The apprentice modeling through reinforcement with a temporal analysis using the Q-Learning algorithm

Marcus Vinicius C. Guelpeli
Universidade Federal dos Vales do Jequitinhonha e Mucuri – (UFVJM) –Diamantina, MG – Brasil
marcus.guelpeli@ufvjm.edu.br

Bruno Santos de Oliveira
Centro Universitário de Barra Mansa (UBM) – Barra Mansa – RJ – Brasil brn.santoos@gmail.com

Márcia Aurélia Pinto
Centro Universitário de Barra Mansa (UBM) – Barra Mansa – RJ – Brasil marcia.aurelia.pinto@hotmail.com

Ruana Carpanzano dos Santos
Centro Universitário de Barra Mansa (UBM)– Barra Mansa – RJ – Brasil ruana_carpanzano @hotmail.com

***Abstract*— This work aims to create the simulations by varying the alpha (a – Learning rate) and Gamma (y – Time reduction) values, such parameters found in the q-learning algorithm, which is possible to analyze the algorithms convergence, on what concerns the variations of these parameters.This work seeks to state that the parameters variations of Alpha and Gamma interfere on the convergence of Q-learning algorithm, thus, in the ITS learning.**

Keywords- ***Machine learning; Learning reinforcement; Q-learning; intelligence tutoring system***

I. INTRODUCTION

The intelligence tutoring system – (ITS) is an evolution of *Computer-Assisted Instruction* – (CAI) systems, improved by artificial intelligence - (AI) techniques. The ITS system interacts with the learner, where the cognitive modeling is progressive and constant. An ITS needs to model the learner in order to provide a personalized teaching. This modeling allows that teaching strategies may be associated to the cognitive state of each learner.

The machine learning – (ML) is a subfield of AI dedicated to the development of algorithms and techniques that allow the computers to learn. According to Bacardit[1] (2004) the main ML concern is how to build programs that automatically improve with its experience, in other words, intelligent systems that learn according its useful lifetime.

As fundamental objective of the ITS's is to provide an adapted instruction to the learner, it is necessary to perform a modeling of it in a computational way. The greatest difficulty found is to define the learner's model that is interacting with the ITS. According to Guelpeli (2000)[3], the greatest challenge, in this case, is to adapt the ITS with the ML, so the whole learner's cognitive modeling process can be done in a computational and automatic way. For this modeling, Guelpeli (2000)[3] suggests the use of the Learning through reinforcement – (LTR) technique, also using the Q-Learning algorithm (WATKINS, 1989) [6]

This work aims to create the simulations by varying the Alpha values (α – Learning rate) and Gamma (y – Time reduction), such parameters found in the Q-learning algorithm, which is possible to analyze the algorithm's convergence, on what concerns the variations of these parameters. This work seeks to state that the parameter's variations of Alpha and Gamma interfere on the convergence of the Q-learning algorithm, thus, in the ITS learning.

Along with the computational learning model were also performed several simulations with different Alpha and Gamma values, in order to analyze the model behavior by altering its parameters. To make such analysis possible, simulations were performed with ITS prototype in a proper environment set by Guelpeli (2000) [3]. Results demonstrate that the variations of these rates, in accordance with the chosen pedagogic policy, interfere in the algorithm convergence as well as in the ITS learning.

The present work is structured in sections. On section two it is found related works, on section three it is presented the Learning through reinforcement technique. Section four presents the system structure. On section five there is the methodology used in this experiment, results are showed and also discussed, and on section six conclusions are presented.

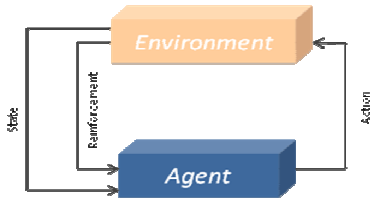
II. COORELATE WORKS

Guelpeli (2000) [3] once did an adaptation between ITS and ML, aiming to obtain dynamic learner modeling through interaction with ITS. After this adaptation performed by Guelpeli (2000) [3], optimization techniques, researches and developing started to be done in this area. It is possible to highlight the study about developing of intelligent systems such as anthology, using learning through reinforcement, which had as its primary goal to develop learning software to teach music history, besides suppressing some of the gaps in music teaching websites, by Boff (BOFF, et al 2004).[2]

III. LEARNING THROUGH REINFORCEMENT

The first picture represents LTR where the agent operates in an environment with several possible states, where on each one can be performed an action between

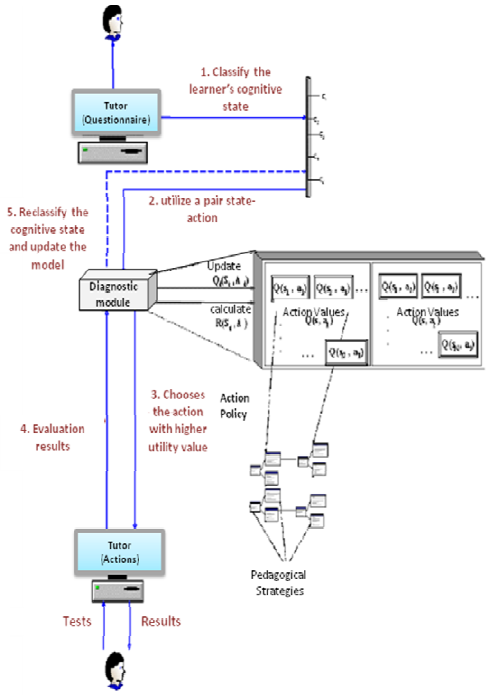
several possible actions receiving reinforcement on every action taken. Such reinforcement is the value of transition from one state to another one, making station-action pairs, with its proper reinforcement values, being generated along the process.



Picture 1: Learning through reinforcement

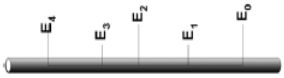
IV. *SYSTEM'S STRUCTURE*

The LTR technique introduced in the diagnostic module by Guelpeli (2000) [3] proposed in picture 2 model, defines how the teaching-learning process will work. It produces an established action policy and defines how the learner's performance is in accordance with the tutor's actions. The diagnostic model will be responsible of sending reinforcements, and producing for each pair of (station, action) a reinforcement value R(s, a) in accordance with the action policy. Thus, for the tutor to reclassify the learner's cognitive state, the utility values Q(s, a) of a pair (state, action) is calculated from reinforcements measured by the learner's cognitive state quality and are updated in the Q table of values.



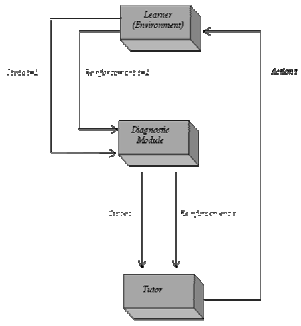
Picture 2: A system with LTR technique (GUELPELI 2000)

Through the learner's modeling showed in picture 2 it will be classified the present learner's cognitive state, thus will be used a pedagogical strategy more suitable following an action policy. In order to know the learner's profile the tutor will initially apply a questionnaire, which will allow a classification of the learner's cognitive state on what concerns different areas of knowledge as it can be observed in picture 3. The system will reclassify the learner with the updates from the pair table (state, action), nevertheless choosing actions which will take him to a better cognitive state than previously. At first it will not be possible to create good actions, because it is necessary to explore, in other words, to visit this pairs of state-action several times, then only this way will the system suggest the best action for a specific cognitive state.



Picture 3: Classification of the Learner's cognitive state

Thus originating according to Guelpeli (2000) [3] an adaptation of LTR showed in the first picture, in other words, for interaction between the environment (learner) and the agent (tutor) the use of a bond between them is defined by the diagnostic module demonstrated on picture four.



Picture 4: Representation of LTR model adapted to the ITS

V. SIMULATION'S METHODOLOGY

The model prototype used was the one proposed on Guelpeli (2000) [3] shown in picture 2, and it is divided in: Tutor module, diagnostic, pedagogical and Q table (s, a). In the prototype was defined that the environment of simulation would be in a matrix 5x10, i.e. the five states and the ten possible actions. In accordance with picture 3

the learner's cognitive level is based on four states, where there are values on each to distinguish states like an evolution of the learner's cognitive level, hence, the states: E_0=> [0,2]; E_1=>]2,4]; E_2=>]4,6] E_3=>]6,8]; E_4=>]8,10].Each state visited this model is a set of ribs and these ribs are snapshots: E0 => R = 1-Bad, E1 => R = 3-Regular, E2 => R = 5-Well, E3 => R 7 = Very Good; E4 => R = 10-Excellent.

Three kinds of model were created, which according to Guelpeli (2003)[4] are M1 model (bad), M2 Model (good), M3 Model (excellent) and can be deterministic and non-deterministic. Two pedagogic policies are also presented, denominated P1 and P2, where P2 is a more restrictive policy than P1 on what concerns model, because in it the intervals between the actions are shorter allowing a bigger action analysis in terms of states, i.e., there will be a larger pool of decisions for each state using the P2 policy. The simulation used in this work uses the M2 model (good) both deterministic and non-deterministic, and the policies P1 and P2. In the Q-Learning algorithm will be performed the alpha and gamma variation aiming to analyze the model's behavior to this variations.

1. *Q-LEARNING ALGORITHM*

Initialize **Q(s, a)**.

For each **t** instant repeat:

1- Observe estate **st** and chose **at** act according with the action policy (μ);

2- Observe the state **st +1** and update $Q^{\mu}_{t}(st, at)$ in accordance with: $Q^{\mu}_{t+1}(st,at) = {}^{\mu}_{t}(st,at) + \alpha t[r(st) + \gamma maxaQ^{\mu}_{t}(st+1,a) - Q^{\mu}_{t}(st,at)]$;

Until **t** achieve step limits.

Where it can be defined:

- $Q^{\mu}_{t+1}(s_t, a_t)$- It's the value (quality) of action a_t in the state s_t following the action policy *(μ)*.

- $r(s_t)$- It is the immediate reinforcement received at state s t

- α – it is the rate of learning

- γ- – it is the rate of temporal reduction

- t – It is a slight sequence of steps in time, i.e., t=0,1,2,3,…

- $max_a Q^{\mu}_{t}(s_{t+1}, a)$– Maximization policy, chooses the action with higher utility value in the future state.

- The factor γ (between 0 and 1) – the closer to 1, higher the importance given to far off reinforcements on time.

VI. EXPERIMENT VARYING ALPHA AND GAMMA

The M2 Model (good) deterministic and non-deterministic was submitted to simulations with a thousand steps, and on each set of simulation was calculated an average on ten accomplishments to obtain final data. The model was not previously known by the tutor, which will estimate through LTR technique an action policy on what concerns the received reinforcements. To analyze the model's behavior to the rate variations were assigned values for variables Alpha and Gamma, showed in tables 1, 2 and 3 about model behavior on rate variations.

In the first table is described alpha=0,1, i.e. closer to zero, making the learner unable to learn, also gamma is varied, analyzing P1 policy in the deterministic model, it is possible to observe that when gamma is also close to zero the algorithm convergence happens, but with an average reinforcement of R=1, i.e. bad, something that does not occur when gamma is 0,5 because the average reinforcement is R=2,5, very close to regular reinforcement, with gamma values = 0,9 the average reinforcements is R=4,5 which is a lot closer to reinforcement R=5, in other words, the good reinforcement, the optimal for the desired state to learn, the state E2, which the states transition varies but the algorithm converges in E2 state. It is also observed in this policy that the E2 state is the most visited in the three variations. For the same model and variations modifying to the policy P2 only, which is more restrictive, it is possible to observe that when gamma=0,1 the average reinforcements is R=4, increasing as gamma approaches to 1 being close to the optimal reinforcement for E2 learning. Although with this variation and policy when gamma=0,1 the variation makes E3 the most visited state.

By analyzing the non-deterministic model the variations are capable of varying the model near to the state desired to be learned, but not learning it, the medium reinforcements applied in P1 and P2 vary due to interference of policy and gamma variation. Happening with P1 as well as P2 the gamma value 0,1 makes E3 the most visited state and not E2, the same phenomenon occurs with gamma=0,5 in P2.

TABLE 1: M2 MODEL DETERMINISTIC AND NON-DETERMINISTIC, USING A=0,1 AND VARYING GAMMA BETWEEN Y=0.1, 0.5, 0.9.

M2 model								
α=0,1 γ=0,1	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	0,41	0,60	4,52	6,53	2,05	2,50	E2 - 868	E3 - 495
P2	3,12	0,61	33,26	6,63	2,32	2,56	E3 - 925	E3 - 589

M2 model								
α=0,1 γ=0,5	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	2,04	2,56	8,01	9,77	2,09	2,05	E2 - 880	E2 - 900
P2	3,39	3,069	3,49	11,67	2,25	2,56	E2 - 682	E3 - 611

M2 model								
α=0,1 γ=0,9	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	3,64	4,48	34,75	38,42	2,06	1,98	E2 - 883	E2 - 825
P2	3,44	4,52	36,57	38,42	2,23	2,01	E2 - 804	E2 - 809

In table 2 alpha posses intermediate value, not so close to zero, something that would not allow the learner to learn anything and not so close to 1 that would make him accomplish the leaning quickly replacing old information for the new ones, it is important to notice that the deterministic model on policy P1 and P2 behaves itself differently due to gamma variation, making model convergence in E2, however with out of optimal reinforcements for the state and with Q(s, a) in far off values for E2 state, which would be between 4 and 6. Visiting gamma 0,5 plus E3 state, followed by E0 and for a very brief moment the E2 state. With the non-deterministic model the variations interfere on ITS behavior in P1 as well as P2 making the E3 state being more visited than the one to be learned.

TABLE 2 M2 MODEL DETERMINISTIC AND NON-DETERMINISTIC, USING A=0,5 AND VARYING GAMMA BETWEEN Y=0.1, 0.5, 0.9.

M2 model								
α=0,5 γ=0,1	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	0,40	0,59	4,45	6,57	2,10	2,46	E2 - 876	E2 - 907
P2	3,38	0,61	36,06	6,77	2,25	2,55	E2 - 761	E3 - 758

M2 model								
α=0,5 γ=0,5	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	2,01	3,02	8,01	11,97	2,09	2,51	E2 - 868	E3 - 554
P2	3,11	3,06	33,04	12,14	2,35	2,56	E3 - 606	E3 - 579

M2 model								
α=0,5 γ=0,9	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	3,65	4,78	38,33	49,72	2,06	2,15	E2 - 869	E2 - 879
P2	3,38	5,23	36,29	53,79	2,25	2,40	E2 - 749	E2 -813

TABLE 3: M3 MODEL DETERMINISTIC AND NON-DETERMINISTIC, USING A=0,9 AND VARYING GAMMA BETWEEN Y=0.1, 0.5, 0.9

M2 model								
α=0,9 γ=0,1	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	0,40	0,56	4,48	6,28	2,09	2,32	E2 - 872	E2 - 574
P2	3,39	0,59	36,06	6,57	2,25	2,45	E2 - 789	E3 - 698

M2 model								
α=0,9 γ=0,5	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	2,02	2,83	8,07	11,35	2,08	2,33	E2 - 892	E2 - 641
P2	3,39	3,05	36,31	10,55	2,24	2,54	E2 - 772	E3 - 547

M2 model								
α=0,9 γ=0,9	average reinforcement		Q(s,a)		States transition		Total visits	
	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic	Deterministic	Non-Deterministic
P1	3,52	4,79	37,80	51,66	2,05	2,16	E2 - 873	E2 - 876
P2	3,42	5,19	36,44	54,94	2,24	2,38	E2 - 728	E3 - 547

In the third table with alpha=0,9 and varying gamma it is possible to observe that independently of policy for deterministic model when alpha and gamma are in balance and closer to 1 the convergence occurs in a more coherent way in accordance to the presented model, because as alpha approaches to 1 older information are replaced for new ones and gamma closer to 1 creates a reinforcement for future rewards of long term seeking for the best state (proposed before), i.e. E2, and not the best choice in the current moment. Highlighting that visit always occurs often to E2 state. With non-deterministic model alpha and gamma variation interfere and the most visited state is E2, though only with P1 policy.

VII. CONCLUSIONS

It is possible to observe that alpha and gamma values exercise influence in the convergence of Q-learning algorithm, naturally, interfering on ITS learning and consequently in module-student diagnosis. In both deterministic and non-deterministic models it is possible to notice the influence of pedagogical policies. We can observe that in P1 there are more visits due to its less restrictive nature in comparison with P2, which can be considered more restrictive, thus it is possible to assure that defining the pedagogical policy is another decisive and determinant factor on what concerns algorithm convergence and also ITS learning.

FUTURE WORK

To study the performances viabilities on what concerns analysis with other pedagogical policies with alpha and gamma variation. To improve algorithm convergence – especially for states with large spaces and with a notable amount of possible tutoring actions using variables based on compact approximations, mainstreaming experience strategies, or action plans obtained through simulation.

REFERENCES

[1] BACARDIT, J.: Peñarroya. Pittsburgh "Machine learning based on genetic data mining era: representations, generalization, and run time." Thesis (Ph.D.) Barcelona, Ramon Llull University, 2004

[2] Boff, R., Vieira, R., and Goulart, R. Teaching music history with the help of ontologies, New Hamburg – RS, FEEVALE University, 2004

[3] Guelpeli, M.V. Use of reinforcement learning for modeling autonomous learner in intelligent tutoring systems. Dissertation. São José dos Campos, São Paulo – SP Division of Computer Science Technolkogical Institute of Aeronautics, 2000.

[4] Guelpeli, M.V. C, Ribeiro, C.H. C., and Omar, N. *Using reinforcement learning for modeling learner autonomy in an intelligent tutor.*. In XIV Brazillian Symposium on Computers in Education., Rio de Janeiro. SBC, 2003.

[5] Guelpeli, M.V. C, Ribeiro, C.H. C., and Omar, N. *Reinforcement Learning to Intelligent Tutoring Systems without na explicit Learner's model*. Computers in Education brazillian magazine - RBIE , SBC index 12 nº 02 , p 69-77, 2004.

[6] Watkins, C. J.C.H., and Dayan, P. *Q-leaning*. *Machine Learning* 8(3/4):279-292, 1992.