

REINFORCEMENT LEARNING ALGORITHM BASED INTELLIGENT SCHEDULING OPTIMIZATION METHOD AND SYSTEM FOR HETEROGENEOUS METADATA

5 TECHNICAL FIELD

[0001] The present invention relates to a heterogeneous data technology, and specifically relates to a reinforcement learning algorithm based intelligent scheduling optimization method and system for heterogeneous metadata.

10 BACKGROUND

[0002] In modern computing environments, scheduling optimization of heterogeneous metadata systems has become an important research direction for improving system performance and resource utilization. With the surge in data volume and the diversification of application scenarios, traditional scheduling methods face increasing challenges. Existing technologies typically rely on static rules or simple heuristic algorithms, making it difficult to adapt to dynamically changing loads and resource conditions.

[0003] Firstly, traditional scheduling methods lack the capability for dynamic analysis of real-time metadata and cannot respond promptly to changes in system status, resulting in lagging scheduling decisions. Secondly, existing technologies often rely on fixed feature sets for feature extraction, failing to fully exploit the potential information of multi-source heterogeneous metadata, which affects the accuracy of scheduling strategies. Finally, existing optimization mechanisms typically focus on a single target and fail to comprehensively consider multi-dimensional factors such as system load balance, access latency, and resource utilization rate, resulting in suboptimal scheduling performance.

25

SUMMARY OF THE INVENTION

[0004] The embodiments of the present invention provide a reinforcement learning algorithm based intelligent scheduling optimization method and system for heterogeneous metadata, which can solve the problems in related technology.

30 [0005] In a first aspect,

[0006] the reinforcement learning algorithm based intelligent scheduling optimization method for heterogeneous metadata provided by the embodiments of the present invention includes:

[0007] constructing a multi-level dynamic feature extraction mechanism, acquiring

real-time metadata information in a multi-source heterogeneous metadata system by a distributed data collection engine, and performing feature decomposition on the real-time metadata information by a self-attention mechanism to obtain a feature matrix set of metadata, the feature matrix set including a type feature matrix, a time-order feature matrix,
 5 a relationship feature matrix and a load feature matrix; inputting the feature matrix set into a transformer encoder for feature fusion to generate a fusion feature vector, and constructing a feature vector space through an adaptive dimension reduction algorithm; training a scheduling strategy model of a dual-channel neural network structure based on the feature vector space, a graph convolution neural network channel extracting spatial
 10 features based on the type feature matrix and the relationship feature matrix, and a gating circulation unit network channel extracting time-order features based on the time-order feature matrix and the load feature matrix and outputting initial scheduling strategy parameters;

[0008] establishing a hierarchical optimization mechanism based on deep
 15 reinforcement learning according to the initial scheduling strategy parameters, and generating a scheduling action candidate set by taking the fusion feature vector as state input; designing a multi-target assessment function to assess the candidate action, respectively computing a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system,
 20 performing weighted combination on the system load balance degree, the access delay score, the resource utilization rate score and the data locality score to obtain a multi-dimensional system performance comprehensive score, and constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score; performing strategy optimization by a dual-target Q network
 25 structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism, and outputting the optimized strategy parameters as a scheduling decision sequence; and

[0009] inputting the scheduling decision sequence into a distributed scheduling
 30 execution engine, and executing a metadata scheduling operation through a two-stage submission protocol; setting an adaptive monitoring threshold based on the fusion feature vector, and deploying a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, performing difference analysis on the monitoring data and the original fusion feature vector,

starting a hierarchical scheduling optimization strategy according to the difference degree, performing an incremental adjustment mode on local abnormality, and triggering strategy re-optimizing for global abnormality; and taking the optimized scheduling effect data as a new training sample, and performing distributed incremental training on the dual-channel
 5 neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model.

[0010] The establishing a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and generating a scheduling action candidate set by taking the fusion feature vector as state
 10 input; designing a multi-target assessment function to assess the candidate action, respectively computing a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system includes:

[0011] constructing a three-level cascade linkage optimization architecture based on
 15 the initial scheduling strategy parameters, the three-level cascade linkage optimization architecture including a strategy level and an execution level, a deep deterministic strategy gradient network of a dual exploration mechanism being built in the strategy level, and a self-correction four-dimensional performance assessment model being built in the execution level; the deep deterministic strategy gradient network being of a four-level
 20 fully-connected structure, the number of nodes at an input layer being the same as the dimension of a fusion feature vector, and the number of nodes at an output layer being the same as the dimension of an action space;

[0012] inputting the fusion feature vector into the deep deterministic strategy gradient network, performing hierarchical Bayesian modeling on the parameter distribution of the
 25 deep deterministic strategy gradient network to obtain hierarchical network weight distribution, acquiring a plurality of groups of network parameter samples from each level of the hierarchical network weight distribution by a Thompson sampling method to form a parameter sample set, and adding Gaussian noise disturbance into the parameter sample set; respectively loading the parameter sample set added with the Gaussian noise disturbance to
 30 the deep deterministic strategy gradient network to generate a plurality of strategy network instances, independently generating the scheduling action candidate set by the plurality of strategy network instances based on the initial scheduling strategy parameters; and

[0013] inputting the scheduling action candidate set into a four-dimensional performance assessment model at the execution level, setting a self-correction threshold for

an assessment result of each dimension by the four-dimensional performance assessment model, and triggering a re-assessment mechanism when the assessment result exceeds the self-correction threshold; and respectively computing the load balance degree score, the access delay score, the resource utilization rate score and the data locality score by the four-dimensional performance assessment model.

[0014] The constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score; performing strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism includes:

[0015] constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score, and computing a difference value between the multi-dimensional system performance comprehensive scores of the current moment and the previous moment to obtain a performance change value; setting the multi-dimensional system performance comprehensive score as a basic reward item, and setting the performance change value as a priority reward item; adaptively adjusting a combination weight of the basic reward item and the priority reward item based on the accumulated reward and the average reward in the training process so as to obtain a comprehensive reward function;

[0016] performing strategy optimization by the dual-target Q network structure, and constructing the online network and the target network which share the network structure; inputting the current system state into the online network to obtain a current action value, and inputting the system state of the next moment into the target network to obtain a target state value; substituting the current action value, the target state value and the comprehensive reward function into a Bellman equation to compute to obtain a time-order difference error;

[0017] computing a sample priority based on an absolute value of the time-order difference error; storing a historical mutual sample and the corresponding sample priority into an experience playback pool; performing importance sampling according to the sample priority to obtain a training sample batch; inputting the training sample batch into the online network, and updating strategy parameters of the online network in real time through a back propagation algorithm based on the time-order difference error; and

[0018] updating the strategy parameters of the target network by a soft update mechanism, and performing weighted average on the updated strategy parameters of the

online network and the updated strategy parameters of the target network to obtain updated parameters of the target network, in which, the weight of weighted average is determined by a preset soft update coefficient, and the stability of updating the target network parameters is kept through the soft update coefficient.

5 [0019] The computing a sample priority based on an absolute value of the time-order difference error; storing a historical mutual sample and the corresponding sample priority into an experience playback pool; performing importance sampling according to the sample priority to obtain a training sample batch; inputting the training sample batch into the online network, and updating strategy parameters of the online network in real time
10 through a back propagation algorithm based on the time-order difference error includes:

[0020] adding the absolute value of the time-order difference error with a priority reference value, and then performing exponential operation to obtain the sample priority, the priority reference value being used for ensuring that all samples have non-zero sampling probability;

15 [0021] forming a sample item by the historical mutual sample and the sample priority and storing the sample item into the experience playback pool that is of a priority queue structure, the priority queue structure including a leaf node and a non-leaf node, the leaf node storing the sample item, and the non-leaf node storing a priority sum of subtrees;

[0022] computing sampling probability based on the sample priority, the sampling
20 probability monotonically increasing along with the sample priority, and performing hierarchical sampling according to the sampling probability to obtain the training sample batch;

[0023] computing an importance weight of each sample in the training sample batch, the importance weight being used for compensating a sample distribution deviation caused
25 by priority sampling, and the importance weight being computed through an adjustable importance sampling index; and

[0024] inputting the training sample batch into the online network, computing a time-order difference loss function with a weight based on the importance weight, and computing a gradient of the loss function about the online network parameters through the
30 back propagation algorithm,

[0025] a gradient computation formula of the network parameters being as follows:

$$[0026] \quad \nabla_{\theta} L = \frac{1}{N} \sum_{i=1}^N w_i (y_i - Q(s_i, a_i; \theta)) \nabla_{\theta} Q(s_i, a_i; \theta);$$

[0027] in which, $\nabla_{\theta} L$ refers to the gradient of the loss function about a parameter θ , N

refers to the size of the experience playback pool, w_i refers to the importance weight of an i^{th} sample, y_i refers to a target Q value of the i^{th} sample, $Q(s_i, a_i; \theta)$ refers to a predicted Q value of the online network about a state-action pair (s_i, a_i) , and $\nabla_{\theta}Q$ refers to the gradient of the Q value about the parameter θ .

5 [0028] The setting an adaptive monitoring threshold based on the fusion feature vector, and deploying a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, performing difference analysis on the monitoring data and the original fusion feature vector, starting a hierarchical scheduling optimization strategy according to the difference degree,
10 performing an incremental adjustment mode on local abnormality, and triggering strategy re-optimizing for global abnormality includes:

 [0029] establishing a dynamic reference line based on the fusion feature vector, computing an adaptive adjustment factor in combination with system operation environment parameters, and multiplying the dynamic reference line by the adaptive
15 adjustment factor to obtain an adaptive monitoring threshold;

 [0030] deploying a three-level distributed monitoring network to collect the real-time system operation state data, the three-level distributed monitoring network being composed of a local monitoring agent, a regional monitoring node and a global monitoring center, and device-level state data collected by the local monitoring agent being gathered by the
20 regional monitoring node and then transmitted to the global monitoring center;

 [0031] performing feature extraction and fusion processing on the received real-time system operation state data by the global monitoring center so as to generate a real-time fusion feature vector, and comparing the real-time fusion feature vector with the adaptive monitoring threshold;

25 [0032] when the real-time fusion feature vector deviates from the adaptive monitoring threshold, computing a standardized difference value of the real-time fusion feature vector relative to the initial fusion feature vector, and computing a comprehensive difference degree based on the standardized difference value; and

 [0033] grading the comprehensive difference degree according to a preset difference
30 degree threshold, when the comprehensive difference degree is less than the difference degree threshold, determining to be the local abnormality and triggering the increment adjustment strategy, and when the comprehensive difference degree is greater than the difference degree threshold or equal to the difference degree threshold, determining to be the global abnormality and triggering the re-optimization strategy.

[0034] The taking the optimized scheduling effect data as a new training sample, and performing distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model includes:

5 [0035] collecting scheduling effect assessment data as a training sample, the scheduling effect assessment data containing system state vectors before and after scheduling execution, an actually executed scheduling action sequence, resource utilization rate data and energy consumption data, and performing normalization preprocessing on the scheduling effect assessment data to obtain a standardized training sample;

10 [0036] distributing the standardized training sample to the local training node by the federated learning framework, the local training node constructing a local model copy containing the dual-channel neural network and the dual-target Q network;

15 [0037] at the local training node, inputting the state feature vector in the standardized training sample into a state representation network of the dual-channel neural network so as to obtain a state code, and generating a new scheduling action by an action generation network of the dual-channel neural network based on the state code;

[0038] comparing the new scheduling action with an actual execution action in the standardized training sample, computing a strategy gradient, and updating parameters of the action generation network;

20 [0039] inputting the state code and the new scheduling action into the dual-target Q network, computing a resource utilization rate Q value and an energy efficiency Q value, weighting the resource utilization rate Q value and the energy efficiency Q value according to performance index values in the standardized training sample to obtain a comprehensive Q value, and updating the parameters of the dual-target Q network based on the
25 comprehensive Q value;

[0040] periodically triggering a federated average algorithm, collecting model parameters of the local training node for global aggregation to obtain global model parameters, and distributing the global model parameters to the local training node to update the local model copy; and

30 [0041] collecting new scheduling effect assessment data as incremental training samples, and continuously optimizing the dual-channel neural network and the dual-target Q network through the reinforcement learning framework.

[0042] The inputting the state code and the new scheduling action into the dual-target Q network, computing a resource utilization rate Q value and an energy efficiency Q value,

weighting the resource utilization rate Q value and the energy efficiency Q value according to performance index values in the standardized training sample to obtain a comprehensive Q value, and updating the parameters of the dual-target Q network based on the comprehensive Q value includes:

5 [0043] splicing the state code and the new scheduling action to obtain a state action vector, the state action vector being used as a common input of a resource utilization rate assessment network and an energy efficiency assessment network;

[0044] inputting the state action vector into the resource utilization rate assessment network, performing nonlinear transformation through a first hidden layer and a second
10 hidden layer of the resource utilization rate assessment network to obtain a resource utilization rate feature, and computing the resource utilization rate Q value based on the resource utilization rate feature;

[0045] a computation formula of the resource utilization rate Q value being as follows:

$$[0046] \quad Q_r(s_t, a_t) = W_3 h_2 + b_3,$$

15 [0047] in which, $Q_r(s_t, a_t)$ refers to the resource utilization rate Q value, s_t refers to the state code at the moment t , a_t refers to the action at the moment t , W_3 refers to a weight matrix at the output layer, h_2 refers to the resource utilization rate feature, and b_3 refers to a bias scalar at the output layer;

[0048] inputting the state action vector into the energy efficiency assessment network,
20 performing nonlinear transformation through the first hidden layer and the second hidden layer of the energy efficiency assessment network to obtain an energy efficiency feature, and computing the energy efficiency Q value based on the energy efficiency feature,

[0049] a computation formula of the energy efficiency Q value being as follows:

$$[0050] \quad Q_e(s_t, a_t) = W_3^e h_2^e + b_3^e;$$

25 [0051] in which, $Q_e(s_t, a_t)$ refers to the energy efficiency Q value, W_3^e refers to the weight matrix at the output layer of the energy efficiency assessment network, h_2^e refers to the energy efficiency feature, and b_3^e refers to the bias scalar at the output layer of the energy efficiency assessment network;

[0052] computing a relative change amount of the resource utilization rate based on
30 performance indexes in a current batch of training samples, inputting the relative change amount of the resource utilization rate into an adaptive weight computation unit, and computing the resource utilization rate and the energy efficiency weight by the adaptive weight computation unit;

[0053] multiplying the resource utilization rate Q value by the resource utilization rate weight to obtain a weighted resource utilization rate Q value, multiplying the energy efficiency Q value by the energy efficiency weight to obtain a weighted energy efficiency Q value, and adding the weighted resource utilization rate Q value to the weighted energy efficiency Q value to obtain the comprehensive Q value; and

[0054] constructing the time-order difference loss function based on the comprehensive Q value, and updating network parameters of the resource utilization rate assessment network and the energy efficiency assessment network at the same time through the time-order difference loss function.

10 [0055] In a second aspect,

[0056] the reinforcement learning algorithm based intelligent scheduling optimization system for heterogeneous metadata provided by the embodiment of the present invention includes:

[0057] a first unit, configured to construct a multi-level dynamic feature extraction mechanism, acquire real-time metadata information in a multi-source heterogeneous metadata system by a distributed data collection engine, and perform feature decomposition on the real-time metadata information by a self-attention mechanism to obtain a feature matrix set of metadata, the feature matrix set including a type feature matrix, a time-order feature matrix, a relationship feature matrix and a load feature matrix; input the feature matrix set into a transformer encoder for feature fusion to generate a fusion feature vector, and construct a feature vector space through an adaptive dimension reduction algorithm; and train a scheduling strategy model of a dual-channel neural network structure based on the feature vector space, a graph convolution neural network channel extracting spatial features based on the type feature matrix and the relationship feature matrix, and a gating circulation unit network channel extracting time-order features based on the time-order feature matrix and the load feature matrix and outputting initial scheduling strategy parameters;

[0058] a second unit, configured to establish a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and generate a scheduling action candidate set by taking the fusion feature vector as state input; design a multi-target assessment function to assess the candidate action, respectively compute a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system, perform weighted combination on the system load balance degree, the access delay score,

the resource utilization rate score and the data locality score to obtain a multi-dimensional system performance comprehensive score, and construct a comprehensive reward function based on the multi-dimensional system performance comprehensive score; perform strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism, and output the optimized strategy parameters as a scheduling decision sequence; and

[0059] a third unit, configured to input the scheduling decision sequence into a distributed scheduling execution engine, and execute a metadata scheduling operation through a two-stage submission protocol; set an adaptive monitoring threshold based on the fusion feature vector, and deploy a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, perform difference analysis on the monitoring data and the original fusion feature vector, start a hierarchical scheduling optimization strategy according to the difference degree, perform an incremental adjustment mode on local abnormality, and trigger strategy re-optimizing for global abnormality; and take the optimized scheduling effect data as a new training sample, and perform distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model.

[0060] In a third aspect,

[0061] an electronic device provided by the embodiment of the present invention includes:

[0062] a processor; and

[0063] a memory configured to store processor executable instructions; and

[0064] the processor is configured to call instructions stored in the memory to execute the abovementioned method.

[0065] In a fourth aspect,

[0066] the embodiment of the present invention provides a computer-readable storage medium having computer program instructions stored therein, when executed by a processor, the computer program instructions execute the abovementioned method.

[0067] The present invention has the following beneficial effects:

[0068] 1. the scheduling efficiency is improved: by constructing the multi-level dynamic feature extraction mechanism and through the adaptive dimension reduction algorithm, the multi-source heterogeneous metadata features can be obtained and fused in

real time, so that the scheduling strategy is optimized, and the scheduling efficiency of the system is remarkably improved.

[0069] 2. The system performance is enhanced: the scheduling action is assessed by the multi-target assessment function, and a plurality of dimensions such as the system load
5 balance, access delay, resource utilization rate and data locality are taken into account, so that the overall performance of the system can be effectively improved, and reasonable utilization of resources is ensured.

[0070] 3. Continuous optimization is realized: distributed incremental training is performed on the scheduling strategy model by the federated learning framework, thus the
10 scheduling strategy can be continuously optimized according to the real-time monitoring data and the scheduling effect data, and adaptive adjustment and continuous improvement of the system can be realized.

BRIEF DESCRIPTION OF THE DRAWINGS

[0071] FIG. 1 is a flowchart of a reinforcement learning algorithm based intelligent
15 scheduling optimization method for heterogeneous metadata according to an embodiment of the present invention; and

[0072] FIG. 2 is a schematic structural diagram of a reinforcement learning algorithm
based intelligent scheduling optimization system for heterogeneous metadata according to
20 an embodiment of the present invention.

DETAILED DESCRIPTION

[0073] In order to make the purpose, technical solution and advantages of the
embodiment of the present invention clearer, the technical solutions in the embodiments of
25 the present invention will be clearly and completely described in combination with the accompanying drawings in the embodiments of the present invention. Definitely, the described embodiments are only a part of the embodiments of the present invention, not all of them. Based on the embodiments in the present invention, all other embodiments obtained by those of ordinary skill in the art without carrying out creative labor fall within
30 the scope of protection of the present invention.

[0074] The technical solutions of the present invention are described in detail by specific embodiments below. The following specific embodiments can be combined with each other, and the same or similar concepts or processes may not be repeated in some embodiments.

[0075] FIG. 1 is a flowchart of a reinforcement learning algorithm based intelligent scheduling optimization method for heterogeneous metadata according to an embodiment of the present invention, as shown in FIG. 1, the method includes:

[0076] S11: constructing a multi-level dynamic feature extraction mechanism,
 5 acquiring real-time metadata information in a multi-source heterogeneous metadata system by a distributed data collection engine, and performing feature decomposition on the real-time metadata information by a self-attention mechanism to obtain a feature matrix set of metadata, the feature matrix set including a type feature matrix, a time-order feature matrix, a relationship feature matrix and a load feature matrix; inputting the feature matrix
 10 set into a transformer encoder for feature fusion to generate a fusion feature vector, and constructing a feature vector space through an adaptive dimension reduction algorithm; training a scheduling strategy model of a dual-channel neural network structure based on the feature vector space, a graph convolution neural network channel extracting spatial features based on the type feature matrix and the relationship feature matrix, and a gating
 15 circulation unit network channel extracting time-order features based on the time-order feature matrix and the load feature matrix and outputting initial scheduling strategy parameters;

[0077] S12: establishing a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and
 20 generating a scheduling action candidate set by taking the fusion feature vector as state input; designing a multi-target assessment function to assess the candidate action, respectively computing a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system, performing weighted combination on the system load balance degree, the access delay
 25 score, the resource utilization rate score and the data locality score to obtain a multi-dimensional system performance comprehensive score, and constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score; performing strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the
 30 comprehensive reward function, a target network keeping training stability through a soft update mechanism, and outputting the optimized strategy parameters as a scheduling decision sequence; and

[0078] S13: inputting the scheduling decision sequence into a distributed scheduling execution engine, and executing a metadata scheduling operation through a two-stage

submission protocol; setting an adaptive monitoring threshold based on the fusion feature vector, and deploying a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, performing difference analysis on the monitoring data and the original fusion feature vector, starting a hierarchical scheduling optimization strategy according to the difference degree, performing an incremental adjustment mode on local abnormality, and triggering strategy re-optimizing for global abnormality; and taking the optimized scheduling effect data as a new training sample, and performing distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model.

[0079] In an optional embodiment, the establishing a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and generating a scheduling action candidate set by taking the fusion feature vector as state input; designing a multi-target assessment function to assess the candidate action, respectively computing a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system includes:

[0080] constructing a three-level cascade linkage optimization architecture based on the initial scheduling strategy parameters, the three-level cascade linkage optimization architecture including a strategy level and an execution level, a deep deterministic strategy gradient network of a dual exploration mechanism being built in the strategy level, and a self-correction four-dimensional performance assessment model being built in the execution level; the deep deterministic strategy gradient network being of a four-level fully-connected structure, the number of nodes at an input layer being the same as the dimension of a fusion feature vector, and the number of nodes at an output layer being the same as the dimension of an action space;

[0081] inputting the fusion feature vector into the deep deterministic strategy gradient network, performing hierarchical Bayesian modeling on the parameter distribution of the deep deterministic strategy gradient network to obtain hierarchical network weight distribution, acquiring a plurality of groups of network parameter samples from each level of the hierarchical network weight distribution by a Thompson sampling method to form a parameter sample set, and adding Gaussian noise disturbance into the parameter sample set; respectively loading the parameter sample set added with the Gaussian noise disturbance to the deep deterministic strategy gradient network to generate a plurality of strategy network

instances, independently generating the scheduling action candidate set by the plurality of strategy network instances based on the initial scheduling strategy parameters; and

[0082] inputting the scheduling action candidate set into a four-dimensional performance assessment model at the execution level, setting a self-correction threshold for an assessment result of each dimension by the four-dimensional performance assessment model, and triggering a re-assessment mechanism when the assessment result exceeds the self-correction threshold; and respectively computing the load balance degree score, the access delay score, the resource utilization rate score and the data locality score by the four-dimensional performance assessment model.

[0083] Firstly, the comprehensive reward function is constructed based on the multi-dimensional system performance comprehensive score. The difference value between the system performance comprehensive scores of the current moment and the previous moment is computed to obtain the performance change value. For example, if the performance comprehensive score of the current moment is 0.85, and the performance comprehensive score of the previous moment is 0.78, the performance change value will be 0.07. The performance comprehensive score is treated as the basic reward item, and the performance change value is treated as the priority reward item. In practical application, the initial basic reward item weight is set to be 0.6, and the priority reward item weight is 0.4. In the training process, when the accumulated reward exceeds 1,000 and the average reward is greater than 0.8, the weight ratio will be dynamically adjusted, the basic reward item weight is adjusted to be 0.5, the priority reward item weight is adjusted to be 0.5, and then the comprehensive reward function is obtained.

[0084] Secondly, the dual-target Q network structure is utilized to perform strategy optimization. The online network and the target network which have the same network structure are constructed, containing the input layer, three hidden layers and the output layer. In the input layer, the current system state information is received, including the features of resource utilization rate, the load distribution and the like. The current state is inputted into the online network to obtain the action value assessment result, for example, the assessment value of a certain scheduling action is 0.92. Moreover, the state of the next moment is inputted into the target network to obtain a target state value of 0.88. The action value, the state value and comprehensive reward are combined to form the time-order difference error.

[0085] Thirdly, the sample priority is computed based on the absolute value of the time-order difference error. For example, if the absolute value of the time-order difference

error is 0.15, the corresponding sample priority will be 0.85. The historical mutual samples and the priority information are stored in the experience playback pool, and the capacity of the playback pool is set to be 10,000. Importance sampling is performed according to the sample priority, and the higher the priority of the sample is, the higher the sampling probability is. A training sample batch with the size of 256 is obtained by sampling and inputted into the online network for training. The online network parameters are updated by the back propagation algorithm, and thus the predicted value gradually approaches a target value.

[0086] Fourthly, the target network parameters are updated by the soft update mechanism. The soft update coefficient is set to be 0.01, and weighted average is performed on the updated online network parameters and the target network parameters. In specific implementation, the new target network parameter value is equal to 0.99 times of the original target network parameter value plus 0.01 times of the online network parameter value. Stable updating of the target network parameters is ensured by a small soft update coefficient, and the training stability is improved.

[0087] Through the solution of the present invention:

[0088] in algorithm performance, the comprehensive reward function is constructed and the weight is dynamically adjusted, thereby improving the adaptability of the model to environmental changes; and the dual-target Q network structure and the soft update mechanism are adopted, which can obviously improve the training stability, and avoid strategy oscillation. In system efficiency, the sample utilization efficiency is improved through the priority-based experience playback mechanism; it is ensured that high-value samples can be fully trained by importance sampling; and the soft update mechanism reduces the fluctuation of parameter update. In practicability, the algorithm design takes into full account the actual application requirements and has relatively high expandability; the parameter configuration is flexible and adjustable, thus facilitating optimization in different scenes; and the training process is stable and controllable and is suitable for engineering practice deployment.

[0089] In an optional embodiment, the constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score; performing strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism includes:

[0090] constructing a comprehensive reward function based on the multi-dimensional

system performance comprehensive score, and computing a difference value between the multi-dimensional system performance comprehensive scores of the current moment and the previous moment to obtain a performance change value; setting the multi-dimensional system performance comprehensive score as a basic reward item, and setting the performance change value as a priority reward item; adaptively adjusting a combination weight of the basic reward item and the priority reward item based on the accumulated reward and the average reward in the training process so as to obtain a comprehensive reward function;

[0091] performing strategy optimization by the dual-target Q network structure, and constructing the online network and the target network which share the network structure; inputting the current system state into the online network to obtain a current action value, and inputting the system state of the next moment into the target network to obtain a target state value; substituting the current action value, the target state value and the comprehensive reward function into a Bellman equation to compute to obtain a time-order difference error;

[0092] computing a sample priority based on an absolute value of the time-order difference error; storing a historical mutual sample and the corresponding sample priority into an experience playback pool; performing importance sampling according to the sample priority to obtain a training sample batch; inputting the training sample batch into the online network, and updating strategy parameters of the online network in real time through a back propagation algorithm based on the time-order difference error; and

[0093] updating the strategy parameters of the target network by a soft update mechanism, and performing weighted average on the updated strategy parameters of the online network and the updated strategy parameters of the target network to obtain updated parameters of the target network, in which, the weight of weighted average is determined by a preset soft update coefficient, and the stability of updating the target network parameters is kept through the soft update coefficient.

[0094] Firstly, the multi-dimensional system performance comprehensive score is the basis of constructing the comprehensive reward function. The score is obtained by assessing the performance of the system in different dimensions. The score of each dimension can be quantified through set standards, such as response time, resource utilization rate and system stability. The comprehensive score of the current moment is compared with the comprehensive score of the previous moment, thus the performance change value can be computed. This change value reflects the improvement or reduction of

the system performance.

[0095] Secondly, the multi-dimensional system performance comprehensive score is treated as the basic reward item, and the performance change value is treated as the priority reward item. The basic reward item provides basic feedback of the system, and the priority
 5 reward item emphasizes dynamic change of the performance. The combined weight of the two reward items is dynamically adjusted by analyzing the accumulated reward and the average reward in the training process, thus making sure that the comprehensive reward function can adapt to different training stages and environmental changes.

[0096] In the aspect of strategy optimization, the dual-target Q network structure is
 10 adopted. The online network and the target network share the same network structure. The online network is used for processing the current system state in real time and outputting the value of the current action, and the target network is used for assessing the system state value of the next moment. The current action value, the target state value and the comprehensive reward function are combined to compute the time-order difference error
 15 through the Bellman equation. The error provides an important basis for subsequent strategy updating.

[0097] In the computation of the sample priority, the importance of the sample is determined based on the absolute value of the time-order difference error. The historical mutual samples and the corresponding priority are stored in the experience playback pool
 20 for subsequent training. Importance sampling is performed according to the sample priority to obtain training sample batches. The samples are inputted into the online network, and strategy parameters are updated in real time by the back propagation algorithm so as to improve the learning efficiency of the network.

[0098] Thirdly, the strategy parameters of the target network are updated by the soft
 25 update mechanism. The weighted average is performed on the updated parameters of the online network and the parameters of the target network, so as to make sure that the parameter updating process of the target network is stable. The weight of weighted average is determined by the preset soft update coefficient, and such mechanism effectively avoids violent fluctuation in the parameter updating process, thereby improving the training
 30 stability.

[0099] Through the solution of the present invention:

[0100] by dynamically adjusting the comprehensive reward function, the system can better adapt to different environment changes, and the learning efficiency and the strategy optimization effect are improved. Due to the application of the dual-target Q network

structure, the stability of strategy optimization is enhanced, and fluctuation in the training process is reduced, so that the system is more reliable in performance in a complex environment. With adoption of the soft update mechanism, stable updating of the target network parameters is ensured, and therefore, the overall performance and the response capability of the system are further improved.

[0101] In an optional embodiment, the computing a sample priority based on an absolute value of the time-order difference error; storing a historical mutual sample and the corresponding sample priority into an experience playback pool; performing importance sampling according to the sample priority to obtain a training sample batch; inputting the training sample batch into the online network, and updating strategy parameters of the online network in real time through a back propagation algorithm based on the time-order difference error includes:

[0102] adding the absolute value of the time-order difference error with a priority reference value, and then performing exponential operation to obtain the sample priority, the priority reference value being used for ensuring that all samples have non-zero sampling probability;

[0103] forming a sample item by the historical mutual sample and the sample priority and storing the sample item into the experience playback pool that is of a priority queue structure, the priority queue structure including a leaf node and a non-leaf node, the leaf node storing the sample item, and the non-leaf node storing a priority sum of subtrees;

[0104] computing sampling probability based on the sample priority, the sampling probability monotonically increasing along with the sample priority, and performing hierarchical sampling according to the sampling probability to obtain the training sample batch;

[0105] computing an importance weight of each sample in the training sample batch, the importance weight being used for compensating a sample distribution deviation caused by priority sampling, and the importance weight being computed through an adjustable importance sampling index; and

[0106] inputting the training sample batch into the online network, computing a time-order difference loss function with a weight based on the importance weight, and computing a gradient of the loss function about the online network parameters through the back propagation algorithm,

[0107] a gradient computation formula of the network parameters being as follows:

$$\nabla_{\theta} L = \frac{1}{N} \sum_{i=1}^N w_i (y_i - Q(s_i, a_i; \theta)) \nabla_{\theta} Q(s_i, a_i; \theta);$$

[0109] in which, $\nabla_{\theta} L$ refers to the gradient of the loss function about a parameter θ , N refers to the size of the experience playback pool, w_i refers to the importance weight of an i^{th} sample, y_i refers to a target Q value of the i^{th} sample, $Q(s_i, a_i; \theta)$ refers to a predicted Q value of the online network about a state-action pair (s_i, a_i) , and $\nabla_{\theta} Q$ refers to the gradient of the Q value about the parameter θ .

[0110] Firstly, the historical mutual samples are collected, and the time-order difference error of each sample is computed. The time-order difference error refers to the difference between the current state and the next state and reflects the performance of the model under the current decision. The absolute values of the errors are processed to obtain the priority of each sample. The computation mode of the priority is to add the absolute values to a preset priority reference value and then perform exponential operation to ensure that all samples have the non-zero sampling probability.

[0111] Secondly, the historical mutual samples are combined with the computed sample priorities into the sample item and stored into the experience playback pool. The experience playback pool is of a priority queue structure, the leaf nodes are used for storing the sample items, and the non-leaf nodes are used for storing the priority sum of subtrees. This structure can effectively manage the priority of the samples, and therefore, the samples with higher importance can be preferentially selected during sampling.

[0112] After the samples are stored, the sampling probability is computed based on the sample priority. The sampling probability refers to the possibility of selecting the samples from the experience playback pool, and the sampling probability will monotonically increase along with the increase of the sample priority. In this way, hierarchical sampling can be realized, and important samples can be used more frequently in the training process.

[0113] Thirdly, the importance weight of each sample in the training sample batch is computed. The importance weight is used for compensating the sample distribution deviation caused by priority sampling. The computation mode of the weight can be flexibly adjusted through adjustable importance sampling indexes so as to meet different training requirements.

[0114] After the training sample batch is inputted into the online network, the time-order difference loss function with the weight is subjected to weight computation based on the importance. The loss function reflects the difference between model prediction and an actual target, and the gradient of the loss function about online network

parameters can be computed by the back propagation algorithm. In this process, the model parameters can be effectively updated, so that the learning effect of the model is improved.

[0115] Through the solution of the present application,

[0116] the sample utilization efficiency can be improved: the important samples are
 5 used more frequently by priority sampling, so that the learning process of the model is accelerated. The stability of the model can be enhanced: the added importance weights can effectively compensate the deviation of sample distribution and reduce the fluctuation in the training process. The training effect can be optimized: the loss function with weights makes the model pay more attention to the important samples when updating parameters,
 10 so that the overall prediction accuracy is improved.

[0117] In an optional embodiment, the setting an adaptive monitoring threshold based on the fusion feature vector, and deploying a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, performing difference analysis on the monitoring data and
 15 the original fusion feature vector, starting a hierarchical scheduling optimization strategy according to the difference degree, performing an incremental adjustment mode on local abnormality, and triggering strategy re-optimizing for global abnormality includes:

[0118] establishing a dynamic reference line based on the fusion feature vector, computing an adaptive adjustment factor in combination with system operation
 20 environment parameters, and multiplying the dynamic reference line by the adaptive adjustment factor to obtain an adaptive monitoring threshold;

[0119] deploying a three-level distributed monitoring network to collect the real-time system operation state data, the three-level distributed monitoring network being composed of a local monitoring agent, a regional monitoring node and a global monitoring center, and
 25 device-level state data collected by the local monitoring agent being gathered by the regional monitoring node and then transmitted to the global monitoring center;

[0120] performing feature extraction and fusion processing on the received real-time system operation state data by the global monitoring center so as to generate a real-time fusion feature vector, and comparing the real-time fusion feature vector with the adaptive
 30 monitoring threshold;

[0121] when the real-time fusion feature vector deviates from the adaptive monitoring threshold, computing a standardized difference value of the real-time fusion feature vector relative to the initial fusion feature vector, and computing a comprehensive difference degree based on the standardized difference value; and

[0122] grading the comprehensive difference degree according to a preset difference degree threshold, when the comprehensive difference degree is less than the difference degree threshold, determining to be the local abnormality and triggering the increment adjustment strategy, and when the comprehensive difference degree is greater than the difference degree threshold or equal to the difference degree threshold, determining to be the global abnormality and triggering the re-optimization strategy.

[0123] Firstly, the basis of the fusion feature vector is established. The multi-dimensional feature extraction is performed on the system operation state data to form a comprehensive feature vector. The feature vector contains information such as device operation parameters, environmental factors and historical data, thus ensuring the comprehensiveness and accuracy of the feature vector.

[0124] Secondly, the adaptive monitoring threshold is set. A dynamic reference line is computed according to the fusion feature vector and the system operation environment parameters. The dynamic reference line is set according to the normal device operation state and can reflect the performance change of the device in different environments. The adaptive adjustment factor is adjusted according to the fluctuation condition of the real-time monitoring data to ensure that the monitoring threshold can reflect the actual operation state of the device in real time.

[0125] Thirdly, the three-level distributed monitoring network is deployed. The network is composed of the local monitoring agent, the regional monitoring node and the global monitoring center. The local monitoring agent is used for collecting device-level state data and transmitting the data to the regional monitoring node. The regional monitoring node summarizes and preliminarily analyzes the data from the plurality of local monitoring agents and finally transmits the processed data to the global monitoring center.

[0126] In the global monitoring center, the received system real-time operation state data is subjected to feature extraction and fusion processing to generate the real-time fusion feature vector. This vector is compared with the adaptive monitoring threshold to determine whether the device operation state is normal or not.

[0127] When the real-time fusion feature vector deviates from the adaptive monitoring threshold, it is needed to compute a standardized difference value between the real-time fusion feature vector and the initial fusion feature vector. The difference value reflects the deviation degree between the current state and the normal state. Based on this difference value, the comprehensive difference degree is further computed to assess the abnormal condition of the device.

[0128] The comprehensive difference degree is graded according to the preset difference degree threshold. When the comprehensive difference degree is less than the difference degree threshold, it is determined to be local abnormality, and the incremental adjustment strategy is triggered to perform small-range parameter adjustment so as to
 5 recover the normal state of the device; and when the comprehensive difference degree is greater than or equal to the difference degree threshold, it is determined to be the global abnormality, and the re-optimization strategy is triggered to perform comprehensive system optimization and adjustment.

[0129] Through the solution of the present invention,

10 [0130] the real-time system monitoring capability can be improved, the abnormal condition of the device can be found and processed in time, and thus the failure rate is reduced. Through the set adaptive monitoring threshold, the flexibility and adaptability of the monitoring system are ensured, and the system can adapt to the operation states in different environments. Through the implementation of the hierarchical scheduling
 15 optimization strategy, the local abnormality and the global abnormality can be handled more efficient, and therefore, the overall operation efficiency of the system is improved.

[0131] In an optional embodiment, the taking the optimized scheduling effect data as a new training sample, and performing distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so
 20 as to realize continuous optimization of the scheduling strategy model includes:

[0132] collecting scheduling effect assessment data as a training sample, the scheduling effect assessment data containing system state vectors before and after scheduling execution, an actually executed scheduling action sequence, resource utilization rate data and energy consumption data, and performing normalization preprocessing on the
 25 scheduling effect assessment data to obtain a standardized training sample;

[0133] distributing the standardized training sample to the local training node by the federated learning framework, the local training node constructing a local model copy containing the dual-channel neural network and the dual-target Q network;

[0134] at the local training node, inputting the state feature vector in the standardized
 30 training sample into a state representation network of the dual-channel neural network so as to obtain a state code, and generating a new scheduling action by an action generation network of the dual-channel neural network based on the state code;

[0135] comparing the new scheduling action with an actual execution action in the standardized training sample, computing a strategy gradient, and updating parameters of

the action generation network;

[0136] inputting the state code and the new scheduling action into the dual-target Q network, computing a resource utilization rate Q value and an energy efficiency Q value, weighting the resource utilization rate Q value and the energy efficiency Q value according to performance index values in the standardized training sample to obtain a comprehensive Q value, and updating the parameters of the dual-target Q network based on the comprehensive Q value;

[0137] periodically triggering a federated average algorithm, collecting model parameters of the local training node for global aggregation to obtain global model parameters, and distributing the global model parameters to the local training node to update the local model copy; and

[0138] collecting new scheduling effect assessment data as incremental training samples, and continuously optimizing the dual-channel neural network and the dual-target Q network through the reinforcement learning framework.

[0139] Firstly, the scheduling effect assessment data are collected as the training samples. The data include the system state vectors before and after scheduling execution, the actually executed scheduling action sequence, the resource utilization rate data and the energy consumption data. And normalized preprocessing is performed on the data to obtain the standardized training sample. The purpose of normalization processing is to eliminate the influence of different dimensions and ranges on model training, so that the data are compared and learned under the same standard.

[0140] Secondly, the standardized training samples are distributed to the local training nodes by the federated learning framework. At each local training node, the local model copy containing the dual-channel neural network and the dual-target Q network is constructed. According to this step, model training can be independently performed on each node, and moreover, the data privacy is protected.

[0141] In the local training node, the state feature vector in the standardized training sample is inputted into the state representation network of the dual-channel neural network to obtain the state code. Based on the state code, the action generation network of the dual-channel neural network generates a new scheduling action. In this process, the current scheduling decision is optimized by learning the historical scheduling effect.

[0142] Thirdly, the generated new scheduling action is compared with the actual execution action in the standardized training sample, the strategy gradient is computed, and the parameters of the action generation network are updated. In this way, the model can

continuously adjust and optimize the decision strategy of the model to improve the scheduling effect.

[0143] Fourthly, the state code and the new scheduling action are inputted into the dual-target Q network, and the resource utilization rate Q value and the energy efficiency Q value are computed. According to the performance index value in the standardized training sample, the resource utilization rate Q value and the energy efficiency Q value are weighted to obtain the comprehensive Q value. The parameters of the dual-target Q network are updated based on the comprehensive Q value, thereby ensuring that the model reaches the optimal balance between resource utilization and energy efficiency.

[0144] The federated average algorithm is triggered regularly, and the model parameters of the local training node are collected for global aggregation, so as to obtain global model parameters. The global model parameters are distributed to the local training node, and the local model copy is updated. In this process, it is ensured that models of all nodes can share learning results, thus improving the performance of the whole model.

[0145] Fifthly, new scheduling effect assessment data is collected to serve as the incremental training sample, and the dual-channel neural network and the dual-target Q network are continuously optimized by the federated learning framework. By continuous iteration and updating, the model can adapt to new scheduling environments and requirements, and the efficient scheduling strategy is kept.

[0146] Through the solution of the present invention:

[0147] the accuracy and efficiency of the scheduling strategy can be improved, the resource utilization rate and energy efficiency can be optimized, and the energy consumption can be reduced. The data privacy is protected by federated learning, and the security risk caused by centralized storage of data is avoided. Continuous optimization of the model can be realized, the scheduling environment of dynamic change is adapted, and thus the flexibility and response capability of the system are improved.

[0148] In an optional embodiment, the inputting the state code and the new scheduling action into the dual-target Q network, computing a resource utilization rate Q value and an energy efficiency Q value, weighting the resource utilization rate Q value and the energy efficiency Q value according to performance index values in the standardized training sample to obtain a comprehensive Q value, and updating the parameters of the dual-target Q network based on the comprehensive Q value includes:

[0149] splicing the state code and the new scheduling action to obtain a state action vector, the state action vector being used as a common input of a resource utilization rate

assessment network and an energy efficiency assessment network;

[0150] inputting the state action vector into the resource utilization rate assessment network, performing nonlinear transformation through a first hidden layer and a second hidden layer of the resource utilization rate assessment network to obtain a resource utilization rate feature, and computing the resource utilization rate Q value based on the resource utilization rate feature;

[0151] a computation formula of the resource utilization rate Q value being as follows:

$$[0152] \quad Q_r(s_t, a_t) = W_3 h_2 + b_3,$$

[0153] in which, $Q_r(s_t, a_t)$ refers to the resource utilization rate Q value, s_t refers to the state code at the moment t, a_t refers to the action at the moment t, W_3 refers to a weight matrix at the output layer, h_2 refers to the resource utilization rate feature, and b_3 refers to a bias scalar at the output layer;

[0154] inputting the state action vector into the energy efficiency assessment network, performing nonlinear transformation through the first hidden layer and the second hidden layer of the energy efficiency assessment network to obtain an energy efficiency feature, and computing the energy efficiency Q value based on the energy efficiency feature,

[0155] a computation formula of the energy efficiency Q value being as follows:

$$[0156] \quad Q_e(s_t, a_t) = W_3^e h_2^e + b_3^e ;$$

[0157] in which, $Q_e(s_t, a_t)$ refers to the energy efficiency Q value, W_3^e refers to the weight matrix at the output layer of the energy efficiency assessment network, h_2^e refers to the energy efficiency feature, and b_3^e refers to the bias scalar at the output layer of the energy efficiency assessment network;

[0158] computing a relative change amount of the resource utilization rate based on performance indexes in a current batch of training samples, inputting the relative change amount of the resource utilization rate into an adaptive weight computation unit, and computing the resource utilization rate and the energy efficiency weight by the adaptive weight computation unit;

[0159] multiplying the resource utilization rate Q value by the resource utilization rate weight to obtain a weighted resource utilization rate Q value, multiplying the energy efficiency Q value by the energy efficiency weight to obtain a weighted energy efficiency Q value, and adding the weighted resource utilization rate Q value to the weighted energy efficiency Q value to obtain the comprehensive Q value; and

[0160] constructing the time-order difference loss function based on the comprehensive Q value, and updating network parameters of the resource utilization rate

assessment network and the energy efficiency assessment network at the same time through the time-order difference loss function.

[0161] Firstly, the state code and new scheduling action are processed. The state code vector and the scheduling action vector are spliced to generate the state action vector. For example, the dimension of state code is 128, the dimension of scheduling action is 64, and the state action vector with dimension of 192 is obtained after splicing. The vector is used as the common input feature of the resource utilization assessment network and the energy efficiency assessment network.

[0162] For the resource utilization assessment network, the three-level neural network structure is adopted. In the first hidden layer, it contains 256 neurons, and ReLU activation function is utilized to perform nonlinear transformation; in the second hidden layer, it contains 128 neurons, and the ReLU activation function is also utilized to obtain the resource utilization feature vector. The feature vector is inputted to the output layer with 1 neuron so as to obtain the resource utilization Q value assessment result. In practical application, it is assumed that the resource utilization Q value of a certain state action combination is 0.85.

[0163] In the energy efficiency assessment network, it uses the same network architecture. The energy efficiency feature vector is obtained by nonlinear transformation at the two hidden layers, and is inputted to the output layer to obtain the energy efficiency Q value. For example, the energy efficiency Q value of the same state action combination is 0.72. Two assessment networks share the parameters at the input layer, but have independent hidden layer and output layer parameters.

[0164] The relative change of the resource utilization rate is computed based on the current training sample batch. If the average resource utilization rate of the current batch of samples is 75%, and the average resource utilization rate of the previous batch of samples is 70%, the relative change amount will be 7.14%. The variable quantity is inputted into the adaptive weight computation unit, and the weight coefficients of the two networks are computed according to the preset weight adjustment rule. When the resource utilization rate obviously increases, the weight of resource utilization rate assessment is increased. For example, it is computed to obtain the resource utilization rate weight of 0.6 and the energy efficiency weight of 0.4.

[0165] The resource utilization rate Q value is multiplied by the corresponding weight to obtain the weighting result of 0.51, and the energy efficiency Q value is multiplied by the weight to obtain the weighting result of 0.29. The two weighted Q values are added to

obtain a comprehensive Q value of 0.80. The comprehensive Q value reflects the resource utilization efficiency and also takes into account the energy consumption level.

[0166] Finally, the time-order difference loss function is constructed, and the difference between the current comprehensive Q value and the target Q value is taken as the optimization target. The gradient is computed by the back propagation algorithm, and the parameters of the two assessment networks are updated at the same time. The parameter optimization is performed by the Adam optimizer, and the learning rate is set to be 0.001.

[0167] Through the solution of the present invention:

[0168] in optimization effect: the joint optimization of the resource utilization rate and the energy efficiency is realized through the dual-target Q network structure; the dynamic balance of the two optimization targets is ensured by the adaptive weight mechanism; and redundant computation of feature extraction is reduced by the shared input layer design. In algorithm performance: the extraction capability of state action features is improved by the deep neural network; the expression capability of the model is enhanced by multi-level nonlinear transformation; and the professionalism of different optimization targets is ensured by the independent assessment network. In engineering practicability, the network structure design is reasonable and facilitate actual deployment; the weight adaptive mechanism can be flexibly adjusted according to actual requirements; and the selection of assessment indexes meets the requirements of actual application scenarios.

[0169] FIG. 2 is a schematic structural diagram of a reinforcement learning algorithm based intelligent scheduling optimization system for heterogeneous metadata according to an embodiment of the present invention, as shown in FIG. 2, the system includes:

[0170] a first unit, configured to construct a multi-level dynamic feature extraction mechanism, acquire real-time metadata information in a multi-source heterogeneous metadata system by a distributed data collection engine, and perform feature decomposition on the real-time metadata information by a self-attention mechanism to obtain a feature matrix set of metadata, the feature matrix set including a type feature matrix, a time-order feature matrix, a relationship feature matrix and a load feature matrix; input the feature matrix set into a transformer encoder for feature fusion to generate a fusion feature vector, and construct a feature vector space through an adaptive dimension reduction algorithm; and train a scheduling strategy model of a dual-channel neural network structure based on the feature vector space, a graph convolution neural network channel extracting spatial features based on the type feature matrix and the relationship feature matrix, and a gating circulation unit network channel extracting time-order features based on the time-order

feature matrix and the load feature matrix and outputting initial scheduling strategy parameters;

[0171] a second unit, configured to establish a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and generate a scheduling action candidate set by taking the fusion feature vector as state input; design a multi-target assessment function to assess the candidate action, respectively compute a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system, perform weighted combination on the system load balance degree, the access delay score, the resource utilization rate score and the data locality score to obtain a multi-dimensional system performance comprehensive score, and construct a comprehensive reward function based on the multi-dimensional system performance comprehensive score; perform strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism, and output the optimized strategy parameters as a scheduling decision sequence; and

[0172] a third unit, configured to input the scheduling decision sequence into a distributed scheduling execution engine, and execute a metadata scheduling operation through a two-stage submission protocol; set an adaptive monitoring threshold based on the fusion feature vector, and deploy a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, perform difference analysis on the monitoring data and the original fusion feature vector, start a hierarchical scheduling optimization strategy according to the difference degree, perform an incremental adjustment mode on local abnormality, and trigger strategy re-optimizing for global abnormality; and take the optimized scheduling effect data as a new training sample, and perform distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model.

[0173] In a third aspect,

[0174] an electronic device provided by the embodiment of the present invention includes:

[0175] a processor; and

[0176] a memory configured to store processor executable instructions;

[0177] the processor is configured to call instructions stored in the memory to execute

the abovementioned method.

[0178] In a fourth aspect,

[0179] the embodiment of the present invention provides a computer-readable storage medium having computer program instructions stored therein, when executed by a processor, the computer program instructions execute the abovementioned method.

[0180] The present invention can be a method, an apparatus, a system, and/or a computer program product. The computer program product can include a computer-readable storage medium that contains computer-readable program instructions for executing the present invention in various aspects.

[0181] Finally, it is to be noted that: the above embodiments are only used for illustrating the technical solution of the present invention, not to limit it. Although the present invention is described in detail with reference to the aforementioned embodiments, those of ordinary skill in the art should understand that it may still modify the technical solutions recorded in the aforementioned embodiments, or replace some or all of the technical features thereof. These modifications or substitutions do not make the essence of the corresponding technical solution out of the scope of the technical solutions of each embodiment of the present invention.

CLAIMS

We claim:

1. A reinforcement learning algorithm based intelligent scheduling optimization method for heterogeneous metadata, comprising:

5 constructing a multi-level dynamic feature extraction mechanism, acquiring real-time metadata information in a multi-source heterogeneous metadata system by a distributed data collection engine, and performing feature decomposition on the real-time metadata information by a self-attention mechanism to obtain a feature matrix set of metadata, the feature matrix set comprising a type feature matrix, a time-order feature matrix, a relationship feature matrix and a load feature matrix; inputting the feature matrix set into a transformer encoder for feature fusion to generate a fusion feature vector, and constructing a feature vector space through an adaptive dimension reduction algorithm; training a scheduling strategy model of a dual-channel neural network structure based on the feature vector space, a graph convolution neural network channel extracting spatial features based on the type feature matrix and the relationship feature matrix, and a gating circulation unit network channel extracting time-order features based on the time-order feature matrix and the load feature matrix and outputting initial scheduling strategy parameters;

 establishing a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and generating a scheduling action candidate set by taking the fusion feature vector as state input; designing a multi-target assessment function to assess the candidate action, respectively computing a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system, performing weighted combination on the system load balance degree, the access delay score, the resource utilization rate score and the data locality score to obtain a multi-dimensional system performance comprehensive score, and constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score; performing strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism, and outputting the optimized strategy parameters as a scheduling decision sequence; and

 inputting the scheduling decision sequence into a distributed scheduling execution engine, and executing a metadata scheduling operation through a two-stage submission protocol; setting an adaptive monitoring threshold based on the fusion feature vector, and

deploying a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, performing difference analysis on the monitoring data and the original fusion feature vector, starting a hierarchical scheduling optimization strategy according to the difference degree, performing an incremental adjustment mode on local abnormality, and triggering strategy re-optimizing for global abnormality; and taking the optimized scheduling effect data as a new training sample, and performing distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model.

2. The method according to claim 1, wherein the establishing a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and generating a scheduling action candidate set by taking the fusion feature vector as state input; designing a multi-target assessment function to assess the candidate action, respectively computing a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system comprises:

constructing a three-level cascade linkage optimization architecture based on the initial scheduling strategy parameters, the three-level cascade linkage optimization architecture comprising a strategy level and an execution level, a deep deterministic strategy gradient network of a dual exploration mechanism being built in the strategy level, and a self-correction four-dimensional performance assessment model being built in the execution level; the deep deterministic strategy gradient network being of a four-level fully-connected structure, the number of nodes at an input layer being the same as the dimension of a fusion feature vector, and the number of nodes at an output layer being the same as the dimension of an action space;

inputting the fusion feature vector into the deep deterministic strategy gradient network, performing hierarchical Bayesian modeling on the parameter distribution of the deep deterministic strategy gradient network to obtain hierarchical network weight distribution, acquiring a plurality of groups of network parameter samples from each level of the hierarchical network weight distribution by a Thompson sampling method to form a parameter sample set, and adding Gaussian noise disturbance into the parameter sample set; respectively loading the parameter sample set added with the Gaussian noise disturbance to the deep deterministic strategy gradient network to generate a plurality of strategy network instances, independently generating the scheduling action candidate set by the plurality of

strategy network instances based on the initial scheduling strategy parameters; and

inputting the scheduling action candidate set into a four-dimensional performance assessment model at the execution level, setting a self-correction threshold for an assessment result of each dimension by the four-dimensional performance assessment model, and triggering a re-assessment mechanism when the assessment result exceeds the self-correction threshold; and respectively computing the load balance degree score, the access delay score, the resource utilization rate score and the data locality score by the four-dimensional performance assessment model.

3. The method according to claim 1, wherein the constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score; performing strategy optimization by a dual-target Q network structure, an online network updating strategy parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism comprises:

constructing a comprehensive reward function based on the multi-dimensional system performance comprehensive score, and computing a difference value between the multi-dimensional system performance comprehensive scores of the current moment and the previous moment to obtain a performance change value; setting the multi-dimensional system performance comprehensive score as a basic reward item, and setting the performance change value as a priority reward item; adaptively adjusting a combination weight of the basic reward item and the priority reward item based on the accumulated reward and the average reward in the training process so as to obtain a comprehensive reward function;

performing strategy optimization by the dual-target Q network structure, and constructing the online network and the target network which share the network structure; inputting the current system state into the online network to obtain a current action value, and inputting the system state of the next moment into the target network to obtain a target state value; substituting the current action value, the target state value and the comprehensive reward function into a Bellman equation to compute to obtain a time-order difference error;

computing a sample priority based on an absolute value of the time-order difference error; storing a historical mutual sample and the corresponding sample priority into an experience playback pool; performing importance sampling according to the sample priority to obtain a training sample batch; inputting the training sample batch into the online network, and updating strategy parameters of the online network in real time

through a back propagation algorithm based on the time-order difference error; and

updating the strategy parameters of the target network by a soft update mechanism, and performing weighted average on the updated strategy parameters of the online network and the updated strategy parameters of the target network to obtain updated parameters of the target network, in which, the weight of weighted average is determined by a preset soft update coefficient, and the stability of updating the target network parameters is kept through the soft update coefficient.

4. The method according to claim 3, wherein the computing a sample priority based on an absolute value of the time-order difference error; storing a historical mutual sample and the corresponding sample priority into an experience playback pool; performing importance sampling according to the sample priority to obtain a training sample batch; inputting the training sample batch into the online network, and updating strategy parameters of the online network in real time through a back propagation algorithm based on the time-order difference error comprises:

adding the absolute value of the time-order difference error with a priority reference value, and then performing exponential operation to obtain the sample priority, the priority reference value being used for ensuring that all samples have non-zero sampling probability;

forming a sample item by the historical mutual sample and the sample priority and storing the sample item into the experience playback pool that is of a priority queue structure, the priority queue structure including a leaf node and a non-leaf node, the leaf node storing the sample item, and the non-leaf node storing a priority sum of subtrees;

computing sampling probability based on the sample priority, the sampling probability monotonically increasing along with the sample priority, and performing hierarchical sampling according to the sampling probability to obtain the training sample batch;

computing an importance weight of each sample in the training sample batch, the importance weight being used for compensating a sample distribution deviation caused by priority sampling, and the importance weight being computed through an adjustable importance sampling index; and

inputting the training sample batch into the online network, computing a time-order difference loss function with a weight based on the importance weight, and computing a gradient of the loss function about the online network parameters through the back propagation algorithm,

a gradient computation formula of the network parameters being as follows:

$$\nabla_{\theta} L = \frac{1}{N} \sum_{i=1}^N w_i (y_i - Q(s_i, a_i; \theta)) \nabla_{\theta} Q(s_i, a_i; \theta);$$

in which, $\nabla_{\theta} L$ refers to the gradient of the loss function about a parameter θ , N refers to the size of the experience playback pool, w_i refers to the importance weight of an i^{th} sample, y_i refers to a target Q value of the i^{th} sample, $Q(s_i, a_i; \theta)$ refers to a predicted Q value of the online network about a state-action pair (s_i, a_i) , and $\nabla_{\theta} Q$ refers to the gradient of the Q value about the parameter θ .

5. The method according to claim 1, wherein the setting an adaptive monitoring threshold based on the fusion feature vector, and deploying a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, performing difference analysis on the monitoring data and the original fusion feature vector, starting a hierarchical scheduling optimization strategy according to the difference degree, performing an incremental adjustment mode on local abnormality, and triggering strategy re-optimizing for global abnormality comprises:

15 establishing a dynamic reference line based on the fusion feature vector, computing an adaptive adjustment factor in combination with system operation environment parameters, and multiplying the dynamic reference line by the adaptive adjustment factor to obtain an adaptive monitoring threshold;

20 deploying a three-level distributed monitoring network to collect the real-time system operation state data, the three-level distributed monitoring network being composed of a local monitoring agent, a regional monitoring node and a global monitoring center, and device-level state data collected by the local monitoring agent being gathered by the regional monitoring node and then transmitted to the global monitoring center;

25 performing feature extraction and fusion processing on the received real-time system operation state data by the global monitoring center so as to generate a real-time fusion feature vector, and comparing the real-time fusion feature vector with the adaptive monitoring threshold;

30 when the real-time fusion feature vector deviates from the adaptive monitoring threshold, computing a standardized difference value of the real-time fusion feature vector relative to the initial fusion feature vector, and computing a comprehensive difference degree based on the standardized difference value; and

grading the comprehensive difference degree according to a preset difference degree

threshold, when the comprehensive difference degree is less than the difference degree threshold, determining to be the local abnormality and triggering the increment adjustment strategy, and when the comprehensive difference degree is greater than the difference degree threshold or equal to the difference degree threshold, determining to be the global abnormality and triggering the re-optimization strategy.

6. The method according to claim 1, wherein the taking the optimized scheduling effect data as a new training sample, and performing distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model comprises:

collecting scheduling effect assessment data as a training sample, the scheduling effect assessment data containing system state vectors before and after scheduling execution, an actually executed scheduling action sequence, resource utilization rate data and energy consumption data, and performing normalization preprocessing on the scheduling effect assessment data to obtain a standardized training sample;

distributing the standardized training sample to the local training node by the federated learning framework, the local training node constructing a local model copy containing the dual-channel neural network and the dual-target Q network;

at the local training node, inputting the state feature vector in the standardized training sample into a state representation network of the dual-channel neural network so as to obtain a state code, and generating a new scheduling action by an action generation network of the dual-channel neural network based on the state code;

comparing the new scheduling action with an actual execution action in the standardized training sample, computing a strategy gradient, and updating parameters of the action generation network;

inputting the state code and the new scheduling action into the dual-target Q network, computing a resource utilization rate Q value and an energy efficiency Q value, weighting the resource utilization rate Q value and the energy efficiency Q value according to performance index values in the standardized training sample to obtain a comprehensive Q value, and updating the parameters of the dual-target Q network based on the comprehensive Q value;

periodically triggering a federated average algorithm, collecting model parameters of the local training node for global aggregation to obtain global model parameters, and distributing the global model parameters to the local training node to update the local

model copy; and

collecting new scheduling effect assessment data as incremental training samples, and continuously optimizing the dual-channel neural network and the dual-target Q network through the reinforcement learning framework.

5 7. The method according to claim 6, wherein the inputting the state code and the new scheduling action into the dual-target Q network, computing a resource utilization rate Q value and an energy efficiency Q value, weighting the resource utilization rate Q value and the energy efficiency Q value according to performance index values in the standardized training sample to obtain a comprehensive Q value, and updating the parameters of the
10 dual-target Q network based on the comprehensive Q value comprises:

splicing the state code and the new scheduling action to obtain a state action vector, the state action vector being used as a common input of a resource utilization rate assessment network and an energy efficiency assessment network;

inputting the state action vector into the resource utilization rate assessment network,
15 performing nonlinear transformation through a first hidden layer and a second hidden layer of the resource utilization rate assessment network to obtain a resource utilization rate feature, and computing the resource utilization rate Q value based on the resource utilization rate feature;

a computation formula of the resource utilization rate Q value being as follows:

20
$$Q_r(s_t, a_t) = W_3 h_2 + b_3 ;$$

in which, $Q_r(s_t, a_t)$ refers to the resource utilization rate Q value, s_t refers to the state code at the moment t, a_t refers to the action at the moment t, W_3 refers to a weight matrix at the output layer, h_2 refers to the resource utilization rate feature, and b_3 refers to a bias scalar at the output layer;

25 inputting the state action vector into the energy efficiency assessment network, performing nonlinear transformation through the first hidden layer and the second hidden layer of the energy efficiency assessment network to obtain an energy efficiency feature, and computing the energy efficiency Q value based on the energy efficiency feature,

a computation formula of the energy efficiency Q value being as follows:

30
$$Q_e(s_t, a_t) = W_3^e h_2^e + b_3^e ;$$

in which, $Q_e(s_t, a_t)$ refers to the energy efficiency Q value, W_3^e refers to the weight matrix at the output layer of the energy efficiency assessment network, h_2^e refers to the energy efficiency feature, and b_3^e refers to the bias scalar at the output layer of the energy

efficiency assessment network;

computing a relative change amount of the resource utilization rate based on performance indexes in a current batch of training samples, inputting the relative change amount of the resource utilization rate into an adaptive weight computation unit, and
 5 computing the resource utilization rate and the energy efficiency weight by the adaptive weight computation unit;

multiplying the resource utilization rate Q value by the resource utilization rate weight to obtain a weighted resource utilization rate Q value, multiplying the energy efficiency Q value by the energy efficiency weight to obtain a weighted energy efficiency Q value, and
 10 adding the weighted resource utilization rate Q value to the weighted energy efficiency Q value to obtain the comprehensive Q value; and

constructing the time-order difference loss function based on the comprehensive Q value, and updating network parameters of the resource utilization rate assessment network and the energy efficiency assessment network at the same time through the time-order
 15 difference loss function.

8. A reinforcement learning algorithm based intelligent scheduling optimization system for heterogeneous metadata, applied to the method according to any one of claims 1 to 7, comprising:

a first unit, configured to construct a multi-level dynamic feature extraction
 20 mechanism, acquire real-time metadata information in a multi-source heterogeneous metadata system by a distributed data collection engine, and perform feature decomposition on the real-time metadata information by a self-attention mechanism to obtain a feature matrix set of metadata, the feature matrix set including a type feature matrix, a time-order feature matrix, a relationship feature matrix and a load feature matrix; input the feature
 25 matrix set into a transformer encoder for feature fusion to generate a fusion feature vector, and construct a feature vector space through an adaptive dimension reduction algorithm; and train a scheduling strategy model of a dual-channel neural network structure based on the feature vector space, a graph convolution neural network channel extracting spatial features based on the type feature matrix and the relationship feature matrix, and a gating
 30 circulation unit network channel extracting time-order features based on the time-order feature matrix and the load feature matrix and outputting initial scheduling strategy parameters;

a second unit, configured to establish a hierarchical optimization mechanism based on deep reinforcement learning according to the initial scheduling strategy parameters, and

generate a scheduling action candidate set by taking the fusion feature vector as state input; design a multi-target assessment function to assess the candidate action, respectively compute a system load balance degree score, an access delay score, a resource utilization rate score and a data locality score by an adaptive weight system, perform weighted
 5 combination on the system load balance degree, the access delay score, the resource utilization rate score and the data locality score to obtain a multi-dimensional system performance comprehensive score, and construct a comprehensive reward function based on the multi-dimensional system performance comprehensive score; perform strategy optimization by a dual-target Q network structure, an online network updating strategy
 10 parameters in real time based on the comprehensive reward function, a target network keeping training stability through a soft update mechanism, and output the optimized strategy parameters as a scheduling decision sequence; and

a third unit, configured to input the scheduling decision sequence into a distributed scheduling execution engine, and execute a metadata scheduling operation through a
 15 two-stage submission protocol; set an adaptive monitoring threshold based on the fusion feature vector, and deploy a distributed monitoring network to collect system operation state data in real time; when the monitoring data deviates from the adaptive monitoring threshold, perform difference analysis on the monitoring data and the original fusion feature vector, start a hierarchical scheduling optimization strategy according to the
 20 difference degree, perform an incremental adjustment mode on local abnormality, and trigger strategy re-optimizing for global abnormality; and take the optimized scheduling effect data as a new training sample, and perform distributed incremental training on the dual-channel neural network and the dual-target Q network through a federated learning framework, so as to realize continuous optimization of the scheduling strategy model.

25 9. An electronic device, including
 a processor; and
 a memory configured to store processor executable instructions; and
 the processor is configured to call instructions stored in the memory to execute the method according to any one of claims 1 to 7.

30 10. A computer-readable storage medium having computer program instructions stored therein, wherein when executed by a processor, the computer program instructions execute the method according to any one of claims 1 to 7.