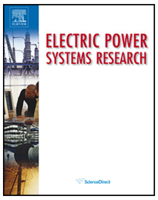
Electric Power Systems Research 141 (2016) 179–190

Contents lists available at ScienceDirect

Electric Power Systems Research

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



Reinforcement learning approach for congestion management and cascading failure prevention with experimental application

Sina Zarrabian, Rabie Belkacemi*, Adeniyi A. Babalola

Electrical and Computer Engineering Department, Center for Energy Systems Research (CESR), Tennessee Technological University, United States

ARTICLE INFO

Article history:
Received 16 February 2016
Received in revised form 28 May 2016
Accepted 25 June 2016

Keywords:
Blackout
Cascading failure
Q-learning
Real-time
Reinforcement learning
Smart grids

ABSTRACT

This article proposes a method based on the reinforcement learning (RL) for preventing cascading failure (CF) and blackout in smart grids by acting on the output power of the generators in real-time. The proposed research work utilizes the Q-learning algorithm to train the system for the optimal action selection strategy during the state-action learning process by updating the action values based on the obtained rewards. The trained system then is able to relieve congestion of transmission lines in real-time by adjusting the output power of the generators (actions) to prevent consecutive line outages and blackout after N-1 and N-1-1 contingency conditions. The proposed RL-based control is validated through experimental implementation as well as simulation studies on the IEEE 118-bus test system for different contingency case studies. The results obtained from the experimental and simulation studies show the accuracy and robustness of the proposed approach in preventing cascading failure and blackout.

1. Introduction

The smart grid concept based on communication and information technology infrastructures has significantly improved the performance of modern and wide-area power systems in the past few decades. However, due to the large interconnections, complexity of power grids, and sophisticated control structures, the major concern with smart grids is still dealing with these vulnerabilities to enhance stability, reliability, and security. One of the catastrophic challenges in power systems is cascading failure (CF), where a single fault or contingency in the system can initiate a series of unexpected outages and disturbances that can lead to total wide-area blackout. For instance, the 2003 North American blackout [1,2] or the 2011 Southwest blackout in the USA [3] demonstrates the lack of appropriate infrastructure in the system components for taking rapid and accurate actions to prevent the spreading of failures. This crucial challenge in power systems imposes extravagant maintenance and restoration costs on governments and electric power industries. Hence, there is great interest in the modeling and prevention of cascading failures to mitigate the negative impacts associated with this phenomenon and to improve the stability of power grids.

As a means to investigate and analyze cascading failures and their different aspects in power systems, various methods and strategies have been developed [4–12] (e.g., risk assessment and probability reduction of cascading failures and blackouts [4,5], SASE model [6], AC and DC power flow model [7], interaction model [8], improved OPA model [9]).

In the SASE model [6], the authors presented a reduced dynamic model of an extensive power system, considering limited state variables to include critical characteristics of the grid. Authors in Ref. [8] introduced an interaction model for cascading failures. This probabilistic model identifies the critical components of the system that propagate cascading failures, and an interaction matrix is acquired based on the interaction of the component failures. This model investigates the risk of cascading failures and provides online decision making. In the OPA model [9], the cascading failure was approximated by considering the dynamics of demands and DC load flow. Linear programming was utilized to re-dispatch the generation and loads after a random line outage. The drawback of the OPA model is that the timing of failures is ignored, which cannot be suitable for the protective coordination studies. The CASCADE model introduced in [11] is based on load dependency. In this model, all elements of the power system are considered to be identical, and failure of each element has an equal impact on the other elements. The CASCADE model does not include all electrical features of the grid. Branching Process is another probabilistic model for analyzing cascading failures [12]. This model relies on the probability distribution for investigating the total component failures and does not provide sufficient individual dynamic variables to incorporate all dynamic characteristics associated with cascading failures and blackouts in power systems.

* Corresponding author.
E-mail address: rbelkacemi@tntech.edu (R. Belkacemi).

http://dx.doi.org/10.1016/j.epsr.2016.06.041
0378-7796/© 2016 Elsevier B.V. All rights reserved.

In addition to the above-mentioned methods in modeling cascading failures for predictive purposes, recent research works have been presented based on load shedding strategies for cascading failure prevention [13,14]. The main drawback associated with the load shedding method is that customers will be deprived of power, which makes the various stakeholders in the power industry incur losses.

On the other hand, some emerging algorithms in Artificial Intelligence (AI) and machine learning such as the multi-agent systems have been widely utilized recently to enhance the power system stability, reliability, and performance [15–19]. The main characteristics of intelligent systems are controllability, adaptability, simplicity, and fast response even for complicated structures. Among the machine learning-based methods, the reinforcement learning (RL) approach is a powerful method that has recently had various applications in power system control [20–24]. The literature review shows that using AI and machine learning approaches for cascading failure and blackout prevention is a novel topic and is in its early steps of development with few research works in this area of study [25,26].

This article proposes a novel RL control approach based on the Q-learning algorithm for adaptive adjustment of the generated power from different generators to prevent cascading transmission line outages and blackout after N-1 and N-1-1 contingency conditions without using any load shedding. In fact, for transmission line contingency in the grid, the system can learn how to adjust the power within the voltage, frequency, angle, and power flow constraints and manage the congestion of transmission lines to prevent initialization of further cascading outages in a continuous and smooth manner rather than discrete and sudden drop of loads. In addition to the computer simulations, this article presents an experimental implementation of the proposed approach on a hardware replica of the high-voltage (HV) side of the IEEE 30-bus system to fill the gap in the literature in terms of experimental analysis and validation. The applicability of the proposed RL method for the large-scale IEEE 118-bus power system is verified by simulation studies as well. The advantages of this method are its accuracy in targeting the congested lines, rapid response, reliability, and adaptiveness.

This article is organized as follows. In Section 2, an introduction to RL method, Q-learning algorithm, and its implementation process for this work are presented. In Section 3, experimental testbed and case studies are discussed. In Section 4, experimental and simulation results are presented and discussed. Finally, the conclusion is presented in Section 5.

2. Reinforcement learning approach

2.1. Reinforcement learning method and Q-learning algorithm

Reinforcement Learning is a type of machine learning method in which an agent (controller unit) interacts with the environment (process) by means of states, actions, and rewards to learn an optimal policy to reach a pre-defined target. In fact, at each time step, by transition from state to action, a reward is received by the agent. The main objective of the reinforcement learning is to discover an optimal policy in which the expected cumulative rewards are maximized [27].

The reinforcement learning process associated with a set of states and actions (state and action space) in addition to the reward function is defined as a Markov decision process (MDP) [27]. An MDP with finite state and action space (FMDP) is presented by a tuple $S, A, P^a_{s^n s^{n+1}}, \rho^a_{s^n s^{n+1}}$, where S is the state space, A is the action space, $P^a_{s^n s^{n+1}} : S^n \times A \times S^{n+1} \to [0, 1]$ is defined as the transition probability function presenting transition from the n^{th} state s^n to the next state s^{n+1}, $\rho^a_{s^n s^{n+1}} : S^n \times A \times S^{n+1} \to \mathbb{R}$ is the reward function that defines the immediate reward after the state transition. At each step of iteration t, the agent receives the state signal of the environment $s_t \in S$ and an action $a_t \in A$ is selected based on the action selection probability $p(s_t, a_t)$ which is the policy that action a_t is selected in state s_t. Then, according to $P(s_t, a_t, s_{t+1})$, state transits from s_t to $s_{t+1} \in S$ and a reward r_{t+1} is generated based on $\rho(s_t, a_t, s_{t+1})$.

In the RL theory, several actions might be selected by each agent in each specific state. Then, for each selected action, a reward is generated for the evaluation of the taken action. Next, the goal is to maximize the expected discounted returns (R_t) including the cumulative rewards during the interaction by:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum\nolimits_{n=0}^{\infty} \gamma^n r_{t+n+1} \tag{1}$$

where $\gamma \in [0, 1]$ is discount factor.

The essential part in almost all RL algorithms is to estimate the value functions consisting of state-value and action-value functions [28]. The state-value and action-value functions are defined by (2) and (3), respectively:

$$V^{\pi}(s) = E_{\pi}\left\{\sum_{n=0}^{\infty} \gamma^n r_{t+n+1} | s_t = s\right\} \tag{2}$$

$$Q^{\pi}(s, a) = E_{\pi}\left\{\sum_{n=0}^{\infty} \gamma^n r_{t+n+1} | s_t = s, a_t = a\right\} \tag{3}$$

The state-value function is described by the expected following returns regarding the policy π in the state s. However, the action-value function is presented by the expected returns for the state s and chosen action a and then is subsequently followed by the policy $\pi(s, a)$.

The Q-learning algorithm is one of the well-known model-free techniques based on Temporal Difference (TD) learning for solving the RL problem. In Q-learning algorithm, the action-value function (Q-value) defined in (3) is obtained. Eventually, the derived expression of Q-learning based on the Bellman optimality function is described by:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[r_{t+1} + \gamma \underset{a}{max} Q_t(s_{t+1}, a) - Q_t(s_t, a_t)] \tag{4}$$

where $\alpha \in [0, 1]$ is the learning rate. $Q_t(s_t, a_t)$ is initialized (estimated) first and action a_t is selected in state s_t. Then, the following state s_{t+1} is obtained with an immediate reward r_{t+1} and the $\underset{a}{\max} Q_t(s_{t+1}, a)$ associated with the new state s_{t+1} is calculated. Next, the error is calculated as $r_{t+1} + \gamma \underset{a}{\max} Q_t(s_{t+1}, a) - Q_t(s_t . a_t)$ and then the Q-value is updated to $Q_{t+1}(s_t, a_t)$. Basically, a lookup Q-table (Q-matrix) is utilized for storing and updating the expected future values (Q-function). The Q-function is typically stored in a Q-table and indexed by state and action. Then, by initializing arbitrary values, the optimal Q-function is approximated iteratively. The table entry for state-action is updated according to (4). In order to obtain an optimal Q-value, there should be a balance between the exploration and exploitation which is called exploration-exploitation trade-off [27]. This means that all possible actions should be considered in every state with nonzero probability. To ensure appropriate trade-off, the softmax (Boltzmann exploration) action selection strategy is used where the probability of selected actions is weighted based on their Q-values by:

$$P(a|s) = \frac{e^{\frac{Q(s,a)}{T}}}{\sum_a e^{\frac{Q(s,a)}{T}}} \tag{5}$$

where $T \geq 0$ is the temperature parameter. Lower values of T will lead the action selection policy to more greedy strategy and higher values will cause more random strategy. Practically, the value of T