Predatory Simulation Using Q-GA: Evolutional Learning Taking Over Experiences

Yuko Ishiwaka
Institute of Hakodate
National College of Technology,
14-1 Tokuracho,Hakodate City,
Hokkaido, Japan
ishiwaka@hakdoate-ct.ac.jp

Takamasa Sato
Future University Hakodate
116-2 Kameda Nakano,
Hakodate city,
Hokkaido, Japan

Masashi Furukawa
Institute of Asahikawa
National College of Technology,
2-1-6, Syunkoudai 2jyou,
Asahikawa city, Hokkaido, Japan
mack@asahikawa-nct.ac.jp

Abstract—Recently many researchers in artificial life are interested in the interrelationship between evolution and learning. The purpose of this research is for individuals to form the prey-predator relationship using Q-GA, which is one of evolutionary reinforcement learning methods. We simulate on the grid world that baits are generated at fixed intervals. In the simulated world there is just one kind of species at first period, then the specie learning how to obtain the baits and leave their progeny. According to their learning, we can show the process to evolve and specialize from the single specie to the multi species and to form the prey-predator relationship.

I. Introduction

Q-learning [1] which is a kind of reinforcement learning learns by trial and errors. An agent is in a lot of state-action pairs and in each state an agent can choose several actions per state in Q-learning. Each value of state-action is called Q-value and the set of Q-value called Q-table. Remembering rewards that it has collected, Q-values are updated for each state-action pair. Thus each agent learns the best action from any given state. On the other hand genetic algorithm (GA) is one method to adapt the environment and figure out approximate solution by iterating crossover, selection and mutation. Individuals for GA are implanted with genes that constructed Q-tables, and therefore each individual is able to learn the behavior with Q-learning algorithms as an agent. The individual implanted Q-learning structure with genes is taken over the learned Q-values from the parents by GA, i.e. crossover, selection and mutation. For a crossover, half genes are behanded down in descendants per parent; therefore the gene structure should be extendable. We employ predation for the application in 2-dimentional grid world. The gene parameters of each agent are followings; field of vision, movement speed, duration of life and bulk which have the relationship as trade-off, and Q-tables which is pared by state-action values, and they are needed for leaning the predatism. As simulation results, many new race are evaluated and the individual whose gene structure is suitable to Q-leaning survived. In multi-agent system, our proposed new learning algorithm Q-GA is able to take over the experiences, and it gives suggestions as parametric setting in Q-learning for each problem.

II. Evolutionary Reinforcement Learning

Some kinds of evolutionary computation (EC) are represented as algorithms which imitate the evolution and adaptation to environment of nature. In real nature, it is assumed that the learning interactions between individuals are quite important factor for evolution of species. This is called the Baldwin effect [2][3].This effect includes two transitions,

First stage: plastic individuals can develop new behaviors and can leave more offsprings,

Second stage: the characteristics which once must be acquired by each individual through learning come to be produced in the normal course of development without learning, i.e. genetic assimilation.

There are some related works about ERL. Ackley[4], Suzuki[5] proposed ERL algorithms. We proposed here new ERL algorithm Q-GA which is combined genetic algorithm with Q-learning.

III. Q-GA Algorithms

We proposed new ERL algorithms, and we call it Q-GA. Each individual is generated like genetic algorithms and it has Q-table in gene structure. By iterating crossover and mutation, new generation is generated and Q-table which one of its parents learned is also handed down. The size of Q-tables depends on the somatic characteristics, i.e. their own sight. In following subsections, the details of Q-GA will be explained.

A. Setting of Individuals

Each individual has gene as parameters which define physical description as followings;

- **Sight**: the range of observable space. Each individual can observe the square range of sight x 2 + 1 centered as an individual. The sight of an individual is shown in **Fig. 1**.

- **Speed:** the time interval from taking one action until taking next action. If this parameter set smaller,

then the individual can move faster.

- **Size and Color:** elements of apparent somatic characteristics. They can be available for the classification of race of artificial life.

- **Life:** When the individual takes one action, the life is decremented. If the life becomes zero, the individual deceases and is removed.

- **Food:** The characteristics of food for each individual. Each individual can not eat all foods but it depends on the characteristics of foods. The characteristics of food are represented by size and color.

Speed has a trade-off relationship with Sight. The trade-off is satisfied with the next equation,

$$speed + sight < SMAX, \qquad (1)$$

, where SMAX is constant number. Each individual has another parameter Energy which represents the reciprocal number of emptiness. Energy is not included in gene, i.e. the parameter Energy is not inherited. If the individual prey the other, the value Energy is added preyed individual Energy to preying individual.

B. Part of genetic algorithms in Q-GA

The gene structure is shown in **Fig.2** and each parameter is already explained above. The total size of coding gene is 56 bit.

The rule of prey depends on the characteristics of foods, i.e. color and size. Each individual has their own food characteristics (see **Fig.2**). When the individual and the other individual share the one same cell, if the physical description of the other individual corresponds to the characteristics included by gene, the individual figures out food and prey it. The crossover also depends on the size and color. When individuals get into touch with the other individual whose color and size are same as themselves, then they cross over to each other and generate new individuals. The new generated gene takes over their parent gene which is determined by given rule for crossover which is explained later.

Information for making a decision of individual recognition is only physical description, i.e. color and size. The gene information of their own appearance is compare with the appearance of other individual, and they determine the foods or homogeneity. If the energy of the individual is enough, the individual puts crossover ahead of prey, and vice versa.

Mutation happens to all individuals with the constant probability. Any 1 bit of gene is turned around.

C. Part of Q-Learning in Q-GA

We employ Q-learning, which is reinforcement learning. Learning architecture is constructed in each individual. Updating Q-values in learning agents is defined in the following equation:

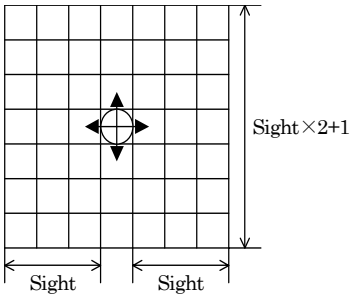

Fig.1 The sight and actions of Individuals

$$Q_i(s_i(t), o_i(t)) \leftarrow (1-\alpha)Q_i(s_i(t), o_i(t)) + \alpha\left[r + \gamma \max_{o_i' \in O_i} Q_i(s_i(t+1), o_i(t+1)) \right] \qquad (2)$$

,where s(t) is the state at time t, o(t) is the taken action at time t, α is the step-size parameter, γ is the discount-rate parameter and r is the reward from the environment.

The states of each individual are given as relative coordinate and species of which the nearest other individual exists in its sight. All individuals take four kinds of actions, i.e. up, down, left, right, and one by one. They can move only 1 cell per step. The following equation shows the state-action in Q-learning,

$$s_i \in \{X_i, Y_i, Geneus_i\}$$
$$a_i \in \{up, down, left, right\} \qquad (3)$$

, where s_i means the state set of individual *i*, a_i means the taken action set of individual *i*, X_i, Y_i means relative coordinate between the individuals and the nearest individual, and $Geneus_i$ shows the species of the nearest individual from it.

The reward *r* is obtained when the individual prey the other and cross over.

The one step is determined by speed of gene information. One action of each individual is corresponded to 1 step in Q-learning.

D. Q-GA algorithm

Q-GA algorithm is shown in Fig. 3.

Population is generated. Each gene of individuals in the population has the parameter of determined rule which has already mentioned at previous section. The parameter of Speed induces the new parameter t whose mean is latency time. Each individual takes an action according to the parameter time t equals to zero. The purpose of each individual is to find the same race in order to cross over or food in order to get the Energy. The optimal route to the other individuals will be learned by Q-learning. For the state of Q-learning each individual has the position and the race in the sight, and then it is possible for the individual to learn the optimal route. Each individual is required to adapt to not only environment but also the world for Q-learning. This means that unsuitable individuals to Q-learning are not selected for

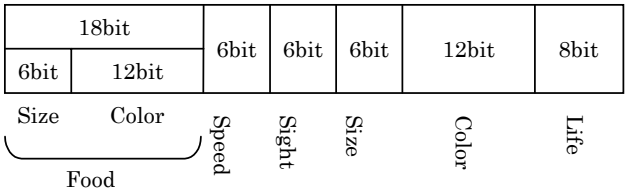

Fig.2 Genetic Coding

next generation. They will become extinct, and then the population will develop into suitable to Q-learning. Of course, it is not real world, but by this simulated world it gives suggestions as parametric setting in Q-learning for each problem.

Crossover rule in Q-GA is quite similar to genetic algorithms. One-point crossover is applied here. The crossover point is selected randomly, and the parent's genes are exchanged at the point. The parameter Energy of both parents reduces to the half of their Energy in that time. A generated individual obtains the rest Energy from both parents, i.e. parents share their Energy half-and-half for their child. Q-table is also taken over to their child. The generated individual takes over Q-table from one of its parents which is selected randomly.

There are two kinds of rewards in Q-GA. One of them is given when they cross over, and the other is given when they prey. Let the total reward r, then r is described as following equation,

$$r = r_{co} + r_f \tag{4}$$

, where r_{co} and r_f mean the rewards when an individual crosses over, and when an individual catch foods respectively. The crossover reward r_{co} is defined as followings,

$$r_{co} = \begin{cases} -1 & \text{: hungry} \\ 1 & \text{: the others.} \end{cases} \tag{5}$$

After taking action, the latency time t is set again. The termination conditions are one of followings, i.e. preyed by other individuals, Life is exhausted, and Energy is zero. Until terminal state, any individuals repeat taking actions.

IV. Environmental Setting

The environment of simulation world is two dimensional torus world whose size is $m \times m$. There is only one individual in one cell simultaneously. The simulation world is described in Fig.4.

For the initial state, a certain number of foods and initial individuals are assigned to the environment. Foods cannot move, but for each cell, they are generated with a constant probability per step.

The genes of initial individuals involve the size and color of their foods' characteristics as predation.

Q-GA Algorithm

Create Initial Population

For Each Agent

  Initialize Q, s from described in Genome

  Wait time $t \leftarrow T$ from described in Genome

Repeat Forever

  For Each Agent

    If t != 0 Then

      $t = t - 1$

    Else

      Choose action a using policy derived from Q

      Take action a

      Observe s' and r

      $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$

      If the same genus agent is in s' then Crossover

      $s' \leftarrow s$

      $t \leftarrow T$ from described in Genome

Fig.3 Q-GA Algorithm

V. Experiments

A. Parameters

Q-GA algorithm is applied to evolutional simulation whose individual learns. The size of experimental space is 50×50, the number of initial foods 500, and the number of initial individuals is 100. For Q-learning parameters, α is 0.1, γ is 0.8, and initial values of Q-value are 2.0. The event probability of generating new foods for each cell is 0.0005. For crossover the mutation is 0.0001. The parameters of initial individuals are shown in **Table 1**.

Table 1. The parameter of initial gene

Energy	20
Speed	30
Sight	3
Size	10
Food	Init Food
Life	127

B. Experimental results of total number of individuals

Fig.5 shows a result of population of individuals until 1 million steps. All individuals are inherited Q-values. The blue line shows the population of the total number of individuals, the fuchsia line shows the total number of foods, and the raspberry line shows the total number of preyed individuals.

From the point of the number of individuals, at first the number of foods is increased indeed, but after 200 thousands step the number of individuals becomes stable. The whole sum of the preyed individuals increases linearly approximately over steps after 200 thousands step. From this result we suppose that the predators are generated constant.

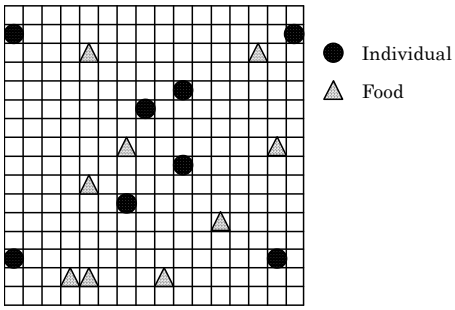

Fig.4 A-Life World

Fig.6 shows the scaled up **Fig. 5** until 200 thousands step. As a result of this figure, it is said that Lotka-Volterra model [6] for two species, which is the famous model of predator-prey interactions, comes out. The general nature of the dynamics resulting from the interaction between two species can be deduced by calculating the zero-growth isoclines for the two species. In this case we can find the population dynamics like this model. Because it is obvious that prey and predator populations are in equilibrium when their predator rates of change are decrease because births are equal deaths under this condition.

For the comparison the total numbers of individuals are shown in **Fig.7** when Q-values do not be taken over, i.e. normal GA. After 400 thousands step the individual number becomes stable. This suggests that the Baldwin effect is functional.

C. Experimental results of spices

Fig.9 shows the number of generated spices and the each number of the individuals after 1 million steps. From beginning to 30 thousands step most species of individuals are same as initial species. Around 100 thousands steps differentiation of species has originated. Until about 300 thousands steps 60% of individuals are one or two races, but around 400 thousands steps six kinds of species account for 10% of the population respectively. In this step, it is assumed that each species has co-dependent relationships with others. Around 600 thousands steps, many kinds of races are distributed. This means that there are no winner takes all, so to speak. Over 800 thousands steps, a small minority of species share about 40% of the total individuals again.

D. Experimental results of characteristics

Every 200 thousands steps, the rate and the characteristics of the most prosperous species are shown in **Fig.9**. As the number of steps increases, the parameter Life and Sight evolve and Speed atrophies. The reasons why Life evolved are 1. longer life can prey more foods, 2. the reward of each individual can be maximized, and 3. longer life can produce numbers of offspring. The descent of Speed cause longer life. Because in this world the elapse of Life of each individual, i.e. the process of aging, depends only on the number of taking actions. This means that faster species of speed has shorter life. The reason why sight evolved is the trade-off relationship between Speed and Sight. Speed, i.e. Life is more important

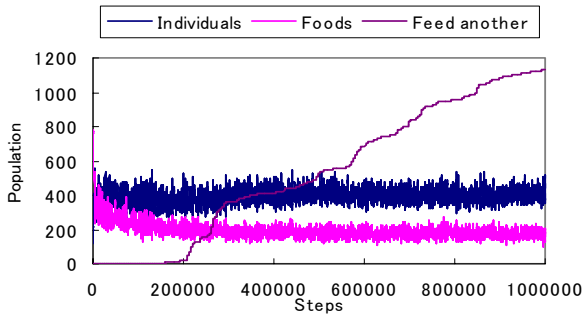

Fig.5 Populations of Individuals
(Inheritance Q value, 1000000steps)

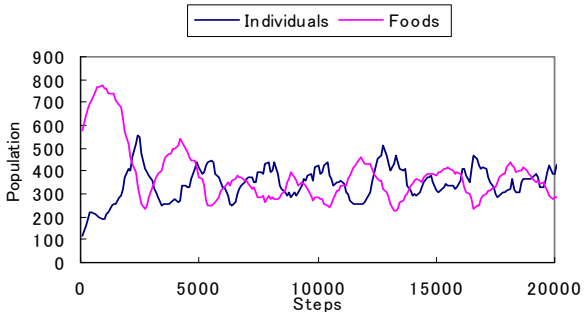

Fig.6 Populations of Individuals
(Inheritance Q value, 20000steps)

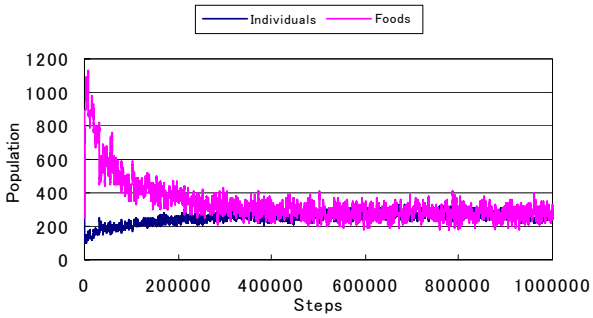

Fig.7 Population of individuals (without taking over Q-tables.)

factor than the other parameters for surviving.

Fig.10 shows the distribution of Speed, Sight and Life every generation by the total number of individuals. In this figure, it is obvious that the process of transition from the evolution of Life and Speed. Concerning Speed, there is no race to evolve to faster Speed race. Mentioned about Sight, around 200 thousands steps most individuals have the little bit smaller sight than initial individuals. From 300 thousand to 700 thousands many species who have various kinds of Sight are born. After 800 thousands steps the race who has pantoscopic Sight has been born and they increase the size of the race until around 1 million steps. For Life, before 100 thousands steps, most individuals already have longer life.

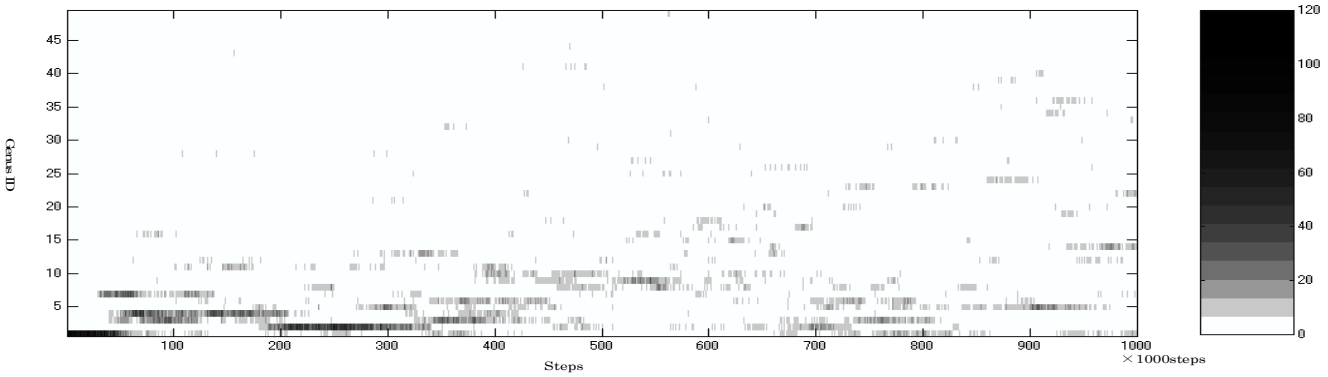

Fig.8 Count of Generated Genus

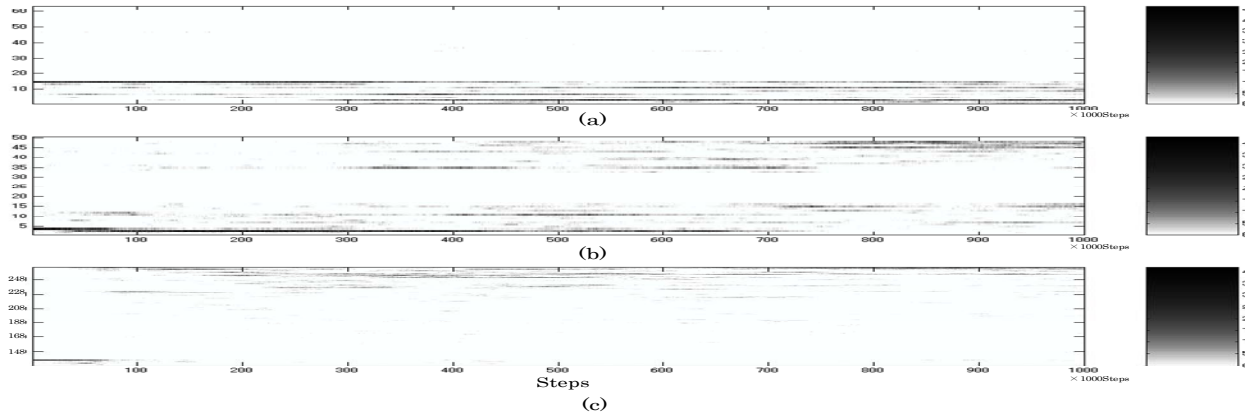

Fig.10 Characteristics of Generation Categorized by (a) the speed, (b) the sight and (c) the length of life

VI. Summary

In this paper we proposed the new evolutionary reinforcement earning algorithm, Q-GA and applied to the evolutional simulation. The purpose of the simulation is the emergence of predatory relationship and adaptation to environment. In the beginning of the simulation there is only one kind of species which has many kinds of parameters for adaptation abilities in its own gene. As a result, various kinds of species are generated from single race, and their abilities change into many types. From predator-prey interactions individuals adapt to the environment and they keep the stable number of population like Lotka-Volterra Model. The emergence and differentiation of species occurred, however the hierarchical predator-prey interactions are not emergent. It is assumed that the small number of factors of characteristics cause. For the future work the number of factors which shows the characteristics of individual abilities increase, or expend to continuous world.

REFERENCES

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

[2]Baldwin, J.M.: *A new factor in evolution. American Naturalist*, 30, 441-451, 536-553, 1896, reprinted in Belew and Mitchell, 1996.

[3]Godfrey-Smith, P.: *Between Baldwin skepticism and Baldwin boosterism*, in Weber, B. & Depew, D. (eds), Learning and evolution. The Baldwin effect reconsidered, Cambridge, MA, The MIT Press. ISBN 0-262-23229-4,2003

[4]Ackley, D.E. and Littman, M.L. *Interactions between learning and evolution*. In C.G. Langton, J.D. Farmer, S. Rasmussen, and C.E. Taylor (eds.) Proceedings of the Second Conference on Artificial Life. Addison-Wesley: Reading, MA. ,1991.

[5] Suzuki, R., and Arita, *T. Interaction between evolution and learning in a population of globally or locally interacting agents ,* In Proceedings of the Seventh International Conference on Neural Information Processing, pp. 738-743 ,2000.

[6]Begon, M., J. L. Harper, and C. R. Townsend. Ecology: *Individuals, Populations, and Communities,* 3rd edition. Blackwell Science Ltd. Cambridge, MA, 1996.

Fig.9 The number of generated spices and the each number of the individuals.