2009 World Congress on Computer Science and Information Engineering

Research and Application of Reinforcement Learning Based on Constraint MDP in Coal Mine

ZHAO Xiao-hu, ZHAO Ke-ke, WANG Qing-qing, Ma Fang-qing
xiaohuzhao@126.com, zkhm99@163.com
College of information and electrics, China University of Mining & Technology, Jiangsu
Xuzhou, 221008 China

Abstract

Reinforcement learning is an algorithm without model which is learning what to do--how to map situations to actions--so as to maximize a numerical reward signal. Reinforcement learning provides an available method to the systems, which are very difficult to build up accurate models around complex environment. But now many practical problems demand a maximum reward with not much cost (expense). For example, the production of coal mine is closely correlated with security in that it increases production in the limited range of security situation. On the base of Markov decision process (MDP) and reinforcement learning, the paper introduced constraint Markov decision process into reinforcement learning. The paper improved Q-learning algorithm with adding cost factor and gave a new Q-learning algorithm based on constraint MDP. Finally, according to the constraint between production and safety in coal mine, the paper made the simulation investigation about the action control of coal shearer in coal mine working face. The simulation result had verified the validity of the method.

Key words: *constraint MDP, reinforcement learning, Q-learning, cost, coal shearer*

1. Introduction

Reinforcement learning is an important method of machine learning. It has been applied in fields of intelligent control, robotics, analysis and forecasting, etc. It is an algorithm without model which is learning what to do--how to map situations to actions--so as to maximize a numerical accumulated reward from the environment[1].

The aim of reinforcement learning is to find a strategy $\boldsymbol{\pi}$ or choose rules of action, to make the greatest value of the reward expected. However, in many practical problems, people do not only demand the maximal reward, but also hope that the price (cost) is not too more.

Especially for coal mine industry, the fundamental prerequisite of various subsystems' operation is security. In the course of mine production, many subsystems have such security constraint conditions. For example, coal mining is closely related to the situation of underground security. Generally speaking, the more coal output, the better effectiveness the coal mine has. While the more output of the work face, the more gas emission it has and the worse security situation becomes. So mine production and security is a mutual restricted unity.

For this type of problem, it is that various systems should operate more rationally and effectively under the conditions of security. The description of using reinforcement learning is constraint optimization problem of control the costs within a certain scope and pursuing the greatest rewards.

2. Constraint model of MDP

Definition of constraint model of MDP is expressed as formula (1)[2,3].

$$\text{CMDP}=\{S, A, P, r, c, \omega, V, C\} \tag{1}$$

The definition of parameters in formula (1) is expressed as follows.

S: Non-empty state set composed of all the possible statuses of system, i.e. state set of the environment.

A: System action set, $A(i)$ is a usable decision set at state i, $i \in S$ .Usually use the symbol of a to express decision(action) of $A(i)$.

P: Transformation function of states. $P(s, a, s')$ expresses the probability of the system in the state s adopts action a to transform the state to s'.

r: Expectation of state action. $r(s,a,s')$is instantaneous reward value gotten from the system that adopts action a in the state of s to transform the state to s'.

978-0-7695-3507-4/08 $25.00 © 2008 IEEE
DOI 10.1109/CSIE.2009.587

c: Cost of state action. *c(s,a,s')* is instantaneous cost value paid from the system that adopts action a in the state of s to transform the state to *s'*.
ω: Given real number, that expresses the maximum that expectation allowable cost.
V: Guidelines function, that expresses the total expectation discount remuneration.
C: Guidelines function, that expresses the total expectation discount cost.

Unlimited stage total expectation discount remuneration and cost under strategy $\pi(\pi \in \prod)$ are respectively:

$$V(s,\ \pi)=E_{\pi}[\sum_{t=0}^{\infty}\beta^{t}\cdot r(s_t,\ \pi(s_t)\,|\,s_0=s)],\ \ i\in S \tag{2}$$

$$C(s,\ \pi)=E_{\pi}[\sum_{t=0}^{\infty}\beta^{t}\cdot c(s_t,\ \pi(s_t)\,|\,s_0=s)],\ \ i\in S \tag{3}$$

$\prod_a=\{\pi\in\prod: C(\pi)\leq\omega\}$is viable strategy set. If $\pi^*\in\prod_a$ satisfies $V(\pi^*)\geq V(\pi)$ $(\pi\in\prod)$, π^* is called constraint optimal strategy.

The goal of constraint MDP is to identify an optimal strategy $\pi^*(s)$ under any state *s*, and maximize reward under conditions of *ω*.

3. Q-learning based on Constraint MDP

Q-learning, brought forward by Watkins in 1989, is the most representative reinforcement learning method, which is similar to dynamic programming algorithm. It provides intelligent system with a learning ability in Markov environment by using experienced action sequences to choose the best action.

Provided the state of system at t moment is S_t, the action that agent chooses is a_t. When the system shifts to next state S_{t+1}, the instantaneous remuneration and the instantaneous cost of agent are respectively r_t and c_t. The goal of agent is to select optimized action, and maximize accumulated discount rewards of the whole process within the cost constraint condition. Provided discount factor β $(0\leq\beta\leq1)$, from the moment *t,* the accumulated discount sum of agent's remuneration and cost are respectively formula(4)and (5).

$$r_{(t)}=r_t+\beta r_{t+1}+\beta^2 r_{t+2}+\cdots+\beta^n r_{t+n}+\cdots \tag{4}$$

$$c_{(t)}=c_t+\beta c_{t+1}+\beta^2 c_{t+2}+\cdots+\beta^n c_{t+n}+\cdots \tag{5}$$

For each state and action $S_t,\ a_t$, total discount cost is formula(6).

$$C(s_t,\ a_t)=r(s_t,\ a_t)+\beta\sum_{s_{t+1}\in S}p(s_{t+1}\,|\,s_t,a_t)C(s_{t+1}) \tag{6}$$

If the condition of $C(S_t,,\ a_t) < \omega$ is satisfied, total expectation discount remuneration is formula(7).

$$Q^*(s_t,\ a_t)=r(s_t,\ a_t)+\beta\sum_{s_{t+1}\in S}p(s_{t+1}\,|\,s_t,a_t)\{\max_{a_{t+1}\in A}[Q^*(s_{t+1},\ a_{t+1})]\} \tag{7}$$

$$\pi^*(s_t)=\arg\max_{a_t\in A}Q^*(s_t,\ a_t) \tag{8}$$

$Q^*(S_t, a_t)$ is the sum of expectation discount reward according to constraint optimal strategy by choosing action a_t at state S_t. If using the method of Q-learning, formula (6) and (7) could be modified as follows[4].

$$C(s_t,\ a_t)=(1-\alpha)C(s_t,\ a_t)+\alpha[c(s_t,\ a_t)+\beta C(s_{t+1})] \tag{9}$$

$$Q^*(s_t,\ a_t)=(1-\alpha)Q^*(s_t,\ a_t)+\alpha[r(s_t,\ a_t)+\beta V^*(s_{t+1})]$$
(10)

Here α is learning factor, and β is discount factor.

According to the above description, we modify the algorithm of Q-learning. Q-learning algorithm based on constraint MDP is as follows:

```
(1) Set parameters $\alpha,\beta$ and set $t=0$,initial set $A$ of action set, reward matrix $r$, and cost matrix $c$.
(2) Initialize matrix $Q$ and matrix $C$.
(3) For each episode
      $S_t$ ←Current state,
      Do while (Do while not reach goal state)
        ● Select a feasible action $a$ of the present state.
        ● Execute action $a$ to enter next state $S_{t+1}$, where $r$ is accepted reward and $c$ is produced cost. If $C(s_t,,\ a_t) \geqslant \omega$, to select action a again.
        ● Calculate $C$ ($s_t$, $a_t$) and $Q$ ($s_t$, $a_t$) as follows.
$C(s_t,\ a_t)=(1-\alpha)C(s_t,\ a_t)+\alpha[c(s_t,\ a_t)+\beta C(s_{t+1},\ a_{t+1})]$
In the condition of $C(s_t,\ a_t) < \omega$, set
$Q^*(s_{t+1},\ a_{t+1})=\max_{a_{t+1}\in A}\sum Q(s_{t+1},a_{t+1})$
$Q(s_t,\ a_t)=(1-\alpha)Q(s_t,\ a_t)+\alpha[r(s_t,\ a_t)+\beta Q(s_{t+1},\ a_{t+1})]$
        ● Set next state as present state.
      End do
    End for
```

4. Application of Q-learning based on constrained MDP in working face system

Working face is the forefront of coal mine production. The production of working face relates directly to the economic benefits of coal mine. Moreover, working face production is closely related to safety. In coal mining, the geological condition and production condition are different in every coal mine working face, and even in different working faces of the same coal mine. This paper takes S4-3 working face of a coal mine in Shanxi province in China to investigate.

4.1. Parameters of working face

Coal seam thickness of S4-3 working face is 6.27 meters, mining height is 3.5 meters, length of coal wall is 3.5 meters, obliquity of coal seam is 2°-6°, and mining length is 1216 meters. Average obliquity tends to be horizontal. The working face is designed to produce 10000 tons coal everyday.

4.2. Mining method

S4-3 working face uses fully-mechanized coal winning technology of longwall mining with sublevel caving, adopting MGTY400/930 electric haulage coal shearer.

Mining coal shearer is the main production equipment of mining at present. Nowadays, various coal mines mainly use electrical traction shearer, most of which adopt frequency conversion control that can adjust traction speed of shearer according to the actual situation. The parameters of MGTY400/930 shearer are as follows. Mining height is 3.5 ± 0.1m, cutting depth is 0.8m, caving average height is 2.77 m, the ratio between mining and caving is 1:1.27, average caving section is 177 m, bottom slice recovery ratio is 98%, and top coal recovery rate is *85%*. One caving and mining is a cycle.

In actual mining, average velocity of shearer is 3.1 m/min. Therefore, it needs about one hour to complete a cycle of mining.

4.3. Model construction

To simplify the problem, the simulation of this paper assumed that there was no faultage in the coal mining region, mine pressure was well and equipments worked normally. The main consideration was using Q-learning to control the traction speed of the shearer in one mining cycle under constraints of production environment security.

According to formula (11):

$$P=60HJ\rho Vq \quad (11)$$

In the formula:

P--actual production capacity of shearer, 730t/h

H--mining height, 3.5m

J--cutting depth, 0.8m

P--entity density of coal, 1.4t/m^3

Vq--actual average traction speed of shearer (m/min)

It can be known from formula (11), that the yield of working face is proportional with the traction speed of shearer when mining technics is identified and the rotate speed of drum is definite[5,6]. Meanwhile, according to the gas production mechanism, more production of coal and more gas emission, so the security situation becomes worse during this time for the higher of gas thickness.

The optimal strategy of this paper is: under the environmental safety restraint in coal face, the performance of working face production is optimal. That is to say, according to production environmental safety situation, the shearer runs reasonably and effectively (get the highest yield under the guarantee security) with a high boot-strap rate, avoiding the situation of exceeding gas emission limit.

4.3.1. State set. To simplify the problem, gas thickness of the working face is seen as the main factor in model. Provided gas thickness of the working face is T.

According to the specific situation of coal mine production, the state set of system is $S=\{S1,S2,S3,S4\}$, in which:

$S1$: good security situation;

$S2$: general security situation;

$S3$: warning (i.e. within the scope of security, but thickness gas is higher);

$S4$: danger (i.e. gas thickness exceeds limits).

According to provisions of gas thickness in fully mechanized coal mine face, warning value is *0.8%*. Then assume that:

If $T<0.6\%$ then security situation is good,

If $0.6\%\leq T\leq 0.8\%$ then security situation is general,

If $0.8\%<T<1.0\%$ then security situation is warning,

If $1.0\%\leq T$ then security situation is dangerous.

4.3.2. Action set. Action set of system is $A=\{V1,V2,V3,V4\}$,in which:

$V1$: Shearer runs fast.

$V2$: Shearer runs in mid-speed.

$V3$: Shearer runs in low-speed.

$V4$: Shearer stops running.

For working face production, shearer cuts coal in front, meanwhile top caving coal mines behind the support. If shearer's speed is too fast, it would yield great gas emission and incomplete top caving. Consequently the total production is influenced. According to the actual situation at spot, assume that:

Shearer runs fast: $V1=4m/min$,

Shearer runs in mid-speed: $V2=3m/min$,

Shearer runs in low-speed: $V3=1.5m/min$,

Shearer stops running: $V4=0$.

4.3.3. Reward and Cost. Reward value is mainly related with mining capacity. More coal is mined if the speed is fast. Provided reward value of action is: *V1* is 4, *V2* is 3, *V3* is 1.5, *V4* is 0.

Cost value relates with state after the execution of action. Provided cost value of state is: *S1* is 0, *S2* is 5, *S3* is 20, *S4* is 100. In which, if selected action is stopping, produced cost is 0.

4.3. Environment state simulation

Shearer's action has a direct impact on production environment. It directly relates to the cost value produced, as well as the selection of the next action.

According to the coal mine production, the space of working face is definite and wind velocity does not change much in a certain period of time. Therefore, more coal mined in unit time, greater gas thickness is. When the velocity of shearer changes, gas thickness produced is different[5]. This paper used an approximate method to calculate gas thickness after shearer ran as formula (12).

$$T_i' = \frac{C(T_i - T_0)}{V_v} \times (V_i - V_v) + T_i \qquad (12)$$

In which, T_i' is gas thickness produced under chosen running velocity.

T_i is gas thickness produced under average running velocity.

T_o is average gas thickness in examining and repairing time.

V_i is running velocity that shearer selects.

T_v is average traction velocity of shearer, provided here is *3.1 m/min*.

C is gas emission coefficient.

Gas emission of working face is produced by shearer mining and top coal-caving, the simulation of this paper mainly simulated the influence on environment of shearer running, and assumed *C = 0.612* according to the actual situation of this working face. The aim of formula (12) is mainly to reflect the trend of gas thickness changes, and further to simulate environment state.

4.4. Simulation

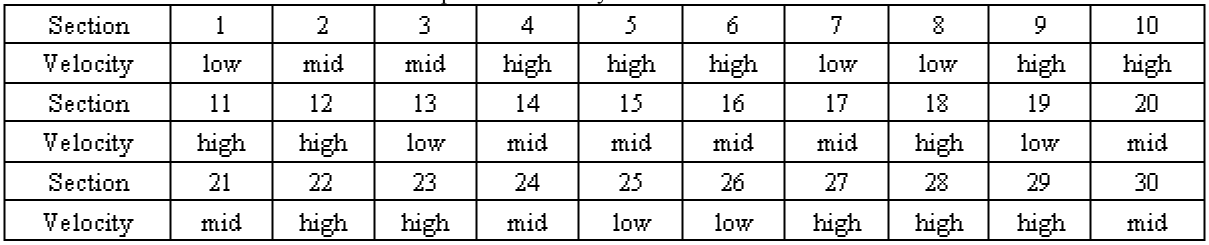
This paper used the data of S4-3 working face on Nov.15th, 2007 to analyze a work cycle on that day. Firstly, gas thickness of working face was ascertained before being mined. According to the production arrangement of the mine, from 0 o'clock to 16 o'clock of every day is production time, and from 16 o'clock to 24 o'clock is examining and repairing time. According to the data of security monitoring system, it can be known that average gas thickness in examining and repairing time (stopping production) is *0.38%*.

For working face, in different areas, gas content of coal seam may be different. The length of S4-3 working face is 192 meters. Combined with geological situation of working face and safe environment monitoring information, the working face is divided into five running areas, of each there are six running sections. There are altogether 30 running sections. The optimal running velocity of shearer in each section is ascertained through constraint Q-learning, under the situation of security.

According to formula (9) and (10), assume *α=0.9*, *β=0.8*, the limit of cost is *60(ω =60)* as far as possible to guarantee that shearer works under safe situation. From security monitoring data, it can be known that, in actual production, gas thickness before shearer enters into each section is respectively *0.7%, 0.72%, 0.86%, 0.71%* and *0.84%*. The average gas thickness when shearer runs in each section is respectively *0.71%, 0.77%, 0.75%, 0.81%* and *0.85%*.

According to actual situation, the optimized running velocity of each section through constraint Q-learning can be seen from Table 1.

Table1. Optimized velocity of coal shearer in each section

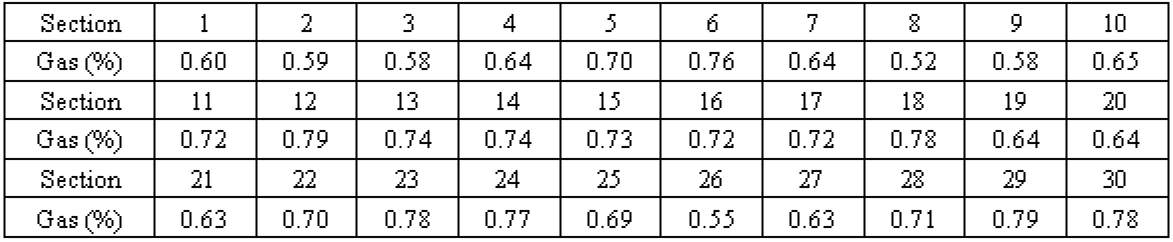


Section	1	2	3	4	5	6	7	8	9	10
Velocity	low	mid	mid	high	high	high	low	low	high	high
Section	11	12	13	14	15	16	17	18	19	20
Velocity	high	high	low	mid	mid	mid	mid	high	low	mid
Section	21	22	23	24	25	26	27	28	29	30
Velocity	mid	high	high	mid	low	low	high	high	high	mid

Security state (gas thickness) that corresponds with each section when shearer uses optimized running velocity can be known from Table 2. It is seen that gas thickness is in normal scope in each running section.

Under the guidance of constraint Q-learning method, shearer can be guaranteed to run continuously in safe state.

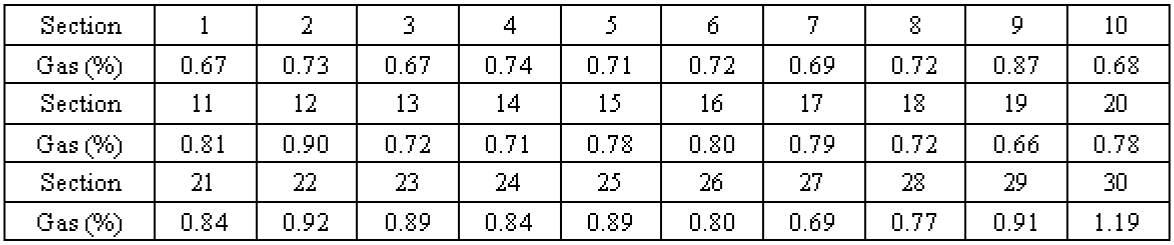
Table2. Gas thickness of each section when using optimized velocity



Section	1	2	3	4	5	6	7	8	9	10
Gas (%)	0.60	0.59	0.58	0.64	0.70	0.76	0.64	0.52	0.58	0.65
Section	11	12	13	14	15	16	17	18	19	20
Gas (%)	0.72	0.79	0.74	0.74	0.73	0.72	0.72	0.78	0.64	0.64
Section	21	22	23	24	25	26	27	28	29	30
Gas (%)	0.63	0.70	0.78	0.77	0.69	0.55	0.63	0.71	0.79	0.78

Security state (gas thickness) that corresponds with each section when shearer actually runs at local can be known from Table 3. It is seen that gas thickness is abnormal in some sections that the working face is in dangerous state.

Table3. Gas thickness of each section when using actual velocity



Section	1	2	3	4	5	6	7	8	9	10
Gas (%)	0.67	0.73	0.67	0.74	0.71	0.72	0.69	0.72	0.87	0.68
Section	11	12	13	14	15	16	17	18	19	20
Gas (%)	0.81	0.90	0.72	0.71	0.78	0.80	0.79	0.72	0.66	0.78
Section	21	22	23	24	25	26	27	28	29	30
Gas (%)	0.84	0.92	0.89	0.84	0.89	0.80	0.69	0.77	0.91	1.19

For traditional shearer control, the driver controls the operation of the shearer according to experience and fixed production arrangement. When the working face pushes into the region that contains more gas, it would fall in dangerous situation such as the gas thickness exceeding the limit. And shearer would be forced to cease running (electrical equipments are power off), and could not mine until the environment state is ensured to be safe. It could be seen from running and stopping records of actual shearer at local and security monitoring data that shearer stopped running for about 30 minutes after gas exceeded the prescribed limit (thickness reaches *1.19%*). Thus coal mine safety could not be effectively guaranteed and the progress of mining could not be completed on time because of the downtime also. As the great gas emission in the region, it needed about 70 minutes to finish a work cycle with optimal running velocity. But shearer was guaranteed in the state of continuous operation in security state without the forced downtime, so production efficiency was increased.

5. Conclusions

The paper studied Markov decision procedure and reinforcement learning. According to the characteristic of many practical problems that demand not only the maximal reward, but also the limit cost (expense), the paper used constraint Markov decision procedure and gave a new Q-learning algorithm based on constraint Markov decision. In the end, the paper made the simulation about control the action of coal shearer in coal mine working face. By comparing, it could be known that Q-learning method based on constraint MDP could pre-guide and judge shearer running according to security situation. This method provided a reasonably scheduling guide for the safe and optimal running of shearer.This paper used the method of environment model calculation to ascertain security degree. In later research, the method will combine with virtual mining technology and get more accurate and effective control strategy.

6. References

[1] R S Sutton and A.G. Barto. Reinforcement Learning. Cambridge, MA:MIT Press, 1998

[2] Hu-Qiying, Liu-Jianyong. Markov introduction to the decision-making process [M].Xi'an: Publisher of Xi'an University of Electronic Science and Technology, 2000

[3] Liu-Ke. Practical Markov decision process [M].Beijing: Publisher of Tsinghua University, 2004

[4] Watkins.J.C.H,Dayan.Qlearing.MachineLearning,1992, 8:279-292

[5] Xu-Yongqi. Mining[M].Xuzhou: Publisher of China University of Mining and Technology,2003

[6] Zhao Xiao-hu. Research on Key Technologies of Execution System of Mine Safe Production[D]. Xuzhou: China University of Mining and Technology, 2007