Knowledge-Based Systems 116 (2017) 26–38

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Accelerating bio-inspired optimizer with transfer reinforcement learning for reactive power optimization

Xiaoshun Zhang[a], Tao Yu[a,*], Bo Yang[b], Lefeng Cheng[a]

[a] College of Electric Power, South China University of Technology, Guangzhou, 510640, China
[b] Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming, 650504, China

ARTICLE INFO

Article history:
Received 16 April 2016
Revised 27 August 2016
Accepted 29 October 2016
Available online 1 November 2016

IKeywords:
Accelerating bio-inspired optimizer
Transfer reinforcement learning
Memory matrix
Cooperating multi-bion
WoLF-PHC
Reactive power optimization

ABSTRACT

This paper proposes a novel accelerating bio-inspired optimizer (ABO) associated with transfer reinforcement learning (TRL) to solve the reactive power optimization (RPO) in large-scale power systems. A memory matrix is employed to represent the memory of different state-action pairs, which is used for knowledge learning, storage, and transfer among different optimization tasks. Then an associative memory is introduced to significantly reduce the dimension of memory matrix, in which more than one element can be simultaneously updated by the cooperating multi-bion. The win or learn fast policy hill-climbing (WoLF-PHC) is also used to accelerate the convergence. Thus, ABO can rapidly seek the closest solution to the exact global optimum by exploiting the prior knowledge of the source tasks according to their similarities. The performance of ABO has been evaluated for RPO on IEEE 118-bus system and IEEE 300-bus system, respectively. Simulation results verify that ABO outperforms the existing artificial intelligence algorithms in terms of global convergence ability and stability, which can raise one order of magnitude of the convergence rate than that of others.

1. Introduction

Nonlinear programming is ubiquitous in power system operation, such as reactive power optimization (RPO) [1], units commitment (UC) [2], optimal power flow (OPF) [3], economic dispatch (ED) [4], etc. Several conventional optimization approaches have been employed to solve this issue, including Newton method [5], quadratic programming [6], interior-point method [7], etc. However, they require an accurate system model and may merely obtain a local optimum if system nonlinearities, discontinuous functions and constraints, and functions with multiple local-minima exist [8].

In order to reduce the full dependence of an accurate system model, an enormous variety of elegant artificial intelligence (AI) algorithms, such as artificial bee colony (ABC) [9], group search optimizer (GSO) [10], ant colony system (ACS) [11], particle swarm optimization (PSO) [12], genetic algorithm (GA) [13], and reinforcement learning (RL) [14], have been applied for the optimal operation of power systems, which can improve the convergence of global optimum. In general, these algorithms can be classified into the following four types [15]: (a) evolution-based algorithms; (b) physics-based algorithms; (c) swarm-based algorithms; and (d) human-based algorithms. The evolution-based algorithm are inspired by the natural evolution rules, e.g., GA inspired by Darwinian evolution [16], biogeography-based optimization (BBO) inspired by the geographical distribution of biological species [17]. The physics-based algorithms are originated from the physical rules in the universe, e.g., gravitational search algorithm (GSA) derived from the law of Gravity and the notion of mass interactions [18], big bang-big crunch (BB-BC) derived from the theory of the universe evolution [19], ray optimization derived from the Snell's light refraction law [20]. The swarm-based algorithms stem from the social behavior of groups of animals, e.g., ABC [9], ACS [10], and PSO [11] are inspired from the social behavior of bee colony, ant colony, and bird flocking, respectively. The human-based algorithms motivated by the human behaviors, e.g., teaching-learning-based optimization (TLBO) based on the philosophy of the teaching-learning process [21], exchange market algorithm (EMA) based on the shares trading in the stock market [22], group counselling optimizer (GCO) enlightened from the human social behaviour in solving social problems through counselling within a group [23]. Unfortunately, as most of these approaches is incapable of recording the prior knowledge, a relatively long computation time is resulted in when dealing with a new optimization task. Consequently, it is difficult to achieve a fast dynamic opti-

* Corresponding author.
E-mail addresses: xszhang1990@sina.cn (X. Zhang), taoyu1@scut.edu.cn, yutao99@tsinghua.org.cn (T. Yu), yangbo_ac@outlook.com (B. Yang), chenglf_scut@163.com (L. Cheng).

http://dx.doi.org/10.1016/j.knosys.2016.10.024
0950-7051/

mization of a large-scale power system which optimization tasks may vary along with the time.

Recently, transfer learning becomes more and more popular in data mining and machine learning due to its merits of fast resolution of similar tasks via exploiting the prior knowledge [24]. It has been found that RL can increase the learning rate through a transfer learning [25,26], thus AI and behaviour psychology inspired transfer reinforcement learning (TRL) has been developed, which can be classified into a behaviour transfer and an information transfer [27]. In order to accelerate the learning rate of the reinforcement learning tasks, a right inter-task mapping was constructed for a transfer learning between a task and a different but relevant task with different actions and state variables [28]. Moreover, a novel transfer learning was proposed in [29] based on subgoal discovery and subtask similarity. In addition, a transfer method was presented in [30] which attempts to leverage the weights from function approximators specifying action-value functions via inter-task mapping method. Furthermore, the transfer was achieved based on the idea that related tasks usually share some common features [31]. In [32], a generalized policy was proven to be a more effective approach to transfer learning compared to policy library by experiment based. Besides, the knowledge transfer was also adopted to accelerate agent's learning rate and coordination ability for multiagent reinforcement learning by [33]. As one of the most famous RL, Q-learning [34] can also be used for the behaviour transfer, in which the Q-value table is defined as the memory matrix for the knowledge learning, storage, and transfer in this paper. However, it consumes a long time to obtain an optimal memory matrix as only a single RL agent is activated for the state space exploration in the environment, which may even lead to the curse of dimension when solving a complex task [35]. To tackle this problem, this paper adopts a modified Q-learning [36] to achieve a higher learning rate of behaviour transfer via combining the conventional Q-learning with the win or learn fast policy hill-climbing (WoLF-PHC). Moreover, the use of cooperative learning of Ant-Q [37] attempts to accelerate the update of memory matrix.

Based on the optimization mechanism of existing AI algorithms, a novel fast optimization method with TRL called accelerating bio-inspired optimizer (ABO) is proposed, which has the following four advantages against to the existing AI algorithms as follows:

- The bions can obtain knowledge through a consistent interaction with the external environment, in which the knowledge can be fully stored in the memory matrix and transferred to different optimization tasks.
- The cooperating learning and WoLF-PHC can jointly improve the update efficiency of memory matrix, while the use of associative memory can effectively avoid the curse of dimensionality of RPO in a large-scale power system.
- ABO with TRL can efficiently exploit the prior knowledge for online optimization according to the deviations of active power demand between source tasks and a new task.
- Compared with the approximate ideal multi-objective solution Q(λ) (AIMS-Q(λ)) learning [38], this paper focus on the convergence acceleration for complex tasks by introducing a memory matrix, cooperating learning, WoLF-PHC, and TRL, respectively. In contrast, AIMS-Q(λ) is developed for multi-objective optimization by only combing the conventional Q(λ)-learning with the improved technique for order preference similar to an ideal solution (TOPSIS) method, which usually encounters a relatively low learning efficiency and the curse of dimension as its performance is completely depended on the conventional Q(λ)-learning.

The remaining of this paper is organized as follows. Section 2 presents the basic principles of ABO. ABO with TRL for fast RPO is developed in Section 3. Simulation results obtained on IEEE 118-bus system and IEEE 300-bus system are given in Section 4. Finally, Section 5 concludes the paper.

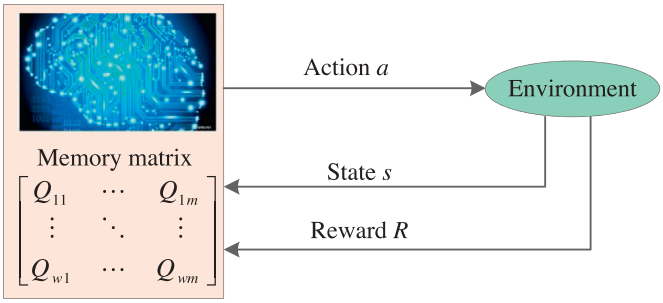

Fig. 1. The principle of memory matrix used in Q-learning.

2. Accelerating bio-inspired optimizer

2.1. Memory matrix and TRL

In this section, three main parts of ABO, including the memory matrix, associative memory, and TRL, are introduced as follows: (a) The memory matrix is established via persistent interactions between the bions and the environment, which is adopted for knowledge learning, storage, and transfer; (b) The associative memory is employed to effectively handle the curse of dimension by decomposing the extremely large-scale action set into multiple small-scale action set; and (c) TRL is adopted to realize the knowledge transfer.

2.1.1. Bion with memory matrix

A bion is an intelligent agent mimicking the cooperation of creatures in nature, which is proposed to achieve associative memory, knowledge learning, storage, and transfer. Due to such promising features, the bion has the ability of self-learning and knowledge learning in a dynamic environment compared with that of other popular swarm intelligence (SI) algorithms, i.e., ABC [9], ACS [11], and PSO [12].

The Q-value table is defined as the memory matrix of each bion illustrated in Fig. 1, while each element of the memory matrix represents a memory of the corresponding state-action pair, i.e., $\mathbf{Q}(s,a)$, which can be updated with the feedback reward from the interaction between the bions and the environment [34]. The memory of each state-action pair is used to estimate the discounted sum of future rewards started from the current state and action policy. Moreover, each bion specifies a stimulus-response pattern to select its action based on the memory matrix, such that the expected long-term rewards in each state can be maximized. In a given state, a higher memory of the element indicates an action which tends to obtain a larger reward. After the bions undergo sufficient actions in the state space, an optimal memory matrix will be obtained and adopted for the knowledge transfer.

2.1.2. Associative memory for dimension reduction

It can be found from Fig. 2 that the curse of dimension will emerge if the number of controllable variables grows too large in conventional Q-learning. Assume the number of alternative actions for a controllable variable x_i is m_i, then the dimension of action set $|A| = m_1 m_2 \cdots m_n$, where n is the number of controllable variables. If n increases significantly, the dimension of the memory matrix will become extremely high, which inevitably results in a slow convergence or even a calculation failure. Currently, the hierarchical reinforcement learning (HRL) [39,40], has been designed to avoid the curse of dimension via decomposing a complicated task