An Improved Immune Q-Learning Algorithm

Zhengqiao JI, Q.M.Jonathan.Wu, and Maher Sid-Ahmed

Abstract—**Reinforcement learning is a framework in which an agent can learn behavior without knowledge on a task or an environment by exploration and exploitation. Striking a balance between exploration and exploitation is one of the key problems of action selection in reinforcement learning. Exploitation causes the agent to reach a locally optimal policy quickly, whereas excessive exploration degrades the performance of the algorithm, though it may improve the learning performance and escape from a locally optimal policy. Recently the human immune systems have aroused researcher's interest due to its useful mechanisms which can be exploited for information processing in a complex cognition system. In this paper, we transplant some immune mechanisms into the basic Q-learning algorithm and convert Q-learning algorithm into a search for the optimum solution in combinatorial optimization. Experiments show that the improved Q-learning converges more quickly than Q-learning or Boltzmann exploration, and easily obtains the global solution set.**

I. Introduction

Reinforcement learning is a Framework in which an autonomous agent can learn behavior without knowledge on a task or an environment by progressively improving its performance based on given rewards when the learning task is performed. As one of the most rapidly developing machine learning methods [1], reinforcement learning includes the temporal difference algorithm proposed by Sutton [2] and the Q-learning of Watkins [3], [4]， which have been extensively used in many fields such as industrial control, time sequence schedule [5], robot navigation [6], and many more.

Although it can learn even in an unknown environment, reinforcement learning requires hundreds of trials of the given task. Additionally, large computation costs are expended by such exploration-oriented methods as Q-learning [7], which brings about the problem of needing a long time to converge to learning results. At the same time, people are looking for a way to reduce the convergence time when apply the Q-learning algorithm on some applications, such as robot navigation and online decision system. That is why to find the proper balance between exploration and exploitation in Q-learning is one of the most important things. Exploitation occurs if the action selection strategy is based purely on current values of the state-action pairs (SAPs), i.e., when the selection is greedy. In the case of most of the optimization problems, this will lead to locally optimal policies instead of a globally optimal one. On the contrary, exploration is the strategy based on the assumption that the agent selects a nonoptimal action in the current situation and allows it to neglect the locally optimal policies in order to reach the globally optimal one instead. The excessive exploration will drastically decrease the performance of a learning algorithm, and in some cases might have potentially disastrous consequences with respect to the learning results themselves [8].The trade-off dilemma between exploration and exploitation has been investigated by a lot of researchers previously. There are a lot of methods addressing this balance problem, such as dynamic-programming, learning automata and randomized strategies (including Boltzmann exploration) [9]. However, dynamic-programming approach provides formal guarantees only for the case of immediate reward and are not extensible in an obvious way to the delayed reinforcement case, and randomized strategies easily get no direction in exploration and converge unnecessarily slowly when the values of the actions are not separated enough.

A simple strategy proposed to deal with this problem on Q-learning is the simple ε-greedy (with $0 < \varepsilon< 1$), with larger ε corresponding to larger probability of exploration [8]. It shows good results on some problems when ε is set to nonzero. However, excessive exploration becomes unnecessary on fixed ε after a period of an initial interaction between the agent and the environment (assuming, of course, that the state-action-pair values are constant). While reinforcement learning is usually adopted as a real time learning method, excessive exploration will unavoidably cause the decreasing performance of reinforcement learning.

Therefore, it is meaningful to explore the possibility of improving the Q-learning approach by appropriately adjusting greedy search steps during the learning process. Through this method, it will not only improve the ability of the agent to acquire new knowledge from environment, but will also enhance the performance of the algorithm if it is set to the constant value of ε (and thus constant probability of exploration) [10]. In other words, for efficient and believable online performance, an exploration strategy also has to avoid cycling through previous solutions and know when to stop without getting stuck in a local optimum.

In this paper, the task of finding the optimal policy in Q-learning is transformed into search for an optimal solution in a combinatorial optimization problem. Then the hypermutation mechanism from immune system is applied to the search procedure in order to keep the balance between exploration and exploitation.

Manuscript received March 16, 2007.

Zhengqiao Ji is with the Electrical Engineering Department, University of Windsor, Windsor, Ontario N9b2m2 Canada (corresponding author to provide phone: 519-253-3000-4861; e-mail: ji9@ uwindsor.ca).

Q.M.Jonathan.Wu, was with National research center Canada. He is now with the Department of Electrical engineering, University of Windsor, Windsor, Ontario N9b2m2 Canada (e-mail: jwu@uwindsor.ca).

Maher Sid-Ahmed is with the Electrical Engineering Department, University of Windsor, Windsor, Ontario N9b2m2 Canada (e-mail: a@ uwindsor.ca).

1-4244-0991-8/07/$25.00/©2007 IEEE

The rest of the paper is organized as follows. Section II introduces the overview of reinforcement learning briefly and its main approach, Q-learning algorithm. Section III gives the outline of the immune mechanism which will be used in this paper. Section IV presents the improved immune algorithm based on immune scheme. Section V shows the experiments and simulation results and finally, we give our conclusions and suggestion for future work in Section VI.

.

II. Reinforcement Learning and Q-Learning Approach

In reinforcement learning, a learning agent interacts wit its environment by taking actions and accepting input from the environment. Input from the environment constitutes the state of the environment followed by an immediate reward. State information passed the agent summarizes all currently relevant information about the environment. The object of reinforcement learning is to maximize the rewards that a learning agent receives by improving its behaviors through the interaction between the agent and an environment using a reinforcement signal; these rewards are considered given by the environment.

Q-learning was presented as one of the most important reinforcement learning algorithms by Watkins in 1989 [4]. In this learning method, the learning process needs to update the state-action value Q(s, a) based on the reward received and the simplest One-step Q-learning can be described through the following Q-function:

$$Q(s,a) \equiv Q(s,a) + \Delta Q(s,a) \qquad \text{while}$$

$$\Delta Q(s,a) \equiv \alpha(r(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

$$s' = \delta(s,a)$$

Where *Q*(*s*, *a*) is the value function of the state-action pair (s, a) at moment t. αand γrefer to the learning rate and discount factor respectively, and r, s' and a' means the reward value function, next state and next action, respectively [8].

The one step Q learning algorithm can be described with following pseudo:

```
Initiate arbitrarily all Q(s,a) values;
For i=1: $N_e$
  Rand(s);
Do while ($s \neq S$)
    r=r(s, a)   ,$a \in A(s)$;
  $s' =$ δ(a,t);
  $Q(s,a) \equiv Q(s,a) + \alpha(r(s,a) + \gamma \max_{a'} Q(\delta(s',a') - Q(s,a))$
  $s = s'$;
end
  end
```

Rand: Choose a random (initial) state;

S: one of the goal states;

N_e : The desired numbers of episodes have been investigated.

The procedure for Q-learning, or the procedure for solving for *Q*(*s*, *a*) is to infinitely repeat the two-step loop: 1.select a particular state *s* and *a* particular action *a*, observe the immediate reward *r* for this state-action pair, and observe the resulting state *s*'; 2. keep adjusting Δ *Q*(*s*, *a*) until choosing the optimal action *s* among all possible actions that can be taken in the successor state *s*' leading to the greatest *Q*-value. For any stationary Markov Decision Problem, the Q-learning procedure is guaranteed to converge to the correct values, provided that the learning rate parameter is decreased over time with an appropriate schedule. Each state-action pair must be visited sufficiently often for learning to converge since convergence of reinforcement learning stochastically depends on visiting every state infinitely often. The mechanism in this one step Q-learning algorithm adopts a greedy strategy with pure exploitation which basically selects the optimal action according to the current state-action pair. It tries to maintain a better speed of convergence with a simple calculation. However, it is not hard to see that it generally difficult to obtain satisfactory results and after some iterations the search progress will be trapped in local optima since agent chooses the next action based on instant reward. Although the algorithm modified with ε-greedy strategy can search for nonoptimal actions to make the agents gain a more accurate knowledge, it may results in the decrease of system performance due to the over exploration with the fixed probability ε. Thus it is necessary to improve the Q-learning algorithm by adjusting greedy search adaptively, which helps with convergence not only in running towards global optima but also in guiding when to stop for saving searching time.

III. Immune Mechanism

The immune system is a complex of cells, molecules and organs with the primary role of limiting damage to the host organism by antigens (Ag), which elicit an immune response. One type of response is the secretion of antibody molecules by B cell. Antibodies (Ab) are Y-shaped receptor molecules bound on the surface of a B cell with the primary role of recognizing and binding, through a complementary match, with an antigen. Details about the immune network theory and terminology are found in [11] and [12].

In the normal course of the immune system evolution, the strength and specificity of the Ag-Ab interaction are measured by the affinity of their match [13]. The initial exposure to an Ag that stimulates an adaptive immune response is handled by a small number of low-affinity B cells. The effectiveness of the immune response to secondary encounters is enhanced considerably by the presence of memory cells associated with the first infection, capable of producing high-affinity Ab's just after subsequent encounters. From this procedure, we can see that learning in the immune system involves raising the population size and the affinities of those lymphocytes (B cells) that have proven themselves to be valuable during the antigen recognition phase. Thus, the immune repertoire is raised from a random base to a repertoire that more clearly reflects the actual antigenic environment [14].

In a T-cell dependent immune response, the repertoire of Ag-activated B cells is diversified basically by two mechanisms: hypermutation and receptor editing, which are used to introduce diversity during an immune response. Receptor editing offers the ability to escape from local optima on an affinity landscape. If a particular Ab (Ab1) is selected during a primary response, then somatic mutations allow the immune system to explore local areas around Ab1 by making small steps toward an Ab with higher affinity, leading to a local optima (Ab1*). Because mutations with lower affinity are lost, the Abs tend to go up the hill. Receptor editing allows an Ab to take large steps through the landscape, landing in a locale where the affinity might be lower (Ab2). However, occasionally the leap will lead to an Ab on the side of a hill where the climbing regions is more promising (Ab3), reaching the global optimum. From this locale, hypermutations can drive the Ab to the top of the hill (Ab3*). In conclusion, somatic mutations are good at exploring local regions, while editing may rescue immune responses stuck on unsatisfactory local optima [15]. In our implementation, it was assumed that the number of clones generated for all these n selected antibodies was given by equation:

$$Nc = \sum_{i=1}^{n} round\ (\frac{\beta \bullet N}{i})$$

Where Nc is the total number of clones generated for each of the Ag's, β is a multiplying factor, N is the total number of Ab's, and round(.) is the operator that rounds its argument toward the closest integer. The affinity measure takes into account the Hamming distance (D) between an antigen Ag and an antibody Ab.

$$D = \sum_{i=1}^{L} \delta_i, \quad where \quad \delta_i = \begin{cases} 1, & if \quad Ab_{ki} \neq Ag_{ji} \\ 0, & otherwise \end{cases}$$

IV. Improved Q-Learning Algorithm

Reinforcement learning is to find the optimal policy (or path) given an initial state and a goal state in some appropriate state space (e.g., game problems). It is based on observing the environment changes from one state to another as a result of actions taken by an agent. In this process, the value function of the state (or the state-action pair) is calculated so that the optimal policy can be found. In a typical Q learning problem, we can describe the policy space as set $R = \prod_{i=1}^{n} A(s_i)$, where s_i (i =1, 2,…, n) are the possible environment states, $A(s_i)$ are the sets of actions allowed in each of the state s_i, respectively. A problem policy will be represented as an n-tuple $(a_{1,} a_2, ..., a_n)$, with $a_i \in A(s_i)$. Then the change of any element in an n-tuple $(a_{1,} a_2, ..., a_n)$ corresponds to the transition from one solution to another in the policy space R. For example, assume that the *i*th element is changed to be b_i, $b_i \in A(s_i)$.Then the new corresponding n-tuple becomes $(a_{1,} a_2, ..., a_{i-1,} b_i, a_{i+1}, ..., a_n)$.

The policy R is a function that maps the system state to an action to be taken by the agent. Through interaction with the environment, the agent learns both an action-value function of a policy denoted as $Q(s,a)$, which maps state s and action a to a "long-term" reward as follows [16]:

$$Q^R(s,a) = \sum_{k=0}^{\infty} \gamma^t r^{t+1} \qquad 0 < \gamma \leq 1$$

And the value function V (R) which is used to evaluate the policy is

$$V(R) = \sum_{k=1}^{n} Q^R(s_k, a_k)$$

It is the sum of all the Q-values obtain for given policy.

Immunity-based techniques are gaining popularity in wide area of applications, and emerging as a new branch of Artificial intelligence. They have been applied to search for the optimum in the solution space of combinatorial optimization problems (like, e.g., the traveling salesman problem). Since the policy space is defined in the discrete Markov decision process environment, it becomes possible to transplant some immune mechanisms into the Q-learning approach.

In the proposed algorithm, the algorithm does not obey the policy learned so far like traditional Q-learning algorithm, but also attempts to explore by increasing chances of selecting actions which may connect to global optimal policy based on the hypermutation mechanism of immune system. In some sense, we make the algorithm move towards another direction instead of its original one for greedy search. The pseudo is presented as follows:

```
Initiate arbitrarily all Q(s,a) values;
For i=1: $N_e$
  Rand(s);
Do while ($s \neq S$)
fit=0;
$a = a_R$
For i=1: $N_c$
$a_g = rand(A(s));$
    If $|(Q(s,a_g) - Q(s,a_R))| > fit$
        $a = a_g$
        fit=$|(Q(s,a_g) - Q(s,a_R)|$
end
end
r=r(s, a);
$s' =$ δ(a,t);
$Q(s,a) \equiv Q(s,a) + \alpha(r(s,a) + \gamma \max_{a'} Q(\delta(s',a') - Q(s,a))$
$s = s'$;
end
    update $N_c$ ;
end
```

Rand: Choose a random (initial) state;

S: one of the goal states;

N_e : The desired numbers of episodes to be investigated.

N_c : hypermutation number

In our Algorithm, we regard that the policy R_1 is superior to the policy R_2 if $V(R_1) > V(R_2)$ according to the reward maximization rule. We compare the value of a policy to the mutation state in immune progress and it will be used in the algorithm to decide the probability of next action in combination with hypermutation mechanism. Through this implementation, the Q-learning algorithm has been converted into the combinatorial optimization problem in the policy space and the introduction of the immune mechanism is rational.

Comparing the description of one step Q-learning algorithm, there are only two additional steps in the proposed algorithm: the mutation selection of action and the fitness values of hypermutation. Since we replaces the difference of the value functions between the policies with $Q(s,a_g) - Q(s,a_R)$ in order to reduce the amount of computation, there is no substantial increase of complexity amongst them as the two steps require constant time to evaluate. In addition, the proposed Q-learning algorithm eliminates the disadvantage of a common ε-greedy approach and the exploration will easily converge to global set with the hypermutation scheme.

V. EXPERIMENT

In our experiment, we use a simple maze problem taken from [17] to test the proposed algorithm and compare with the results of basic Q-learning algorithm equipped with the ε-greedy strategy and of the Boltzmann exploration.

In this 22*17 puzzle problem (Fig. 1) above, there are six goal states (marked by G's) and the bold lines represent the walls. Any state can be a starting state. The task is to construct an optimal policy for moving from any state to a goal state.

The agent is located in one of the squares of the grid (the current state) and has the choice among eight possible actions. The four one-step actions are: NORTH, SOUTH, WEST, and EAST. As their effect the agent moves one square forward in the corresponding direction. Once encountering the wall, the agent has no possibility of moving in the given direction—the resulting state after such action is the original one. In the case of the four compound operators: NORTH-TO-WALL,SOUTH-TO-WALL,WEST-TO-WAL L, and EAST-TO- WALL, the agent moves in the stated direction until it reaches a wall. These operators have different costs. The single-step operators cost 1 unit; the compound operators cost 3 units; and the wall-following operators cost 5 units. The robot receives a reward of 100 units when it reaches the goal. The goal is to find the policy that maximizes the total reward received by the robot and the optimal path to the goal states.

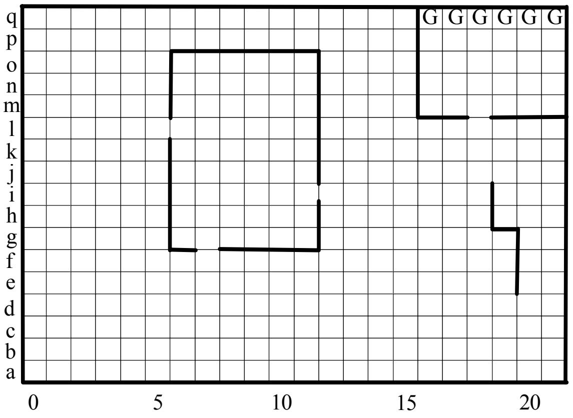

Fig. 1. A small maze problem.

In order to evaluate its performance, the number of solution sets learned by proposed Q-learning algorithm was compared to the results of other searching methods. The ability of the solution found, in terms of differences between the paths proposed by the policy learned by the algorithm, and the optimal paths for this problem, found by other algorithms as a function of the episode.

The comparison has been done for the Boltzmann exploration algorithm, the basic Q-learning algorithm (i.e., the 0-greedy strategy), and the proposed Q-learning algorithm.

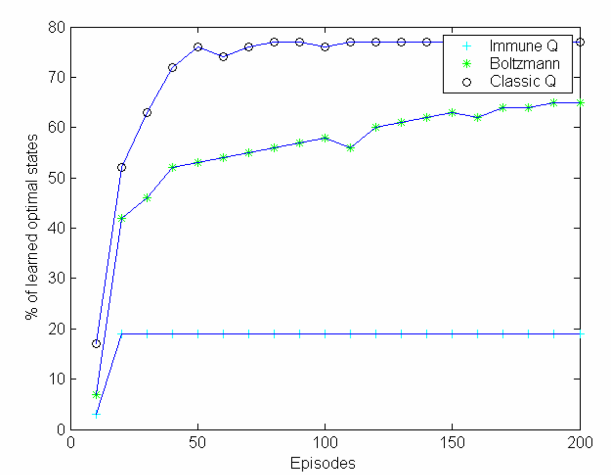

Fig. 2. Optimal solution learned by algorithms.

The graph in Fig. 2 shows the results found by these algorithms. The quality is measured by taking the ratio between the number of steps on the optimal path from the initial state to a goal state, and the number of steps on the path proposed by the policy learned so far. We can see clearly that this ratio is higher for the proposed Q-learning algorithm than for the Boltzmann exploration and much better than Basic Q learning. This result may probably caused by the lack of searching direction on convergence. For the proposed Q-learning algorithm, however, the chances of converging to nonoptimal policies are much smaller due to initial exploration and to its subsequent greedy search based on hypermutation scheme.

TABLE 1 ITERATIONS VS. NC

Nc	ITERATIONS
1	80
2	66
3	62
4	55

The immune mechanism presented here adopts the natural immune response mechanisms of mutation and proliferation. Implementation of this algorithm on the problems of Q-learning shows that some parameters might determine the performance of the applied method such as convergence speed, the computational complexity and its capability to fulfill an optimal search. The table.1 presents the experiment results on the different value of N_c .It clearly shows the trade-off between average iterations and N_c and indicates that the convergence speed increases when N_c increases. This result meets our expectations since larger number of N_c refers to larger number of available clones which accelerates the speed of convergence in terms of program iterations. The computational time per iteration also increases linearly with N_c

VI. CONCLUSIONS

In this paper, we solved the optimal policy learning in Q-learning algorithm as a combinatorial optimization problem by introducing some immune mechanisms to strike a balance between exploration and exploitation in the Q-Learning progress. Theoretical results show that the improved Immune Q-learning is superior to the standard Q-learning algorithm with respect to learning results, such as convergence speed and greed for global optima. Extensive experiments indicate that the proposed Q-learning shows superior results when compared to classic Q-learning algorithm and also exhibits improved performance in comparison with other random learning algorithms such as the Boltzmann exploration.

Further work in the area can be dedicated to analyze the convergence speed of the proposed Q-learning algorithm in relation to the hypermutation parameters. Also the boundary on use of hypermutation factor to guarantee the global optima needs to be further analyzed.

.

REFERENCES

[1] M. Z. Guo, B. Chen, X. L. Wang, and J. R. Hong, “A summary on reinforcement learning” (in Chinese), Computer Science, vol. 25, no. 3, pp. 13–15, 1998.W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] R. S. Sutton, “Learning to predict by the method of temporal difference” Mach. Learn., vol. 3, no. 1, pp. 9–44, 1988.

[3] C. J. C. H.Watkins, “Learning from delayed rewards,” Ph.D dissertation, Psychol. Dept., Cambridge Univ., Cambridge, U.K., 1989..

[4] C. J. C. H. Watkins and P. Dayan, “Q-learning,” Mach. Learn., vol. 8, no. 3, pp. 279–292, 1992.

[5] L. Yang, J. R. Hong, and T. Y. Huang, “Combining the methods of temporal differences with neural network for real-time modeling and prediction of time series” (in Chinese), Chinese J. Comput., vol. 19, no. 9, pp. 695–700, 1996

[6] M. Asada, E. Uchibe, and K. Hosoda, “Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development,” . Intell., vol. 110, pp. 275–292, 1999.

[7] Yamaguchi, T., Miura, M., and Yachida, “M. Multi-agent Reinforcement Leaming with Adaptive Mimetism”, 5th IEEE Int.

Conf. on Emerging Technologies and Factory Automation (ETFA-96), at Kauai, Hawaii.

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

[9] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," J. AI Res., vol. 4, pp. 237–285, 1996.

[10] N. Metropolis, A.W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of calculations by fast computing machines," J. Chem. Phys., vol. 21, pp. 1087–1092, 1953.

[11] Charles Janeway, "Immmunobiology the immune system in Health and Disease", Garland Pub, 2004.

[12] De Castro,L.N. &Von Zuben, F.J.(1999), " Artificaial immune system : Part I-Basic theory And Applications", Technical Report – RT DCA 01/99,p.90.0.

[13] Leandro N. de Castro, Femando J. Von Zuben, "An Evolutionary Immune Network for Data Clustering",In Proc.of the IEEE SBRN,pp.84-89,Rio De Janeiro, 22-25 November, 2000.

[14] Leandro N.de Castro, Jonathan Timmis "Artificial Immune System: A New Computational Intelligence Approach",ISBN 1-85233-594-7 Springer-Verlag London Berlin Heidelberg.

[15] ZhengQiao Ji, Q.M.Jonathan.Wu, "Application of Artificial Immune Algorithm in Multiple Sensor System", Systems, Man and Cybernetics, 2006 IEEE International Conference on, Volume 3, 6-9 Oct. 2006 Page(s):6 pp. vol.3.

[16] Maozu Guo; Yang Liu; Malec, J, " A new Q-learning algorithm based on the metropolis criterion ", Systems, Man and Cybernetics, Part B, IEEE Transactions on Volume 34, Issue 5, Oct. 2004 Page(s): 2140 - 2143

[17] T. G. Dietterich and N. S. Flann, "Explanation-based learning and reinforcement learning:Aunified view," Mach. Learn., vol. 28, pp. 169–210,1997.