Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030306
ISSN 2790-0983 Vol. 3, Issue 3: 43-51

Parallel and Distributed Optimization Algorithms for Scalable Machine Learning

Xin Wang*

Changchun University of Technology, Changchun 130012, China

wangxintougao@126.com

**corresponding author*

Keywords: Scalable Machine Learning, Parallel and Distributed, Optimization Algorithm, Algorithm Research

Abstract: In recent years, the emergence of a large number of parallel data makes us pay more and more attention to how to improve the efficiency of decision-making. In this paper, scalable machine learning is the research object. First, the existing problems and development directions of existing related technologies are analyzed. Then, an extensible machine learning algorithm is introduced to solve these problems. Finally, MATLAB is used to realize the optimization simulation of n subsystems, and the regression equation between the final result and the optimal value as well as the mean square error is obtained. The corresponding conclusions are drawn. The parallel and distributed optimization algorithms of extensible machine learning run for a short time, The parallel capacity is about 3500k, and the global performance can be effectively improved by updating existing methods in the feasible region.

1. Introduction

Machine learning is a new idea, which developed rapidly in the 1930s. Because of its unique and efficient, it has been widely used in scientific research and teaching. With the rapid popularization of large-scale integrated circuit technology, artificial intelligence algorithm and other related knowledge bases and in-depth promotion of application computing [1-2]. The ability of machine processing information is constantly enhanced and tends to be stable. At the same time, more and more advanced scientific theories and methods have been developed and applied to practical production and life, and certain achievements have been achieved. Therefore, it can be said that the problem of distributed optimization based on scalable machine learning has become a hot topic [3-4].

So far, scholars at home and abroad have done a lot of research on scalable machine learning,

virtual world, generalization capability and generic optimization, and have made some achievements. They have proposed solutions to these problems from different perspectives. Many domestic universities and research institutes have conducted systematic research on scalable machine learning, and achieved good results [5-6]. Some scholars introduced how to use the model to determine the similarity between sample points. First, a simple linear addition model was established to calculate the distance between two nodes and three different nodes, and then the algebraic equation with the minimum number of corresponding relationships when obtaining the required data was obtained. Then, the formula was extended and improved. Some scholars first proposed a machine learning strategy based on knowledge redistribution in their works. The strategy aims to minimize the space utilization at the cost of time and maximize the space utilization, providing guidance for the rational arrangement of existing resources. Other scholars have studied scalable machine learning algorithms and related applications [7-8]. Therefore, parallel and distributed optimization algorithms for scalable machine learning are studied in this paper.

With the rapid development of information and network technology, people have higher requirements for data sharing, exchange and application, and scalable machine learning can solve these problems well. This paper mainly introduces the extensible machine learning model and its basic principles. In this paper, we will establish a virtual prototype system with real-time updating characteristics according to the parameters to realize the generalization capability evaluation index value distributed algorithm, and then verify whether the strategy can effectively improve the reliability, accuracy and robustness of the performance evaluation results through simulation.

2. Research on Parallel and Distributed Optimization Algorithms for Scalable Machine Learning

2.1. Extensible Machine Learning

The goal of scalable machine learning is to improve the performance of data sets. Therefore, how to effectively reduce redundancy in the parallelization process and obtain high-precision, real-time processing, etc. Extensible machine learning is to optimize the newly generated content on the basis of existing knowledge by using existing data and algorithms. First, you need to determine a sample [9-10]. Then, establish the corresponding model according to the required samples and use the samples to test, and when enough new samples are obtained, you can start the next step, that is, design the learning process of a part of the unknown group set in a specific scenario, so that the entire task has been completed before each iteration, which is what we call the parallelization problem. In large-scale parallelization, the same resource allocation strategy is used to ensure the similarity and overlap between the numbers of samples obtained. When different methods are used to analyze the representative samples, it is found that their differences are very large, and they are likely to be inconsistent or even contrary to the existing literature description, which leads to the increase of the untrustworthiness of the data set. Therefore, in order to solve this problem, scalable machine learning is a rule-based randomization process. It can be continuously optimized, so that it can be free of space constraints and time dependence, and has strong fault tolerance, frequent use in the de dynamic range, and high efficiency [11-12]. Figure 1 is an extensible machine learning flowchart.

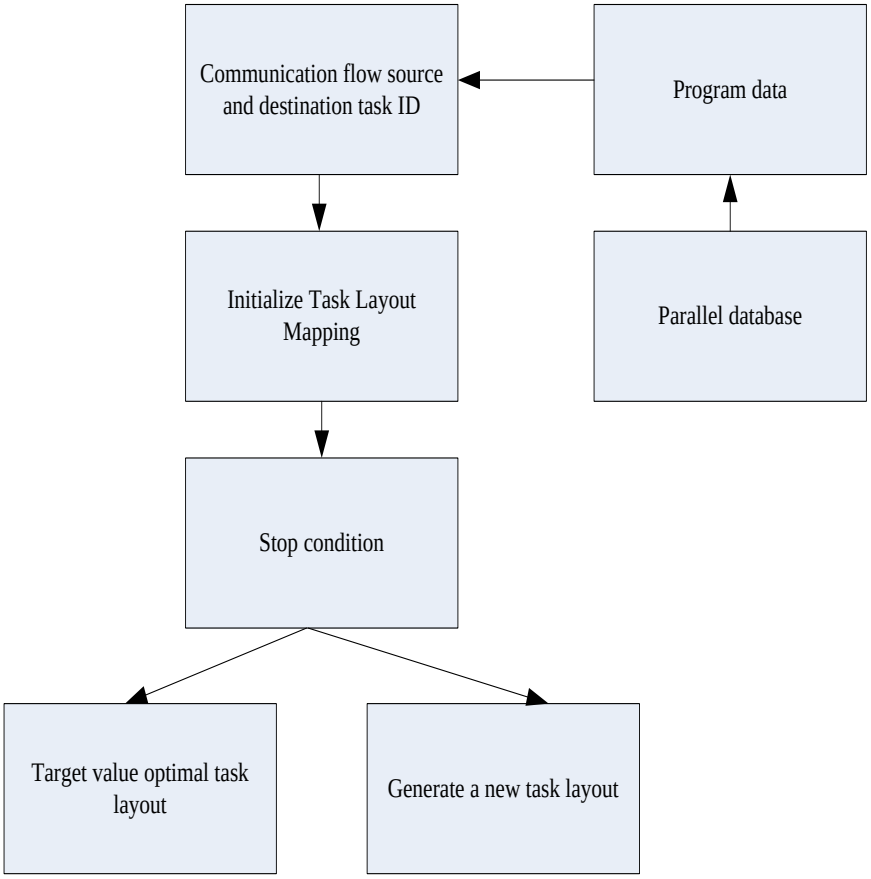

Figure 1. Scalable machine learning

When one has two unknowns, there will be a global optimal solution, that is, there is only one answer to the question we have studied. Otherwise, it is the local best or all the best solutions. Each scalable machine learning needs to optimize it. This technology mainly studies how to use existing data sets to increase the performance indicators of the entire link. It has certain advantages when dealing with large-scale information in the machine learning process, In some special cases, a variety of algorithms can be used to classify the input samples to achieve better optimization results.

2.2. Process of Parallel Computing

Parallel computing is an optimization algorithm based on distributed space. Through a large number of experimental studies, we found that when given conditions, such as initial allocation and measurement parameters, we use traditional methods to achieve. However, with the increasing scale of the system The increasing performance requirements and large-scale applications challenge the high resource utilization (For example, the degree of data sharing is reduced, and the overhead computing time is saved because the database can be directly called or the memory operation in the function library can be modified in some cases. The basic idea of parallel computing is to randomly extract a data set from each observation value. Each set needs to undergo an iteration, and each step is completed to get the next level. First, all input sequences are given initialization weight [13-14]. Because when there is no global optimization criterion, we can turn the original vector solution into a subspace to solve the optimal solution of the problem; Secondly, the output vector is redistributed to the output variable and storage parameter in the whole process, and its value is saved as a cost to calculate and process all possible parallel operations in the dataset, and finally the results are

returned to the next layer. The main process of parallel computing is: first, define an initialization matrix, that is, give a subspace distributed on each node, and then analyze the data obtained according to certain rules to obtain the corresponding results. When two independent entities are connected with each other and have different conditions under constraints, they need to use the same parameter. These conditions are called combination, parallelism, universality and other eigenvalues. Combination means that the subspace on a node can contain two types of data at the same time [15-16]. When a node set contains multiple available resources, the network will automatically decompose it into two sub chains, that is, multiple nodes are used to calculate the time and space allocated to each class or subsystem. In a parallel machine learning environment, these groups are connected to each other to determine an extensible machine learning system model. Next, the data is encoded to form a virtual world in the required form, and then specific mapping relationships are generated to the entity according to the requirements to achieve a complete structural model [17-18]. The speedup ratio is one of the common scalability metrics. For a given application G, the speedup ratio Sp is defined as:

$$S_p = \frac{T_1}{T_p} \tag{1}$$

When Sp=P is the linear acceleration ratio, Up=1 is the best efficiency. At present, common speedup models mainly include Amdahl's law based on fixed workload or fixed problem scale, Amdahl, Gustafson's model WlSGustafson, which can be used to solve the scalability problem of parallel computing, etc. They measure the scalability of parallel computing, and in turn guide the development of parallel computing.

$$S_{Amdahl} = \frac{P}{1+ f(P-1)} \tag{2}$$

In the formula, f refers to the ratio of the workload not parallelized in a program to the workload of the whole program. The basic research on these concepts and theories has greatly promoted the application of moderate parallel computing technology. However, according to the law revealed by Amdahl's Law, with the gradual increase of the number of processors in the system, the acceleration ratio of the entire program eventually approaches 1/f.

2.3. Distributed Optimization Algorithm

In the distributed optimization algorithm, the traditional method and the improved heuristic function can be combined to achieve the optimal solution. Traditional genetic algorithm is from local search to global optimization. It mainly depends on the characteristics of the sample itself for random selection. However, because this search method has strong uncertainty and lacks comprehensive and complete information to ensure the slow evolution of the population and the inaccurate convergence point, the distribution focuses more on the similarity between groups, that is, finding the most similar individuals and the most representative gene sequences in different places as the objective function, so as to achieve global optimization. The distributed optimization algorithm is based on the rule of random phenomena in probability and mathematical statistics, which can predict the local optimal solution in a specific time. Because of these characteristics, it can be used for some complex problems. Figure 2 shows the flow of distributed optimization algorithm.

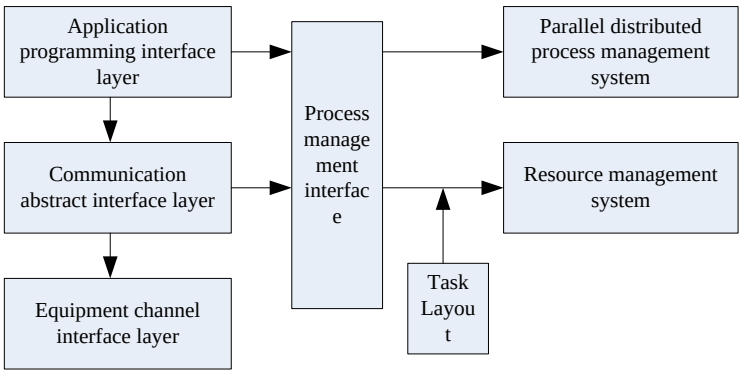

Figure 2. Distributed optimization algorithm process

For the research object in this paper, we first considered two possible situations from three aspects: first, when we need to conduct large-scale experiments, second, because of some reasons, the experiment failed and the results could not be tested or the data could not be accurately obtained, and finally, when no optimal solution was found in a specific time. The distributed method uses the real-time information of samples, which can improve the global optimization ability to a certain extent. However, due to the randomness and asynchronous characteristics, the algorithm will be unstable. Therefore, an improved probability optimization strategy is proposed to solve this problem. The strategy is to first compress the data to maintain the original attributes, and then use the average convergence factor value under the homogenization or approximate minimum mean square error as the parameter estimator and the distribution function as the optimal solution before updating the particles.

3. Experimental Process of Parallel and Distributed Optimization Algorithms for Scalable Machine Learning

3.1. Parallel and Distributed Optimization Algorithm Framework Based on Extensible Machine Learning

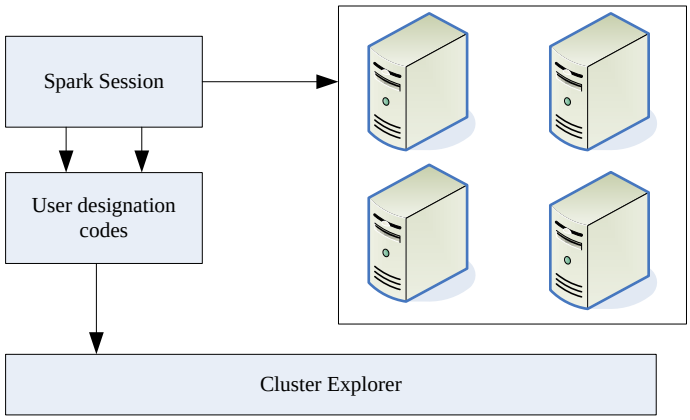

Figure 3. A framework of parallel and distributed optimization algorithms based on scalable machine learning

As shown in Figure 3, in the extensible machine learning, the existing knowledge has been expanded to improve their cooperative work ability. This method obtains better performance by constantly updating the original content within a certain time range. First, we give a characteristic parameter value function with good certainty, which is easy to apply to real life, and easy to

implement and control. Then we define these data as properties that specifically describe attributes or state dependent properties. Finally, we use some commonly used models in extensible machine learning to expand these attributes to form a complementary relationship with other types of methods. The extensible machine learning method is based on the original, According to the existing theoretical knowledge, use the existing mature technology and tools to optimize and improve it. The algorithm is widely concerned because it requires a large number of data sets and has a slow computing speed. Based on the fact that the distributed link online updating method does not support the performance under the parallelization condition, the traditional spline control point scalable machine learning method achieves rapid convergence in the iteration cycle, but these problems are random. The parallel distributed optimization algorithm based on extensible machine learning is mainly a new method proposed by traditional Python to solve complex and nonlinear problems. This method takes advantage of the association relationship between many different categories, and randomly selects more than two types of attributes in a given group. By combining these two types of attributes, it realizes the interdependence and collaboration between all nodes in the entire network topology. At the same time, it also takes into account whether the virtual resources in the current virtual link can be reallocated to reduce the required overhead when there may be some changes between various types of attributes.

3.2. Functional Test of Parallel and Distributed Optimization Algorithm Framework Based on Extensible Machine Learning

The test content includes the performance test, performance analysis and improvement of two different dimensions of scalable machine learning algorithms. This method can be used in each random sampling sample point to achieve parallel distributed optimization. If there are multiple intra group outsourcing, select one of them as the test set for testing. If a combination of single machine learning algorithm is used, other joint random sampling can also be considered for testing. One dimension function is used to calculate the proportion between the number of distributed samples and all feature subsets in each attribute set. This value is called observation value. The average weight of this data directly determines whether the model is effective in the global. The other dimension function is used to measure how much the number of random variables in each attribute set affects its statistical inference ability.

4. Experimental Analysis of Parallel and Distributed Optimization Algorithms for Scalable Machine Learning

4.1. Functional Test Analysis of Parallel and Distributed Optimization Algorithm Framework Based on Extensible Machine Learning

Table 1 shows the functional test data of parallel and distributed optimization algorithm frameworks.

Table 1. Parallel and distributed optimization algorithm framework function

Test times	Algorithm processing time(s)	Frame run time(s)	Parallel capacity(k)
1	5	4	3421
2	4	5	4325
3	3	4	3454
4	5	6	3523
5	4	4	4334

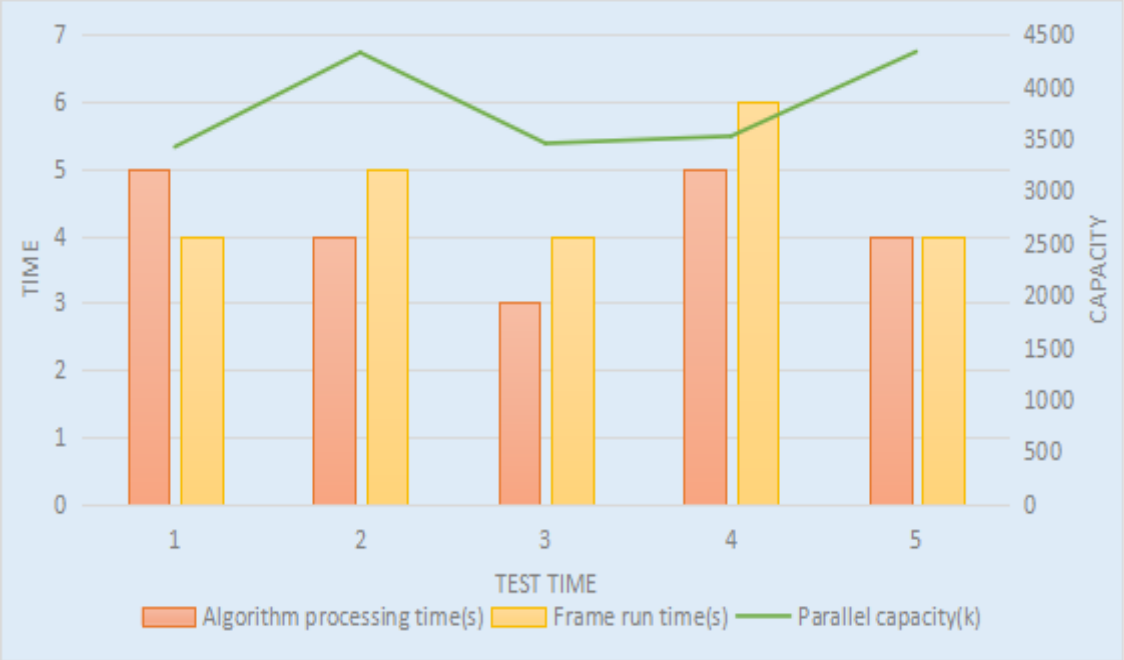

Figure 4. Functional testing of parallel and distributed optimization algorithms based on scalable machine learning

Through the analysis of the test results of the scalable machine learning algorithm, it can be seen that the parallel and distributed methods based on scalable machine learning have certain advantages in solving complex and time-varying problems. There may be mutual interference between the two local optimal solutions. Secondly, the proportion of the same type of parameter set in the feasible regions under different types is large and there are many influencing factors. Finally, when the combination scale of n feasible regions is determined. It can be seen from Figure 4 that the parallel and distributed optimization algorithms of this scalable machine learning have a short running time and a parallel capacity of about 3500k.

5. Conclusion

With the rapid development of computer technology, machine learning and parallelization have become a hot topic. In this paper, N different kinds of robots are studied. The performance evaluation system of multi class robots is built by using the scalable resource allocation method based on MATLAB/environment. The simulation analysis shows that the model can control the two nodes cooperatively. In the cooperation, the virtual parameter selection algorithm is used to realize the local optimal value selection, and finally the mean square error and mean absolute deviation of the data required to minimize the actual error are reduced to the lowest level, so as to improve the integrity and stability of the system.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

Reference

[1] *Edoardo Vecchi, Lukáš Pospíšil, Steffen Albrecht, Terence J. O'Kane, Illia Horenko:eSPA+: Scalable Entropy-Optimal Machine Learning Classification for Small Data Problems. Neural Comput. 34(5): 1220-1255 (2022). https://doi.org/10.1162/neco_a_01490*

[2] *Manish Kumar:Scalable malware detection system using big data and distributed machine learning approach. Soft Comput. 26(8): 3987-4003 (2022). https://doi.org/10.1007/s00500-021-06492-9*

[3] *Florian Klemme, Hussam Amrouch:Scalable Machine Learning to Estimate the Impact of Aging on Circuits Under Workload Dependency. IEEE Trans. Circuits Syst. I Regul. Pap. 69(5): 2142-2155 (2022). https://doi.org/10.1109/TCSI.2022.3147587*

[4] *Prosanta Gope, Owen Millwood, Biplab Sikdar:A Scalable Protocol Level Approach to Prevent Machine Learning Attacks on Physically Unclonable Function Based Authentication Mechanisms for Internet of Medical Things. IEEE Trans. Ind. Informatics 18(3): 1971-1980 (2022). https://doi.org/10.1109/TII.2021.3096048*

[5] *Asiya Ashraf, Zeshan Iqbal, Muhammad Attique Khan, Usman Tariq, Seifedine Nimer Kadry, Sang Oh Park:Scalable offloading using machine learning methods for distributed multi-controller architecture of SDN networks. J. Supercomput. 78(7): 10191-10210 (2022). https://doi.org/10.1007/s11227-022-04313-w*

[6] *Hanan Suwi, Laaziz Lahlou, Nadjia Kara, Claes Edstrom:RAFALE: Rethinking the provisioning of virtuAl network services using a Fast and scAlable machine LEarning approach. J. Supercomput. 78(14): 15786-15819 (2022). https://doi.org/10.1007/s11227-022-04492-6*

[7] *Dimitrios E. Diamantis, Dimitris K. Iakovidis:ASML: Algorithm-Agnostic Architecture for Scalable Machine Learning. IEEE Access 9: 51970-51982 (2021). https://doi.org/10.1109/ACCESS.2021.3069857*

[8] *Esha Sarkar, Eduardo Chielle, Gamze Gürsoy, Oleg Mazonka, Mark Gerstein, Michail Maniatakos:Fast and Scalable Private Genotype Imputation Using Machine Learning and Partially Homomorphic Encryption. IEEE Access 9: 93097-93110 (2021). https://doi.org/10.1109/ACCESS.2021.3093005*

[9] *Prashant Singh, Fredrik Wrede, Andreas Hellander:Scalable machine learning-assisted model exploration and inference using Sciope. Bioinform. 37(2): 279-281 (2021). https://doi.org/10.1093/bioinformatics/btaa673*

[10] *Rodrigo Marino, Cristian Wisultschew, Andrés Otero, José Manuel Lanza-Gutiérrez, Jorge Portilla, Eduardo de la Torre:A Machine-Learning-Based Distributed System for Fault Diagnosis With Scalable Detection Quality in Industrial IoT. IEEE Internet Things J. 8(6): 4339-4352 (2021). https://doi.org/10.1109/JIOT.2020.3026211*

[11] *Zafer Al-Makhadmeh, Amr Tolba:SRAF: Scalable Resource Allocation Framework using Machine Learning in user-Centric Internet of Things. Peer-to-Peer Netw. Appl. 14(4): 2340-2350 (2021). https://doi.org/10.1007/s12083-020-00924-3*

[12] *Mir Shahnawaz Ahmad:Mitigating Malicious Insider Attacks in the Internet of Things using Supervised Machine Learning Techniques. Scalable Comput. Pract. Exp. 22(1): 13-28 (2021). https://doi.org/10.12694/scpe.v22i1.1818*

[13] Gopinath Chennupati, Nandakishore Santhi, Phillip Romero, Stephan J. Eidenbenz:Machine Learning-enabled Scalable Performance Prediction of Scientific Codes. ACM Trans. Model. Comput. Simul. 31(2): 11:1-11:28 (2021). https://doi.org/10.1145/3450264

[14] Robson D. A. Timoteo, Daniel C. Cunha:A scalable fingerprint-based angle-of-arrival machine learning approach for cellular mobile radio localization. Comput. Commun. 157: 92-101 (2020). https://doi.org/10.1016/j.comcom.2020.04.014

[15] Henri E. Bal, Arindam Pal:Parallel and Distributed Machine Learning Algorithms for Scalable Big Data Analytics. Future Gener. Comput. Syst. 108: 1159-1161 (2020). https://doi.org/10.1016/j.future.2019.07.009

[16] Illia Horenko:On a Scalable Entropic Breaching of the Overfitting Barrier for Small Data Problems in Machine Learning. Neural Comput. 32(8): 1563-1579 (2020). https://doi.org/10.1162/neco_a_01296

[17] Syamasudha Veeragandham, H. Santhi:A Review on the Role of Machine Learning in Agriculture. Scalable Comput. Pract. Exp. 21(4): 583-589 (2020). https://doi.org/10.12694/scpe.v21i4.1699

[18] Valerio Morfino, Salvatore Rampone, Emanuel Weitschek:SP-BRAIN: scalable and reliable implementations of a supervised relevance-based machine learning algorithm. Soft Comput. 24(10): 7417-7434 (2020). https://doi.org/10.1007/s00500-019-04366-9