Computer Life
ISSN: 1819-4818 | Vol. 12, No. 1, 2024

The Application and Optimization of Machine Learning in Big Data Analysis

Zehan Wang

University of Maryland, College Park，USA

Abstract: As a branch of artificial intelligence, machine learning enables computers to learn rules and patterns automatically from massive data by training and optimizing algorithms. This paper discusses the important role of machine learning in big data analysis, paying special attention to the application comparison between neural network algorithm and support vector machine (SVM) algorithm. This paper first introduces the close relationship between machine learning and big data analysis, and then compares the differences between neural network and SVM in calculation time and prediction performance through experiments. The results show that when dealing with complex data, the neural network shows higher operating efficiency and prediction accuracy, and its accuracy is improved by more than 20% compared with SVM. This discovery provides a powerful basis for algorithm selection in the field of big data analysis, and indicates the great potential of neural network in processing large-scale data sets.

Keywords: Machine learning; Big data; Neural network; Support vector machine.

1. Introduction

Today, with the digital wave sweeping the world, big data has become an important force to promote social progress and technological innovation. From business decision-making, medical health to urban traffic and environmental monitoring, the application of big data is everywhere, which profoundly changes our lifestyle and working mode [1]. However, the explosive growth of big data has also brought many challenges. How to efficiently process, analyze and mine the value of these data has become a big problem for researchers [2]. In this context, machine learning technology came into being and gradually emerged. As a branch of artificial intelligence, machine learning enables computers to automatically learn rules and patterns from massive data by training and optimizing algorithms, and then make predictions and decisions [3]. Its appearance has greatly improved the intelligence level of data analysis and provided new ideas for solving complex big data problems.

The application of machine learning in big data analysis is extensive and deep. In the field of recommendation systems, machine learning algorithms can accurately push personalized content and services according to users' historical behaviors and preferences [4]; In the forecasting model, machine learning can use historical data to train the model, accurately predict the future trend, and provide strong support for enterprise decision-making [5]; In the field of anomaly detection, machine learning can help us find abnormal values or abnormal patterns in data in time, and provide important clues for risk control and troubleshooting [6]; In clustering analysis, machine learning can classify similar data points into one category and reveal the internal structure and correlation of data [7].

This paper aims to explore the application and optimization of machine learning in big data analysis. Firstly, the basis of machine learning is introduced, including definition, classification, algorithm and evaluation method. Then, it shows its examples and capabilities in big data processing, and then discusses how to optimize its performance, involving distributed framework, parallel computing and algorithm optimization. Finally, the optimization effect is verified by experiments, and the future research direction is prospected.

2. Fundamentals of machine learning

2.1. Basic concepts of machine learning

Machine learning utilizes algorithms to analyze and learn from data, enabling it to make informed decisions and predictions about real-world events [8]. Unlike conventional rule-based or static programs, machine learning systems have the capability to extract valuable insights from data and adapt their internal structures and parameters accordingly, enhancing their proficiency in handling future datasets.

In this paradigm, data is typically segmented into training and test sets. The former serves to train the model by optimizing its parameters to minimize prediction errors, while the latter assesses the model's performance on previously unseen data. The ultimate objective of machine learning is to identify a model that excels in both training and testing scenarios.

2.2. Main types of machine learning

Machine learning can be categorized into three primary learning methods: supervised, unsupervised, and reinforcement learning [9]. In supervised learning, each training sample comes with a pre-assigned label or target value, enabling the algorithm to learn the correlation between inputs and their corresponding labels. Conversely, unsupervised learning involves training samples without labels, requiring algorithms to independently uncover hidden structures, relationships, or patterns within the data. Reinforcement learning, on the other hand, focuses on teaching agents to make optimal decisions by interacting with their environment. This approach involves agents performing actions and learning from the subsequent reward signals provided by the environment, with the ultimate goal of maximizing cumulative rewards. Reinforcement learning algorithms have achieved significant breakthroughs in areas such as automatic control, game AI, and robot navigation.

2.3. Commonly used machine learning algorithms

Many excellent algorithms have emerged in the field of machine learning, and each algorithm has its unique advantages and applicable scenarios. Linear regression is a supervised learning algorithm for predicting numerical data. It describes the relationship between input features and target variables by fitting an optimal straight line or hyperplane. Linear regression has the advantages of simplicity and easy explanation, and has achieved good results in many practical problems. Decision trees are classification and regression techniques that utilize a tree-like structure. They recursively divide datasets into purer subsets to construct the tree. This algorithm is intuitive, easy to implement, capable of handling nonlinear relationships and discrete values, and exhibits robustness when confronted with missing or abnormal data. SVM is a popular supervised learning method for classification and regression tasks. It identifies an optimal hyperplane to separate samples of different classes, aiming to maximize the margin between them. Neural networks mimic the structure of human brain neurons, consisting of interconnected neurons. Each neuron processes input signals from others, generates output signals based on its internal weights and activation function. The network adjusts its weights using the backpropagation algorithm to minimize prediction errors. Neural networks excel in handling complex patterns and large datasets, achieving significant breakthroughs in areas like image and speech recognition.

3. Application of machine learning in big data analysis

3.1. Recommendation system

Recommendation system is a typical application of machine learning in big data analysis. On e-commerce, music, video and other platforms, the recommendation system can recommend personalized products, music or video content for users according to their historical behaviors, preferences and other users' behaviors. Through machine learning algorithm, the system can automatically learn users' interests and preferences, thus providing users with more accurate and personalized recommendations. This not only improves the satisfaction and stickiness of users, but also brings more business opportunities for enterprises.

3.2. Prediction model

Predictive model is another important application field of machine learning. In big data analysis, forecasting models can use historical data to predict future trends and results. For example, in the financial field, machine learning algorithms can predict the trend of stock prices according to historical stock data; In the medical field, machine learning can predict the occurrence probability of diseases and the rehabilitation of patients. These forecast results provide an important reference for enterprise decision-making, which is helpful to reduce risks and improve efficiency.

3.3. Anomaly detection

In large-scale data sets, the detection of outliers or abnormal patterns is an important task. Machine learning algorithm can automatically learn the normal pattern of data and identify abnormal values or behaviors that are inconsistent with the normal pattern. For example, in the field of network security, machine learning can help to detect abnormal behavior in network traffic, so as to find potential network attacks in time; In industrial production, machine learning can monitor the running state of equipment, find abnormal situations in time and give early warning to avoid production accidents.

3.4. Cluster analysis

Cluster analysis is an unsupervised learning method in machine learning, which can classify similar data points into one category. In big data analysis, cluster analysis can help us discover the internal structure and correlation in data. For example, in market segmentation, cluster analysis can classify customers with similar consumption behavior into one category, thus making more accurate marketing strategies for enterprises; In bioinformatics, cluster analysis can be used to analyze gene expression data and reveal the functional relationship between different genes.

4. Optimization strategy

In neural networks, the objective function serves as a metric for gauging the discrepancy between the predicted output and the actual sample label, thereby evaluating the effectiveness of network training. Depending on the specific task, distinct objective functions may be employed. Presently, square loss function and cross entropy loss function are widely utilized. Suppose that the genuine label linked to sample $x^{(i)}$ of type i is $y^{(i)}$, the network's regression prediction is $h\left(x^{(i)};\theta\right)$, and θ represents a network parameter; in this context, the square loss function can be formulated as follows:

$$J(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left\|y^{(i)}-h\left(x^{(i)};\theta\right)\right\|^2 \tag{1}$$

In the context of linear regression, where N denotes the sample count, $h(x)$ represents a linear function. Consequently, $J(\theta)$ becomes a convex function, allowing for the direct computation of an optimal solution using numerical methods to minimize $J(\theta)$. However, in the case of a neural network $h(\theta)$, the scenario is far more complex due to its nonlinear nature, making it impractical to obtain an optimal solution directly. Instead, the network's parameters θ must be iteratively updated using gradient descent techniques to converge on a locally optimal solution. For classification tasks, the cross entropy loss function serves as a widely adopted objective function. This metric effectively quantifies the discrepancy between the output distribution of the Softmax function and the empirical distribution.

The cross entropy loss function is the most frequently employed objective function for classification tasks. It quantifies the discrepancy between the output distribution of the Softmax function and the actual distribution. Suppose that the genuine label linked to sample $x^{(i)}$ of type i is $y^{(i)}$, and there are K distinct classes being outputted (denoted by $y^{(i)}\in\{1,\ldots,K\}$). In this scenario, the cross entropy loss function is expressed as follows:

$$J(\theta)=-\left[\sum_{i=1}^{N}\sum_{k=1}^{K}1\{y^{\{i\}}=k\}\log\frac{\exp(\theta_k x^{(i)})}{\sum_{j=1}^{K}\exp(\theta_j x^{(i)})}\right] \quad (2)$$

$\theta=\{\theta_1,\theta_2,\ldots,\theta_K\}$ represents the parameter generated by applying the Softmax function to the network's output, while N designates the total number of samples. $1\{\cdot\}$ serves as an indicator function, assigning a value of 1 when the enclosed expression holds true and 0 otherwise.

If a sample attribute contributes significantly to the classification system, its information gain will be notably higher. During feature selection, the initial step involves computing the information gain of each attribute. An attribute with the highest information gain value indicates the greatest discriminatory power within the given set. The formula for calculating the information gain of attribute A is as follows:

$$Gain(A)=I(s_1,s_2,\ldots,s_m)-E(A) \quad (3)$$

$I(s_1,s_2,\ldots,s_m)$ is determined by calculating the entropy of the sample and can be defined as follows:

$$I(s_1,s_2,\ldots,s_m)=-\sum_{i=1}^{m}P(C_i)\log_2 P(C_i) \quad (4)$$

In this context, $P(C_i)$ represents the likelihood that any given sample belongs to C_i. m denotes the total number of sample categories present, while s_i signifies the count of samples affiliated with class C_i. Lastly, s is the cumulative number of samples in the entire dataset.

5. Result analysis and discussion

The experiment aims to compare and analyze the performance of neural network and SVM in big data analysis. We use data sets from many fields and implement standard neural network and SVM algorithm. By evaluating the running time, recall and accuracy, the experiment comprehensively measures the calculation efficiency and prediction performance of the two algorithms. The experiment is carried out in a high-performance computing environment, which ensures the accuracy and reliability of the results and provides strong support for subsequent research and application.

As can be seen from Figure 1, in the initial stage, there is little difference in computing time between the neural network algorithm and the SVM algorithm, and even in some cases, the SVM may show a slightly faster computing speed. This may be because the computational complexity of the two algorithms is similar when the data volume is small and the model structure is relatively simple, while SVM may have more advantages in some simple tasks because of its solid mathematical foundation and optimized solution method.

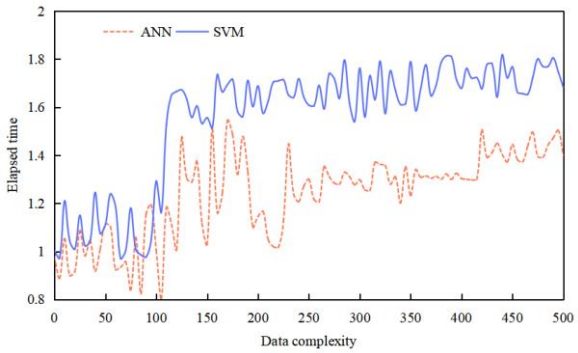

Figure 1 Algorithm calculation time comparison

With the increase of data complexity, neural network algorithm gradually shows its advantages in operating efficiency. This is mainly because neural network, especially deep learning network, has strong representation learning ability and can automatically extract useful features from complex data, thus reducing the demand and time cost of artificial feature engineering. In addition, neural network can further improve the computational efficiency through parallel computing and distributed training, making it more scalable when dealing with large-scale data sets.

Figure 2 and Figure 3 show the comparison of recall and accuracy of neural network algorithm and SVM algorithm in big data analysis and prediction respectively. It can be clearly seen from the figure that the neural network algorithm is significantly superior to the SVM algorithm in recall and accuracy, and the accuracy is improved by more than 20%.

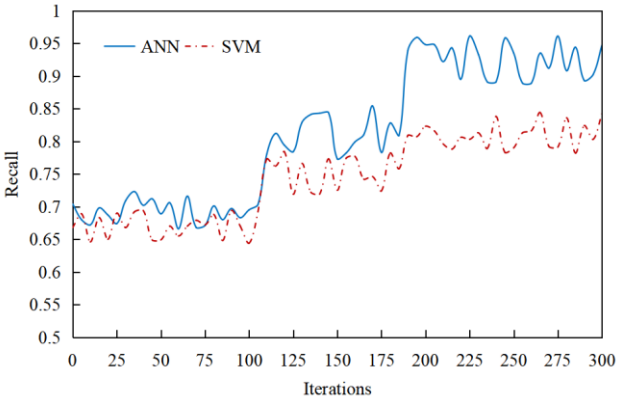

Figure 2 Comparison of recall rate of big data analysis

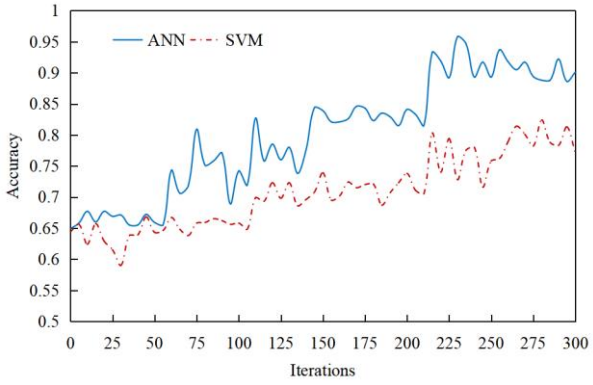

Figure 3 Comparison of accuracy of big data analysis

Neural network has strong nonlinear fitting ability and can capture complex patterns and associations in data. Moreover, neural network can realize more elaborate and complex feature transformation and abstract representation through the design of multi-layer network structure and activation function. In addition, the neural network can also learn and adjust its parameters through back propagation algorithm and

gradient descent optimization method, so that it can better adapt to the training data and improve the prediction performance. Although SVM algorithm is also a powerful machine learning algorithm, it may be limited by the nature of its linear classifier when dealing with complex data. Although nonlinear classification can be realized by mapping data to high-dimensional space through kernel technique, in some cases, this mapping may not fully capture the complex structure in data.

6. Conclusions

This paper delves into the utilization and refinement of machine learning in big data analysis. Through comparative experiments, it examines the efficacy of neural network and SVM algorithms in this context. The findings reveal that the neural network algorithm excels when confronted with intricate and voluminous datasets. As data complexity escalates, the neural network's computational prowess becomes increasingly evident, highlighting its remarkable processing capabilities and scalability for complex information. Additionally, in predictive performance, the neural network algorithm outperforms the SVM, achieving a substantial improvement in recall and accuracy by over 20%. This superior performance is credited to the neural network's robust nonlinear modeling and feature extraction abilities.

In conclusion, the neural network algorithm emerges as a highly efficient and accurate tool for big data analysis, particularly when tackling intricate and extensive datasets. Looking ahead, as computing power continues to advance and data volumes expand, the neural network's potential in big data analysis is poised for even greater realization.

References

[1] Cravero A, Samuel Sepúlveda. Use and Adaptations of Machine Learning in Big Data—Applications in Real Cases in Agriculture[J]. Electronics, 2021, 10(5):552.

[2] Gupta M, Gupta B. Survey of Breast Cancer Detection Using Machine Learning Techniques in Big Data[J]. Journal of Cases on Information Technology, 2019, 21(3):80-92.

[3] Tsaih R H, Kuo B S, Lin T H, et al. The use of big data analytics to predict the foreign exchange rate based on public media: A machine-learning experiment[J]. It Professional, 2018, 20(2):34-41.

[4] Stan L. Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data[J]. Technometrics, 2022, 2022(1):64.

[5] Gui G, Liu F, Sun J, et al. Flight Delay Prediction Based on Aviation Big Data and Machine Learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(1):140-150.

[6] Gui G, Zhou Z, Wang J, et al. Machine Learning Aided Air Traffic Flow Analysis Based on Aviation Big Data[J]. IEEE Transactions on Vehicular Technology, 2020, 69(5):4817-4826.

[7] Wei Y, Chen Y, Xiao M, et al. Protecting Machine Learning Integrity in Distributed Big Data Networking[J]. IEEE Network, 2020, 34(4):84-90.

[8] King P H. Signal Processing and Machine Learning for Biomedical Big Data[J]. IEEE Pulse, 2019, 10(3):34-35.

[9] Liu Y, Bi S, Shi Z, et al. When Machine Learning Meets Big Data: A Wireless Communication Perspective[J]. IEEE Vehicular Technology Magazine, 2020, 15(1):63-72.