manually, so unsupervised learning efficiency is low. Unsupervised learning is mainly applied to k-means clustering and data dimension reduction. Common unsupervised learning algorithms include principal component analysis (PCA).

Whether it is supervised learning or unsupervised learning, by preprocessing the data set and using the appropriate algorithm to establish the prediction model to find the optimal solution of the problem. The higher the accuracy of the prediction on the test data set, the better the model.

3. Research Progress of Machine Learning

With the continuous research and development of machine learning, new learning algorithms have emerged to further promote the development of machine learning. The following mainly introduces the research progress of machine learning in recent years from the common algorithms of machine learning.

The decision tree selects a feature from the features of the training data set as the dividing criterion of the current node, and generates the child nodes recursively from top to bottom according to the evaluation criteria of the selected features, until the construction of the decision tree stops when the data set cannot be divided. [1]proposed a method for synthesizing decision trees which was ID3. [2]based on the improvement of ID3, the IDA algorithm was proposed. [3]presented a decision tree classification algorithm for classifying large data sets and stream data, which was executed in distributed environment and can process stream data on multiple processors. [4]introduced a hybrid algorithm of logical combination of machine learning technique decision tree along with big data analytical platform Hive, improved the efficiency of traditional decision tree classifier by using Map Reduce based framework.

The Naive Bayes classifier was based on probability to select the best category mark. It is widely used because it can obtain the correlation between input and output. One of the most successful cases was the spam filter [5] . [6]introduced the generalized Naive Bayes classifiers for binary classification problems, it can generate accurate predictions through a flexible, non-parametric fitting procedure, while being able to uncover hidden patterns in the data. [7] proposed a new set of models, termed Hierarchical Naïve Bayes models, which extend the model flexibility of Naïve Bayes by introducing latent variables to relax some of the independence statements in the models. [8]proposed a Correlation-based Feature Weighting (CFW) filter for Naive Bayes. In CFW, the weight for a feature was proportional to the difference between the feature-class correlation and the average feature-feature intercorrelation.

The support Vector Machine (SVM) was a learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space [9]. [10] a hierarchical recognition method based on an improved SVM decision tree and the layered feature selection method combining neural network with genetic algorithm was proposed. [11]introduced an application of SVM, to determine the beginning and end of recessions in real time.

The random forest was a classifier that contains multiple decision trees, and the category of its output was determined by the mode of the category of the individual tree output. [12] abstract random forests were one of the most successful ensemble methods which exhibits performance on the level of boosting and support vector machines. The method was fast, robust to noise, did not overfit and offers possibilities for explanation and visualization of its output. [13] combine ideas from on-line bagging, extremely random forest and proposed a novel on-line decision tree growing procedure, additionally, add a temporal weighting scheme for adaptively discarding some trees based on their out-of-bag-error in given time intervals and consequently growing of new trees. [14] proposed system was composed of a Restricted Boltzmann Machine for unsupervised feature learning, and random forest classifier, which were combined to jointly consider existing correlations between imaging data, features, and target variables.

Deep learning was one of the technical and research fields of machine learning, artificial intelligence was realized in the computing systems by establishing Artificial Neural Networks (ANNs) with hierarchical structure.[15] proposed deep learning based framework for age classification task, Deep Convolutional Neural Networks (Deep ConvNets) were used to extract high-level complex age related visual features and predict age range of input face image. [16] presented an exploratory machine learning attack based on deep learning to infer the

functionality of an arbitrary classifier by polling it as a black box, and using returned labels to build a functionally equivalent machine.

4. Application of Machine Learning

Machine learning already exists in every aspects of our lives, such as user behavior prediction, weather regulation, cancer prediction and so on. [17] develop customer churn prediction models by using machine learning technology on big data platforms, and using AUC standard measurements, it can reach 93.3%. [18] used electronic health record data to predicting cardiovascular risk. [19] the emotional state of different speakers can be distinguished from speech based on the decision tree of extreme learning machine (ELM), and achieved 89.6% recognition rate on average. [20] used machine learning to solve problems in the chemical sciences.

5. Summary

This paper gives a brief introduction to machine learning, including supervised learning and unsupervised learning of machine learning. It summarizes the research progress of classical machine learning algorithms, and finally discusses the application of machine learning in various aspects of life.

References

[1] J. R. Quinlan, “Induction of decision trees” Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[2] P.-L. Tu and J.-Y. Chung, “A new decision-tree classification algorithm for machine learning,” in Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI ’92, Arlington, VA, USA, 1992, pp. 370–377.

[3] Ben-Haim, Yael , and E. Tom-Tov . "A Streaming Parallel Decision Tree Algorithm." Journal of Machine Learning Research 11.11(2010):849-872.

[4] K. Ahlawat and A. P. Singh, “A Novel Hybrid Technique for Big Data Classification Using Decision Tree Learning,” in Computational Intelligence, Communications, and Business Analytics, vol. 775, J. K. Mandal, P. Dutta, and S. Mukhopadhyay, Eds. Singapore: Springer Singapore, 2017, pp. 118–128.

[5] D. Heckerman, “Bayesian Networks for Data Mining,” Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 79–119, 1997.

[6] K. Larsen, “Generalized Naive Bayes Classifiers,” SIGKDD Explor. Newsl., vol. 7, no. 1, pp. 76–81, Jun. 2005.

[7] H. Langseth and T. D. Nielsen, “Classification using Hierarchical Naïve Bayes models,” Mach Learn, vol. 63, no. 2, pp. 135–159, May 2006.

[8] L. Jiang, L. Zhang, C. Li, and J. Wu, “A Correlation-Based Feature Weighting Filter for Naive Bayes,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 2, pp. 201–213, Feb. 2019.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” Mach Learn, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[10] Q. Mao, X. Wang, and Y. Zhan, “SPEECH EMOTION RECOGNITION METHOD BASED ON IMPROVED DECISION TREE AND LAYERED FEATURE SELECTION,” Int. J. Human. Robot., vol. 07, no. 02, pp. 245–261, Jun. 2010.

[11] A. James, Y. Abu-Mostafa, and X. Qiao, “Nowcasting Recessions Using the SVM Machine Learning Algorithm,” SSRN Journal, 2018.

[12] M. Robnik-Šikonja, "Improving Random Forests," in Machine Learning: ECML 2004, vol. 3201, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 359–370.

[13] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line Random Forests," in 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 2009, pp. 1393–1400.

[14] S. Pereira et al., "Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation," Medical Image Analysis, vol. 44, pp. 228–244, Feb. 2018.

[15] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," Neurocomputing, vol. 187, pp. 4–10, Apr. 2016.

[16] Yi Shi, Y. Sagduyu, and A. Grushin, "How to steal a machine learning classifier with deep learning," in 2017 IEEE International Symposium on Technologies for Homeland Security (HST), Waltham, MA, USA, 2017, pp. 1–5.

[17] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," Journal of Business Research, vol. 94, pp. 290–301, Jan. 2019.

[18] J. Wolfson et al., "A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data," Statist. Med., vol. 34, no. 21, pp. 2941–2957, Sep. 2015.

[19] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," Neurocomputing, vol. 273, pp. 271–280, Jan. 2018.

[20] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature, vol. 559, no. 7715, pp. 547–555, Jul. 2018.