5th INTERNATIONAL CONFERENCE
TRANSPORT SYSTEMS TELEMATICS TST'05

ZESZYTY NAUKOWE
POLITECHNIKI ŚLĄSKIEJ 2005
TRANSPORT z.59, nr kol. 1691

reinforcement learning, traffic control,
multicriteria decisions making

Włodzimierz FILIPOWICZ [1]

REINFORCEMENT LEARNING IN VESSELS TRAFFIC ENGINEERING

Reinforcement learning is applied when agent or operator interacts with or tries to control more or less random environment. In reinforcement learning there are: a policy, a state or action value function, and a model of the environment involved. A policy specifies set of actions applied for controlling the traffic flow.

UCZENIE ZE WZMOCNIENIEM W INŻYNIERII RUCHU MORSKIEGO

Uczenie ze wzmocnieniem ma zastosowanie w przypadku, kiedy agent albo operator steruje mniej lub bardziej losowym środowiskiem. W uczeniu ze wzmocnieniem występują: polityka, funkcja definiująca nagrody, funkcja oceny wartości stanów lub akcji oraz model sterowanego procesu. Polityka określa zbiór możliwych do podjęcia, w ramach sterowania ruchem statków, akcji.

1. INTRODUCTION

In reinforcement learning there are: a policy, a reward function, a value function, and a model of the environment involved. A policy defines the agent's way of behaving at a given state of the environment. It specifies set of actions applied for controlling the environment. The agent can receive reward or punishment – both called reinforcement. The reinforcement assigns values to states of the environment, indicating the quality of the state. An agent's objective is to maximize the total reward it receives in the long term [3].

A reinforcement function indicates immediate rewards and usually is supplemented by a state value function. The last specifies how good is to take particular action being in given state. The value of a state is the total amount of reward an agent can expect to accumulate over the future states.

Important part of reinforcement learning systems is a model of the environment. This is something that imitates the behavior of the environment. Given a state and action, the model might be used to predict the next state and next reward. The incorporation of models and planning into reinforcement learning systems makes the whole approach more flexible.

A model of the environment enables to simulate of a set of states. Transition to a new state is a result of an agent action. In Markov model important is that the probability of performing some action will result in transitioning to some other state, and the reinforcement

[1] Department of Informatics, Gdynia Maritime University, 81-225 Gdynia, Morska 81/83, wlofil@am.gdynia.pl

the agent might receive from that transition, do not depend on any previous actions the agent has made, or the previous states. In Markov model history doesn't matter most important is the state the environment is at any moment, and the action the agent is performing.

Reinforcement learning can also be seen as a family of algorithms, which enable to generate optimal solutions to Markov models. Reinforcement learning can be applied to continual process-control tasks as well as to episodic processes. Such tasks arise while dealing with traffic control. VTS operator interacts with the traffic within given area, he probably also face a problem of suggesting best route for particular vessel. Within scope of his activity he follows up certain policy, which objective is aimed at improvement of safety standards within the area. Maximizing overall rewards expressed as best safety conditions is his final aim. To make his decisions he is supposed to use an environment model. The fuzzy model, which enables estimation of a proposed set of parameters will be discussed in the presentation. Decision-making is multicriteria one so some hints regarding the approach will be also presented.

2. REINFORCEMENT LEARNING CONCEPT

The agent and environment interact at each moment of a sequence of discrete time $t \in \{t_1, t_2, t_n, \}$. At each time step, the agent receives some representation of the environment's state, $s_t \in S$, where S is the set of possible states, and on that basis selects an action, $a_t \in A(s_t)$, where $A(s_t)$ is the set of actions available in state s_t. One step later, as a consequence of its action, the agent receives a numerical reward, $r_{t+1} \in R$ and the environment transits to a new state, $s_{t+1} = s' = \delta(s_t, a_t)$. The new state depends on the current state s_t and taken decision a_t. In deterministic case the new state is always known since it merely determined by the action. Many real life cases contradict this simple assumption.

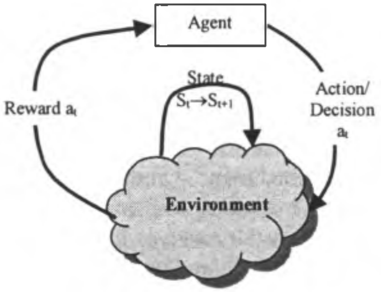

Fig.1. General concept of the reinforcement learning

At each time step, the agent implements a mapping from state representations to probabilities of selecting each possible action. This mapping is called the policy and denoted π_t, where $\pi_t(s, a)$ is the probability that $a_t = a$ if $s_t = s$ [3]. Reinforcement learning methods specify how the agent changes its policy as a result of its experience. The agent's goal is to maximize the total amount of reward it receives over the long run. Profits can be estimated using environment's model. Adequate model is of great value in many real life applications where the discussed methodology could be applied. Basic dilemma of reinforcement learning of exploitation and exploration can be differently perceived when good model is available.

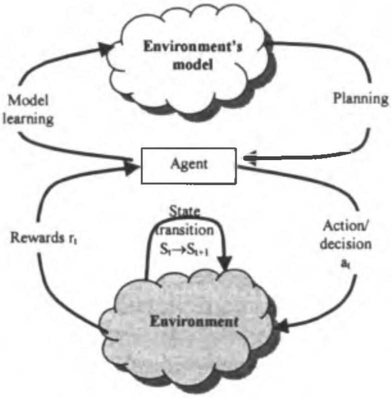

Fig.2. The reinforcement learning scheme with the environment model included

In the reinforcement learning framework, the agent makes its decisions on the basis of the environment's state. A state characteristic that embraces all relevant information is said to be Markov, or the environment that enables such characteristic is said to posses the Markov property. For example, the current configuration of all pieces on the chessboard would serve as a Markov state because it includes everything important about the player situation and its future success or failure. Information about the previous actions, which resulted in a particular configuration, is lost but this does not matter. If an environment has the Markov property, then its one-step dynamics enables to predict the next state and expected future rewards starting from the current state and taking one of possible action.

2.1. VALUE FUNCTIONS, BELLMAN EQUATION AND OPTIMAL POLICY

Reinforcement learning algorithms are based on value functions, functions that estimate how good it is for the agent that environment is in a given state. It can also specify quality of performance of a given action in a given environment state. Quality is expressed in terms of expected future profits. The value of a state *s* under a policy π, denoted $V^{\pi}(s)$, is the expected reinforcement E_{π} when following policy π starting in state *s*. The state value $V^{\pi}(s)$ is used to determine $Q^{\pi}(s, a)$, which evaluates taking action *a* in state *s* under policy π. The expected return is given by formula (1)

$$Q^{\pi}(s,a) = r(s,a) + \gamma V^{\pi}(\delta(s,a)) \tag{1}$$

Q^{π} is called the action-value function for policy π, it delivers direct assessment of the agent activity. It denotes the reward for starting in state *s*, taking action *a* and discounting ($\gamma \leq 1$) new state value. A fundamental property of value functions is that they satisfy particular recursive relationships (2). The formula is known as Bellman equation for action-value function.

$$Q^{\pi}(s,a) = r(s,a) + \gamma\sum_{a'} \pi(\delta(s,a),a') * Q^{\pi}(\delta(s,a),a') \tag{2}$$

It bridges the value of given action and the values of actions available within successor states. Recursive Bellman equation means that being at the state s and taking action a next environment state is likely to be $\delta(s,a)$ and its immediate reaction, reinforcement is $r(s,a)$. For finite Markov Processes, one can define an optimal policy. A policy π is better than or equal to a policy π' if its expected result is greater than or equal to that of π' for all states. It means that: π ≥π'if and only if $V^{\pi}(s) \geq V^{\pi'}(s)$ for all $s \in S$. There is always one policy that is better than or equal to all other policies, this is an optimal policy. Bellman optimality equation must express the maximum of the expected result for any action from given state, formula (3) define the relationship.

$$\begin{aligned} V^{*}(s) &= \max_{a} Q^{*}(s,a) \\ &= \max_{a}(r(s,a) + \gamma V^{*}(\delta(s,a))) \\ Q^{*}(s,a) &= r(s,a) + \gamma \max_{a'} Q^{*}(\delta(s,a),a') \end{aligned} \tag{3}$$

3. TRAFFIC CONTROL

Vessels traffic is monitored by coastal station wherever it is possible. Monitoring aim is to check that everyone obeys imposed rules and traffic separation is not violated. All such measures available within Vessels Traffic Systems significantly contributed to reduction of risk of collision and improved environment safety standards. Further improvement will be possible when VTS operator is able to implement and follow adequate policy aimed at traffic allocation within the area. Avoiding local congestion one can reduce number of encounters and furthermore potential risk of collision. To introduce such measures a few assumptions are to be made. First there must be decision-making body available within VTS structure. Second databases are to be implemented and relevant decision problems to be formulated and solved. The last comes along with proposal of the set of assessment criteria and delivering necessary tools to decision maker. Finally legal aspects related to enforcement of the policy are to be overcome.

Ship's master usually carries out decision-making regarding passage through congested area. Such decision process is doomed to be impaired by lack of actual data regarding other traffic. VTS stations plus reporting system are adequate source of crafts movement data. They are supposed to recommend on itinerary as well as best time of passage.

3.1. ROUTE SELECTION AND ROUTES ASSIGNMENT

Considering vessels traffic there are three main problems to be discussed. First is a route selection (RS for short), the problem arises when master has to decide on itinerary during his sea passage. Second is a routes assignment (RA), the problem appears when dealing with a set of crafts and alternative routes exist in restricted area [2]. The last is a VTS supervisor problem and is related to best possible distribution of the entire passing traffic. VTS operator can be asked for advise on best possible passage for particular vessel, while not interacting with the rest of traffic. Such problem will be called route recommendation and abbreviated as RR. Ship's route can be treated as a sequence of legs joining turning areas. Whenever a route is to be selected or routes should be assigned one has to consider a wide variety of different parameters. For particular vessel and each route, she is assumed to take, time of passage and scheduled traffic are presumably known. Apart from these parameters there are forecasts regarding local traffic as well as rogues or crafts that do not obey imposed rules. To take proper choice, assumed a few options exist one has to compare a handful of parameters of different nature. There are crisp and fuzzy parameters values, forecast empirical sets and probability distribution functions to be dealt with.

Traffic is classified taking into account gross tonnage of a vessel and kind of cargo she has on board. For this reason safety factor has been introduced. Environmentally dangerous freight and huge tonnage increase the factor. The factor vary on an integer scale such that the higher the number the more serious the consequences of an accident. There was range from 1 to 10 suggested [1][2]. It is assumed that safety factor number is assigned to every ship.

Extended, compared to presented in [2], set of criteria embraces

- number of vessels directed to fair „quality" of the waterways. Quality of a waterway should be thought of as excellent, very good, good, fair etc. in other words one use linguistic terms to express the quality of a passage for particular class of vessel. Classes are assumed to be related to safety factors. Criterion is abbreviated as NVDF and remains applicable to RA problem
- total passage time (or maximum delay). It is crisp or fuzzy value, fuzziness is related to unforeseen deviation from the scheduled trajectory because of collision avoidance manoeuvres or pitching due to bad weather and seas. Criterion (TPT) is applicable to RS, RA and RR problems
- total number of encounters, the parameter is calculated based on simulation. Criterion (TNE) is applicable to RA and RR problems
- total number of crossing encounters. Since any ship presence within any area is described by membership function this parameter is to be defuzzified into crisp value. Criterion reflects whole group of crossings, for this reason included numbers specify range of courses difference (example notation TNCE 20, 40). Data is applicable to RA and RR problems
- number of crossing encounters of ships with safety factor greater than given value. Criterion (example of abbreviation TNE SF>5 20, 60) is applicable to RA and RR problems. Meaning of included numbers is the same as before
- number of encounters of ships with safety factor greater than given value which will occur in the area of special concern. Value is to be defuzzified. Criterion (example of abbreviation NE SF>5) is applicable to RA and RR problems
- number of encounters with local traffic forecast for consecutive nodes. This parameter will be in a form of sets of empirical data recorded by the local authority. Theoretical density distributions can be used instead. Criterion (FNE LT) is applicable to RS, RA and RR problems
- number of encounters with unscheduled traffic forecast for consecutive nodes. This parameter will be a set of empirical data recorded by the local authority. Criterion (FNE UT) remains applicable to RS, RA and RR problems

- maximum sum of safety factors of ships present, at the same time, within area A that is an example region of particular concern (such areas are sometimes referred to as Safety Zones). Criterion (example abbreviation SSF A) is mainly applicable to RA problem
- maximum number of ships present, at the same time, within area of particular concern (such areas are sometimes referred to as Safety Zones). Criterion (example abbreviation MNS A) is mainly applicable to RA problem
- weather, current and seas condition along each route. Criteria (WC, CC, SC) are applicable to RS problem. Criteria should be thought of in terms of linguistic expression as excellent, very good, good, fair, poor and very poor

3.2. ENVIRONMENT MODEL

To foresee encounter numbers a timetable of arrival at given points are to be constructed for each scheduled vessel. Timetable of passage, for each vessel, and for given area is a vector of so-called fuzzy slots, which are quads of values that define so called membership or presence in the region function. Latest arrival time in the area (A_L) and the earliest departure time from the area (D_E) of the particular vessel are reference values. From them membership function descends to earliest arrival (A_E) and latest departure time (D_L) respectively. Fuzziness can be associated with difference between earliest and latest moment primarily depends on sea condition and necessary deviation from the prescribed trajectory. To foresee what will take place within given area one has to consider all presence functions greater than zero within all possible time slots. Formally the membership function can be of the form presented by equation (4).

$$f_{S_k}(t) = \begin{cases} f^L(t) = \dfrac{t - A_E}{A_L - A_E} & \text{for all} \quad (A_E \le t \le A_L) \\ 1 & \text{for all} \quad (A_L \le t \le D_E) \\ f^R(t) = \dfrac{t - D_L}{D_E - D_L} & \text{for all} \quad (D_E \le t \le D_L) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Some parameters, for example total number of encounters, can be estimated based on simulation. Basic assumptions of the mathematical model embrace:

- there are ten classes of ships, each vessel is classified with respect of her tonnage and carried cargo
- system of route within the area is fixed, for some directions of flow there are alternative passages. Model is equipped with interface enabling definition of the routes scheme
- interface enabling input of initial positions and intended routes
- route consists of legs linking turning areas
- ships trajectory leads from one turning point to the subsequent one. Turning points are randomly distributed within turning areas
- movement along prescribed trajectory is double screened random Markovian process, no collision avoidance is carried out
- state “save to file” ability, for continuity reasons
- interface enabling ship domain(s) definition

Encounters should be detected when safe limit is violated and subsequently stored for further analysis. Two ships are registered being involved in close approach when it first occurs, their subsequent mutual positions are not considered unless category of encounter is changed. Categories list of encounters embrace: meeting, overtaking and crossing, which is further subdivided regarding angle of crossing.

3.3. STATE VALUE IN TRAFFIC CONTROL

Initial positions of all ships, their gross tonnage cargo intended courses and speeds are vital when considering navigational situation in a region. For this reason following definition of a state in traffic control is adopted: "ships positions at the beginning of a given interval of time, their intention regarding passage itinerary as well as tonnage, speed and kind of cargo characterize condition within the area and consequently environment state as defined in reinforcement learning".

TOPSIS method was adopted for upgrading state values hierarchy. TOPSIS stands for technique for order preference by similarity to an ideal solution. Was introduced as a multi-attribute decision making (MADM) method. Initially the approach was intended for crisp values then extended for fuzzy parameters as well as for sets of empirical data The extension covered fuzziness as well as empirical sets of data is presented in [4]. The method is based on a concept that the best alternative among the available alternative set is the closest to the best possible solution and the farthest from the worst possible solution at the same time. The final TOPSIS ranking is created by sorting, in descending order, the coefficient values assigned to each of the alternatives.

3.4. NUMERICAL EXAMPLE

Let us consider a set of data presented in table 2. There are six different states parameters included. Each state, that characteristic is included in a single row, refers to a possible option when deciding on routes assignment. First of the shaded column contains figures generated by procedure with implemented extended [4] TOPSIS method.

Consecutive columns titles conform mentioned notation. Last shaded ones mean:
ranking output generated by procedure, which implemented TOPSIS method
probability initial probability that particular assignment result in given set of parameters, the value will be subject to further changes in policy improvement
hierarchy final hierarchy

Table 1

Set of six possible assignments with equal probability of occurrence

	TNE	TNE SF>5	NoVDF	TPT	SSF X	MNS X	ranking	probability	hierarchy
1	10	5	1	150	10	medium	0,42/5	1	5
2	9	5	1	144	10	small	0,46/4	1	4
3	10	4	1	130	10	very small	0,53/1	1	1
4	11	3	2	160	7	negligible	0,50/2	1	2
5	12	4	2	170	5	very small	0,48/3	1	3
6	13	5	1	175	10	large	0,37/6	1	6

In case of following greedy policy operator will decide on third option with the highest-ranking value. Since each state reflects particular assignment of routes to each of the vessels any lack of concordance with the calculated scheme leads to different overall situation. While trying to enforce assignment number 3 recommendation disobeyed by one vessel resulted in creating condition similar to this stated in raw 6 (see table 2). For this reason greedy policy must be subject to further modification and improvement based on forecast and real data.

Table 2

Set of forecast and real data observed after selecting option 3 (table 1) for execution

	TNE	TNE SF>5	NVDF	TPT	SSF X	MNS X	ranking	hierarchy
3	10/13	4/5	1/1	130/175	10/10	very small/large	0.53/0,37	1/6

Since greedy policy can fail, following policy improvement scheme could be adopted:

- given state value V^* for arbitrary deterministic policy
- for particular state select adequate action, calculate $Q^*(s,a)$, which does conform with greedy policy
- based on available database estimate probability of occurrence of the particular assignment recalculate $Q^\pi(s,a)$
- if $Q^\pi(s,a) > Q^*(s,a)$ therefore choose new policy π

4. SUMMARY AND CONCLUSIONS

The aim of the VTS observer activity should be pointed at improvement of the formal safety parameters. The operator is assumed to advise vessels regarding passage details. Advises are supposed to be generated based on multi criteria optimization and multi attributes decision selection. TOPSIS method with many practical extensions well suits to upgrade hierarchy in considered problem. Theory of artificial intelligence, in particular reinforcement learning, is used to deal with similar to traffic control. The method is general one and delivers solid theoretical basis for formalization of the approach. Operator follows policy, which is assumed to be a greedy one. This means that best possible solution is to be executed. It is likely that from time to time such policy is to be revised. Adjustment of the policy is inherent feature of the reinforcement learning.

BIBLIOGRAPHY

[1] FILIPOWICZ Wł., „Minimizing Risk Probability For Vessels Traffic Control" – *Proceedings of the 3rd International Conference TST' 03 Transport Systems Telematics*, Katowice 2003, pp. 111-120

[2] FILIPOWICZ Wł., "Vessels Traffic Control Problems" – *Journal of Navigation* vol. 57/1, London 2004, pp. 15-24

[3] SUTTON R. S., BARTO A. G., "Reinforcement Learning: An Introduction", MIT Press, Cambridge, MA, 1998

[4] SZŁAPCZYŃSKA J., „Rozszerzona metoda Fuzzy TOPSIS – przypadki użycia w nawigacji morskiej", VI Międzynarodowe Sympozjum Nawigacyjne, Gdynia

Reviewer: Prof. Bernard Wiśniewski