Applying Expert Heuristic as an a Priori Knowledge for FRIQ-Learning

Tamás Tompa, Szilveszter Kovács

Department of Information Technology, University of Miskolc,
Miskolc-Egyetemváros, H-3515 Miskolc, Hungary
e-mail: tompa@iit.uni-miskolc.hu, szkovacs@iit.uni-miskolc.hu

Abstract: Many Reinforcement Learning methods start the learning phase from an empty, or randomly filled knowledge-base. Having some a priori knowledge about the way as the studied system could be controlled, e.g. in the form of some state-action control rules, the convergence speed of the learning process can be significantly improved. In this case, the learning stage could start from a sketch, from a knowledge-base formed based upon the already existing knowledge. In this paper. the a priori (expert) knowledge is considered to be given in the form state-action fuzzy control rules of a Fuzzy Rule Interpolation (FRI) reasoning model and the studied reinforcement learning method is restricted to be a Fuzzy Rule Interpolation-based Q-Learning (FRIQ-Learning) method. The main goal of this paper is the introduction of a methodology, which is suitable for merging the a priori state-action fuzzy control rule-base to the initial state-action-value function (Q-function) representation. For demonstrating the benefits of the suggested methodology, the a priori knowledge-base accelerated FRIQ-Learning solution of the "mountain car" benchmark is also discussed briefly in the paper.

Keywords: Reinforcement Learning; Heuristically Accelerated Reinforcement Learning; Fuzzy Rule Interpolation; Q-Learning; FRIQ-Learning

1 Introduction

The reinforcement learning (RL) (originally introduced in [22]), is still a popular machine learning algorithm among the devices of the computational intelligence. The RL methods are kind of trial-and-error type algorithms, solving problems without the explicit knowledge about the solution, but based on rewards leading to the targeted behaviour of the system. The rewards (reinforcements) are given by the environment, according to the observed and targeted behaviour, independently from the inner states of the RL agent. The original Q-learning [31] and the Fuzzy Q-learning (FQ-learning) [10], [4], [7] algorithms are starting with an empty knowledge-base and building their approximated Q-function during iterations based on the gained reward values. The Q-function representation in the case of

Q-learning is a Q-table. In the case of FQ-learning it is a fuzzy rule-base describing the Q-function in continuous state and action universes. These RL algorithms automatically build their knowledge-base during the learning process. Therefore, they can be applied in such situations, where there is no initial knowledge about the system to be controlled. These methods are starting from an empty knowledge-base at the beginning of the learning process, and their Q-function approximation is built through iterations. On the other hand, if there is an initial knowledge-base about the operating process may exist, this knowledge could form a draft for the initial RL model.

There are existing solutions for combining the RL methods with an initial expert knowledge-base, which is noted in this case as “heuristic”. These methods can be used in such systems, where the knowledge, or a portion of the knowledge about the system operation already exists, but the full control needs to be extended and adjusted based of the feedback of the working environment. One of these methods is the Heuristically Accelerated Reinforcement Learning, HARL [5]. In the HARL the heuristic is given in a form of a heuristic function ($H_t(s_t,a_t)$). It defines for the agent, which “a_t” action should be selected in the state “s_t”, at the time “t”. The combination of the HARL model with the traditional Q-learning, is called Heuristically Accelerated Q-learning (HAQL) [5]. Another possible solution for describing the heuristic is the formal knowledge representation with a declarative language. The “GOAL” is an agent programming language, which is defining the action selection for the agent by a set of “if then” type conditions [8]. In fuzzy rule-based Q-learning [10] the Q-function is represented by a fuzzy rule-base. In this case, it is straightforward that the a priori information of the expert should be also represented as a fuzzy rule-base. In [21] Pourhassan et al. are proposing a way to incorporate the expert knowledge in the Q-learning by means of fuzzy rules.

The main goal of this paper is to suggest a way for extending the Fuzzy Rule Interpolation (FRI) model-based Reinforcement Learning (FRIQ-Learning) methods to be able to adopt a priori expert knowledge about the problem solution in the form of fuzzy rules.

2 The FRIQ-Learning

The Fuzzy Rule Interpolation-based Q-learning (FRIQ-learning) [26][9] is an extension of the traditional Q-learning with continuous state and action space, represented by a FRI model. For Fuzzy Q-learning (FQ-learning) techniques, usually, the classical 0-order Takagi-Sugeno Fuzzy Inference model is adopted (see, e.g. [1], [3] and [11] for more details). In case of FRIQ-learning adapting FRI methods (see e.g. [2] for a short overview of FRI methods) for the FQ-learning can reduce the size of the fuzzy rule-base, by permitting the use of sparse fuzzy rule-bases for fuzzy knowledge representation. The sparse rule-base built for the FRI can represent the same or nearly the same approximated Q-function as the complete fuzzy rule-base with the classical (e.g. CRI [18]) fuzzy reasoning.

One of the available FRI techniques is the Fuzzy Rule Interpolation based on Vague Environment (FIVE) FRI, which was originally introduced in [13], [14] and [15]. The FIVE is a multidimensional, application-oriented FRI technique, which is based on the Vague Environment (VE) [12] concept. According to the VE concept, the fuzzy partitions of the antecedent and consequent universes can be represented by scaling functions [12]. The similarities of fuzzy sets can be calculated as the scaled distances of crisp points. Therefore, the FIVE can give a crisp conclusion directly without any additional defuzzification step. The combination of the FQ-learning with the FIVE FRI is called FRIQ-learning [26]. In the FRIQ-learning the state-action-value function is described by a sparse fuzzy rule-base and the Q-function is approximated by the FIVE FRI. The form of the i^{th}, $i \in [1, r]$ fuzzy rule in the Q-function rule-base is the following:

If s_1 **is** $S^i{}_1$ **And** s_2 **is** $S^i{}_2$ **And … And** s_n **is** $S^i{}_n$ **And** a **is** A^i **Then** $\tilde{Q}(\boldsymbol{s},a) = q^i$ (1)

where $S^i{}_j$ $j \in [1, n]$ is a label of a fuzzy set in the j^{th} dimension of the n dimensional state space $\boldsymbol{S}$, $\boldsymbol{s} \in \boldsymbol{S}$ is the n dimensional state observation, s_j is the j^{th} dimension of the state observation $\boldsymbol{s}$, A^i is the label of a fuzzy set in the one-dimensional action space U, $a \in U$ is the selected action, $\tilde{Q}(\boldsymbol{s},a)$is the approximated Q-function, q^i is the singleton conclusion of the i^{th} fuzzy rule. Applying the FIVE FRI model for the Q-function representation, according to [16], we get the following formulas:

$$\tilde{Q}(\boldsymbol{s},a) = \begin{cases} q^i & \text{if } (\boldsymbol{s},a) = (\boldsymbol{s}^i, a^i) \text{for some } i, \\ \displaystyle\sum_{i=1}^{r} \left(\left(q^i \Big/ \left(\delta_{\text{v}}^i\right)^{\lambda} \right) \Big/ \left(\sum_{j=1}^{r} 1 \Big/ \left(\delta_{\text{v}}^j\right)^{\lambda} \right) \right) & \text{otherwise.} \end{cases} \tag{2}$$

where q_i is the consequent of the i^{th} rule, $(\boldsymbol{s},a)$ is the crisp observation, λ is the Shepard parameter and r is the number of the rules in the rule-base. The δ_{v}^i is the scaled distance of the actual observed state, selected action $(\boldsymbol{s},a)$ value and the i^{th} fuzzy rule antecedent according to the scaling function v of the corresponding vague environment [12]:

$$\delta_{\text{v}}^i = \delta_{\text{v}}\left((\boldsymbol{s},a),(\boldsymbol{s}^i,a^i)\right) = \left[\sum_{j=1}^{n} \left(\int_{s_j^i}^{s_j} \text{v}_j(s_j)ds_j \right)^2 + \left(\int_{a^i}^{a} \text{v}(a)da \right)^2 \right]^{1/2} \tag{3}$$

where $(\boldsymbol{s},a)$ is the actual state and action, $(\boldsymbol{s}^i,a^i)$ is the antecedent part of the i^{th} rule, s_j is the j^{th} dimension of the n dimensional state universe, $\text{v}_j(s_j)$ is the scaling function of the s_j state universe, $\mathrm{v}(a)$ is the scaling function of the action universe *U*.

Substituting the formulas of the FIVE FRI (2) to the update form of the Q-learning, we get the q_i rule consequent of the *i*th fuzzy rule in the (*k*+1)th iteration in the following form:

$$q_i^{k+1} = \begin{cases} q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) & \text{if } (\boldsymbol{s},a) = (\boldsymbol{s}^i, a^i) \text{ for some } i, \\ q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) \cdot \left(1/\delta_{\mathrm{v},i}^{\lambda}\right) / \left(\sum_{i=1}^{r} 1/\delta_{\mathrm{v},i}^{\lambda} \right) & \text{otherwise.} \end{cases} \tag{4}$$

where $\Delta\tilde{Q}^{k+1}(\boldsymbol{s},a)$ is the (*k*+1)th update value of the Q-function in $(\boldsymbol{s},a)$:

$$\tilde{Q}^{k+1}(\boldsymbol{s},a) = \tilde{Q}^{k}(\boldsymbol{s},a) + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) \tag{5}$$

$$\Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) = \alpha \cdot \left(g(\boldsymbol{s},a,\boldsymbol{s}') + \gamma \cdot \max_{a' \in U} \tilde{Q}^{k}(\boldsymbol{s}',a') - \tilde{Q}^{k}(\boldsymbol{s},a) \right) \tag{6}$$

In this form, as in the original Q-learning [31], γ is the discount factor and $\alpha \in [0,1]$ is the step size parameter. The q_i^{k+1} is the *k*+1 iteration of the singleton conclusion of the *i*th fuzzy rule, taking action *a* in state $\boldsymbol{s}$, $\boldsymbol{s}'$ is the new observed state, $g(\boldsymbol{s},a,\boldsymbol{s}')$ is the observed reward completing the $\boldsymbol{s} \rightarrow \boldsymbol{s}'$ state-transition. The $\tilde{Q}^k$ and the $\tilde{Q}^{k+1}$ are the *k*th and the (*k*+1)th iteration of the Q-function approximated by the FIVE FRI (2).

For the action selection policy, the FRIQ-learning applies the greedy policy (or optionally the ε-greedy policy) [27], which is always selecting the action having the greatest Q value (or in case of ε-greedy, the greatest with ε probability) in the corresponding state. The greedy policy can be described by the following form:

$$\pi(s) = \arg\max_{a \in U} Q^{\pi}(s,a) \tag{7}$$

The FRIQ-learning was also extended with an automatic incremental rule-base creation method [27]. In this technique, based on reinforcements, the Q-function rule-base can be built automatically through iterations. The method starts from an "empty" rule-base, in which the rules are at the corners of the (n+1)-dimensional action-state space hypercube (this is required because of the definition of the interpolation, where n is the dimension of the state). In the further iteration steps, the initial rule-base grows according to the values of the updating rule (4). A new rule is inserted to the rule-base if the updating value of the state-action-value function ($\Delta\tilde{Q}$) is greater than a predefined limit and the existing rules are farther than a given distance. The position of the newly inserted rule is the closest possible (enabled) rule position (see [27] for more details). In case if the update value is smaller than the predefined limit or the given state-action point is close to an existing rule, then only the existing fuzzy rules consequents are updated.

The method has rule-base reduction strategies too. They are based on the approach, that for approximating the Q-function there is no need for all the rules created in the incremental phase. Some of the (redundant) rules could be removed without a relevant change in the Q-function representation. Applying the reduction strategies, these rules can be omitted from the rule-base. There are four different reduction strategies defined in [24], [28] and [29]. Three of them are based on the differences in the close rule consequences (Q-values) [28] [29] and one is based on rule clustering [24].

3 Expert Heuristic as a Priori Knowledge for Defining the Initial Rule-Base of the FRIQ-Learning

In case if there are some a priori knowledge about the system to be controlled, e.g. some kind of expert rules exist, the convergence speed of the Q-learning could be improved by their adoption. In this paper, the suggested way of this adoption is the merging of the expert knowledge to the initial rule-base of the FRIQ-learning. The expert knowledge, as an a priori information, is defined by a human expert before the learning process. In this paper, the suggested way of the expert knowledge expression is in the form of fuzzy rules. This case the a priori rule-base can be directly adapted to the initial fuzzy rule-based Q-function representation of the FRIQ-learning. During the learning process this initial knowledge representation will be tuned and modified (extended or reduced), then at the end of the process the expert rules can also be fetched back e.g. for expert rule validating purposes.

For merging the expert rules to the initial rule-base of the FRIQ-learning, the problem of the different rule representations must be solved. In the case of the FRIQ-learning, the fuzzy rules are state-action-value rules according to form (1), while the expert knowledge is usually expressed in the form of state-action production rules (8). The fuzzy rule consequences of the state-action-value Q-function representation in FRIQ-learning are the Q-values. On the other hand, the fuzzy rule consequences of the state-action production rules are expert-defined actions. The suggested form of the i^{th}, $i \in [1, r]$ expert-defined production fuzzy rule is the following:

$$\textbf{If } s_1 \textbf{ is } \hat{S}_1^i \textbf{ And } s_2 \textbf{ is } \hat{S}_2^i \textbf{ And } \ldots \textbf{ And } s_\mathrm{n} \textbf{ is } \hat{S}_n^i \textbf{ Then } a = \hat{A}^i \tag{8}$$

In structure, the form of (8) is very similar to (1), except the different types of consequents and the missing action antecedent in (8). The i^{th} rule consequent $\hat{A}^i$ is the expert-defined rule action for a given n dimensional state $\hat{\boldsymbol{S}}^i = \left[\hat{S}_1^i, \hat{S}_2^i, \ldots, \hat{S}_n^i\right]$.

For adapting the expert rules (8) in the initial rule-base of the Q-function representation, the missing Q-values must be determined. Considering the initial expert rules, as "valuable" decisions about the actions, and taking into account of the planned greedy action selection policy, the Q-values of the expert rules must be set to relatively higher initial values.

Having a different action selection policy than a greedy one, the given expert rules can be also considered to be a heuristic policy modifier [5]. Considering the expert rules to be always and unquestionably true, the greedy policy of the FRIQ-learning can be turned into a heuristic policy, which obeys the expert rules in the following form:

$$\pi(\boldsymbol{s}) = \begin{cases} a = \hat{A}^i, & \text{if } \boldsymbol{s} = \hat{\boldsymbol{S}}^i, \textit{ for some } i \\ \arg\max\limits_{a \in U} Q^{\pi}(\boldsymbol{s}, a) & \textit{otherwise.} \end{cases} \tag{9}$$

where $\hat{\boldsymbol{S}}^i$ and $\hat{A}^i$ are the state and action fetched from the i^{th} expert rule, and $\boldsymbol{s}$ is the actual state observation. If the actual observation $\boldsymbol{s}$ matches the state $\hat{\boldsymbol{S}}^i$ of one of the expert state-action rules, then the selected action will be the corresponding action $\hat{A}^i$. Otherwise, the greedy action selection of the FRIQ-learning will be followed.

Considering the expert rules to be always and unquestionably true, with the greedy policy for the rest of the state-action space, the FRIQ-learning can only extend the initial rule-base of the expert by additional rules. In this case, the goal of the suggested FRIQ-learning-based methodology is the extension of the expert-defined state-action rules by additional state-action rules for the state space area uncovered by the expert rules.

In case of supposing that the expert rules may be false, or incorrect, the goal of FRIQ-learning-based methodology can be extended by the tuning of the initial expert rule-base (moving, removing, or updating the expert rules).

It is also important to note that because of the incremental manner of the rule-base construction during the learning phase, there is no need for defining state-action expert rules for all the possible states (like an optimal policy), but it is sufficient to give the expert rules only in any states, where the expert has knowledge about the system. Therefore, the expert might define any number of state-action pairs. Thus, if the tuning of the initial expert rule-base is permitted during the learning phase, the quality of the expert-defined a priori information effects only the convergence rate of the FRIQ-learning.

4 Adaptation of the Expert Rule-Base

The suggested expert rule-base adaptation method of the FRIQ-learning is built upon two phases. In the first phase, the Q-value determination method calculates the initial approximated Q-values for the expert-defined rules. In the second phase the rule-base adaptation method combines the expert-defined state-action a priori rule-base (6) with the FRIQ-learning initial rule-base (1). Thereafter the combined rule-base will serve as the initial Q-function approximation rule-base of the FRIQ-learning process.

4.1 Determining the Initial Q-Values of the Expert Rules

The consequents of the expert rules are actions (defining a states-action function). Therefore, the rules of the expert knowledge representation have no Q-values. On the other hand, the rule representation of the FRIQ-learning describes a state-action-quality function (Q-function), where the quality of the state-action pairs must be determined. Therefore, to adopt the a priori expert rules to the Q-function rule representation, the corresponding Q-values must be determined. It must be done before the learning phase as an initialization step of the Q-function rule-base generation.

The goal of the proposed Q-function rule-base initialization method is to determine the initial, estimated Q-value ($\tilde{Q}_{\mathrm{init}}$) for each expert-defined state-action rules before the learning phase. According to the proposed concept, the rule Q-values should be initialized with an expert-defined quality ($\hat{Q}_{\mathrm{init}}$) value. I.e. the Q-values of the expert rules, together with the state reward value definitions are an inherent part of the expert knowledge representation. With full confidence, these values can not be determined independently from the corresponding expert rules. On the other hand, it could happen, that the expert heuristic contains only the worthy production rules, without any additional information related to the initial Q, or reward values. In this case it can be supposed, that the expert knowledge representation contains only the most important correct rules, and if the expert rule Q-values are missing, the initial Q-values of the expert rules can be approximated by a relatively "high" Q-value. The relatively "high" Q-value is an estimation. As a first straightforward estimate, if the initial Q-values definition is missing from the expert knowledge representation, this paper suggests setting the initial Q-value to be the same for all the rules, as a fraction of the estimated maximal achievable Q-value ($\tilde{Q}_{\max}$). Where $\tilde{Q}_{\max}$ can be approximated based on the maximal reward can be given by the environment. The initial Q-value $\tilde{Q}_{\mathrm{init}}$ can estimated by the following formula:

$$\tilde{Q}_{\mathrm{init}} = \eta \cdot \tilde{Q}_{\max} \tag{10}$$

$$\tilde{Q}_{\max} = \lim_{k\to\infty}\tilde{Q}^{k+1}\left(\mathbf{s}^*,a^*\right) = \lim_{k\to\infty}\left(\tilde{Q}^{k}\left(\mathbf{s}^*,a^*\right) + \alpha\cdot\left(g\left(\mathbf{s}^*,a^*,\mathbf{s}^*\right) + \gamma\cdot\tilde{Q}^{k}\left(\mathbf{s}^*,a^*\right) - \tilde{Q}^{k}\left(\mathbf{s}^*,a^*\right)\right)\right)$$

$$\tilde{Q}^{k}\left(\mathbf{s}^*,a^*\right) = \max_{a'\in U}\tilde{Q}^{k}\left(\mathbf{s}^*,a'\right) \text{ and } g\left(\mathbf{s}^*,a^*,\mathbf{s}^*\right) = \max_{\mathbf{s}\in\mathbf{S},a\in U} g\left(\mathbf{s},a,\mathbf{s}'\right) = g_{\max}$$

$$\tilde{Q}_{\max} = \lim_{k\to\infty}\tilde{Q}^{k+1}\left(\mathbf{s}^*,a^*\right) = \lim_{k\to\infty}\left(\tilde{Q}^{k}\left(\mathbf{s}^*,a^*\right) + \alpha\cdot\left(g_{\max} + (\gamma-1)\cdot\tilde{Q}^{k}\left(\mathbf{s}^*,a^*\right)\right)\right) =$$

$$= \tilde{Q}^{k}\left(\mathbf{s},a\right) + \alpha\cdot g\left(\mathbf{s},a,\mathbf{s}'\right) + \alpha\cdot(\gamma-1)\cdot\tilde{Q}^{k}\left(\mathbf{s}',a'\right) = \frac{\alpha\cdot g_{\max}}{-\alpha\cdot(\gamma-1)} = \frac{g_{\max}}{1-\gamma} \tag{11}$$

$$\tilde{Q}_{\text{init}} = \eta\cdot\frac{g_{\max}}{1-\gamma}\text{, in case if } \gamma < 1 \tag{12}$$

where $\tilde{Q}_{\max}$ is the estimated maximal achievable Q-value having $g_{\max}$ being the maximal reinforcement could be given by the environment. $\eta\in[0,1]$ is the discount factor of the $\tilde{Q}_{\text{init}}$ estimation.

There are other initial Q-values approximation methods can be found in the literature, e.g. for discrete state Q-learning in [19] the initial Q-values are determined based on the reward value of the goal state applying a binary reward function, for fuzzy Q-learning in [21] the initial Q-values are determined based on expert knowledge related to the estimated Q values of some states.

4.2 Merging the Expert Rules with the Initial Rule-Base of the FRIQ-Learning

Being in interpolated Q function representation, the initial rules of the FRIQ-learning has to hold the corners (corner rules) of the $(n+1)$-dimensional state-action space [27]. Therefore, the number of the initial fuzzy rules are 2^{n+1}. E.g. in case of two states and one action, it is $2^{2+1}=8$. If the number of the expert rules is $\hat{r}$, the size of the initial merged rule-base has $2^{n+1}+\hat{r}$ rules. According to the FRIQ-learning initial rule definition suggested in [27], the initial rule consequent values of the corner rules are $q_i = 0$. Therefore, the i^th^ corner rule $r^{\square i}$ has the following format:

If s_1 **is** $\mathrm{S}_1^{\square i}$ **And** s_2 **is** $\mathrm{S}_2^{\square i}$ **And…And** s_n **is** $\mathrm{S}_n^{\square i}$ **And** a **is** $\mathrm{A}^{\square i}$ **Then** $\tilde{Q}(\mathbf{s},a)$=0 (13)

where $\mathrm{S}_l^{\square i}\in[\min(S_l),\max(S_l)],\forall i,l$, $\mathrm{A}^{\square i}\in[\min(A),\max(A)],\forall i$ are the corner state and action values and see Eq. (1) for the rest of the notation.

For the expert rules, the initial rule consequent values are $q_i = \tilde{Q}_{\text{init}}$, $i\in[1,\hat{r}]$ hence the i$^{\text{th}}$ expert rule, $\hat{r}^{\,i}$ has the following format:

If s_1 **is** $\hat{S}_1^i$ **And** s_2 **is** $\hat{S}_2^i$ **And…And** s_n **is** S^i_n **And** a **is** $\hat{A}^i$ **Then** $\tilde{Q}(\mathbf{s},a) = \tilde{Q}_{\text{init}}$ (14)

where see Eq. (8) for the notation.

In case if there is an expert rule, which hit the position of the initial corner rules, the overlapping expert rule will replace the corresponding initial corner rule.

The main steps of the suggested Q-function rule-base initialization are summarized on Figure 1.

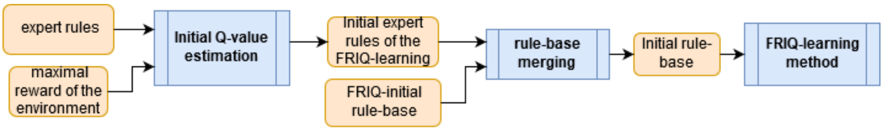
expert rules
maximal reward of the environment
Initial Q-value estimation
Initial expert rules of the FRIQ-learning
FRIQ-initial rule-base
rule-base merging
Initial rule-base
FRIQ-learning method

Figure 1
The suggested FRIQ-learning rule-base initialization

5 FRIQ-Learning and the Mountain Car Problem

The goal of this section is to give an application example for the suggested merging of the a priori state-action expert rules with the initial corner rule-base of the FRIQ-learning (i.e. for the suggested rule-base initialization) in a standard RL task.

The chosen example is the well-known “mountain car” RL benchmark example. The mountain car problem is the task of a car for getting out of a deep valley. Initially, the car is situated in the center of the valley. The goal is to get out of the valley by going to the top of the hill within a given time frame. In this example the problem is considered to be solved if the car gets out of the valley in less than 1000 iteration steps. If the car reached the goal or the 1000 iteration steps elapsed, then an episode is completed. The full learning phase will be done if the Q-update values are smaller to a predefined Q-update limit (e.g. 0.05) through some episodes and if the size of the rule-base is not changed (not adding a new rule).

The problem has two states and one action variable. The states descriptors are the velocity and the position of the car and the action variable is the left, right, or neutral movement of the car:

- s1: velocity of the car
- s2: position of the car
- a: movement of the car (left, right, neutral)

The rule-base initialization is done according to the suggested expert rule-base merging discussed in section 4. The benefit of the suggested rule-base initialization is measured by the achievable performance gain.

In the first example, the performance of the expert rules extended initial rule-base will be compared to the empty initial rule-base (according to Eq. 12). The effect of the expert rule-base quality will be also studied by comparing a well-formed proper initial expert rule-base in the second example to a partially correct and in the third example to a randomly generated initial "expert" rule-base. During the performance investigation of the well-formed proper initial expert rule-base, the effect of the proper initial rule consequent value $\tilde{Q}_{\text{init}}$ selection will be also discussed.

The performance of the FRIQ-learning can be characterized by the convergence rate of the learning. In this paper, the convergence rate is calculated as the average number of episodes required for adapting the Q-function rule-base to be able to solve the mountain car problem. One episode lasts till the car gets out from the valley, or 1000 iteration steps without solution. The averages, the convergence rate and the number of the required rules, are estimated based on independent runs starting from different initial state space positions. The reward given by the environment is g_{max}=100 (an expert suggested value) for the state if the car reaches the goal position (top of the valley in less than 1000 iteration steps). During the iteration, Eq. (6) was applied for the $\tilde{Q}$ updates. The learning parameters were the followings:

- learning rate (α): 0.5
- discount factor (γ): 0.99

If the system starts from an empty rule-base without the expert-defined initial rules (corner rules only), then the average convergence rate of 10 independent run became 28.3 episodes, with 91.7 rules (at the end of the incremental rule-base creation phase, before the rule-base reduction, see Table 1. for the detailes).

Table 1

The results when starting with the empty knowledge-base

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	23	36	34	35	20	34	25	26	29	21	28.3
Rule-base size	80	85	82	96	105	90	89	98	99	93	91.7

The first task of the suggested Q-function rule-base initiation method, is the estimation of the $\tilde{Q}_{\text{init}}$ values according to Eq. (11). For the $\tilde{Q}_{\text{init}}$ values estimation we have to determine a suitable value of the η discount factor (see Eq. (11)). The effect of the η discount factor is problem dependent. In this paper, we study its effect on the mountain car problem in case of having a properly set initial expert rule-base.

In our example the properly set initial expert rule-base was generated by a single run of the automatic incremental rule-base creation technique introduced in [27], together with the rule-base reduction strategies III and IV introduced in [28] and [24]. The remaining 17 rules (see e.g. on Table 2) became the properly set initial expert rule-base of our example.

Table 2

Rules of the well-formed proper initial expert rule-base, where the **a** is the rule consequent

R#	1	2	3	4	5	6	7	8	9
s_1	-0.5	-0.475	-0.475	-0.27	-0.27	-0.475	-0.065	-0.475	-0.68
s_2	0	-0.014	0.014	0.014	-0.014	0.042	-0.014	-0.042	0.042
a	-1	1	1	1	1	1	0	-1	-1

R#	10	11	12	13	14	15	16	17
s_1	-0.065	-0.065	0.14	-0.27	-0.885	-0.65	-1.09	0.14
s_2	0.042	0.014	-0.014	-0.042	0.042	0.042	0.042	-0.014
a	1	0	-1	-1	-1	0	-1	0

The maximal reward value (g_{max}) is defined by the expert. In this example it is set to 100. From the maximal reward value, the suggested $\tilde{Q}_{\text{init}}$ initial Q-value was calculated according to Eq. (11). Table 3 contains the initial rule-base (expert rules merged with the FRIQ initial rules before the learning phase).

Table 3

The initial Q-values rule-base with the well-formed proper initial expert rules, where the *Q* is the rule consequent

R#	1	2	3	4	5	6	7	8	9
s_1	-0.5	-0.475	-0.475	-0.27	-0.27	-0.475	-0.065	-0.475	-0.68
s_2	0	-0.014	0.014	0.014	-0.014	0.042	-0.014	-0.042	0.042
a	-1	1	1	1	1	1	0	-1	-1
Q	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$

R#	10	11	12	13	14	15	16	17
s_1	-0.065	-0.065	0.14	-0.27	-0.885	-0.65	-1.09	0.14
s_2	0.042	0.014	-0.014	-0.042	0.042	0.042	0.042	-0.014
a	1	0	-1	-1	-1	0	-1	0
Q	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$

R#	18	19	20	21	22	23	24	25
s_1	-1.5	-1.5	-1.5	-1.5	0.3450	0.3450	0.3450	0.3450
s_2	-0.07	-0.07	0.07	0.07	-0.07	-0.07	0.07	0.07
a	-1	1	-1	1	-1	1	-1	1
Q	0	0	0	0	0	0	0	0

The first 17 rules in Table 3 are the rules of the expert rule-base with the suggested $\tilde{Q}_{\text{init}}$ values estimation. The last $2^3=8$ rules (18…25) are the initial corner rules of the empty Q-function rule-base, according to Eq. 13.

The next step is the η discount factor estimation (see Eq. (11), (12)) by checking its effect to the convergence rate having the properly set initial expert rule-base. Table 4 demonstrates the dependency of the convergence rate from the value of the η discount factor with the corresponding initial $\tilde{Q}_{\text{init}}$ values according to Eq. (12).

Table 4

The convergence rate in case of different η discount factor values (γ=0.99, $g_{\max}$ =100)

η	$\tilde{Q}_{\text{init}}$	convergence rate (episodes)
1	10000	23
0.75	7500	23
0.6	6000	30
0.37	3700	29
0.075	750	25
0.00015	1.5	27

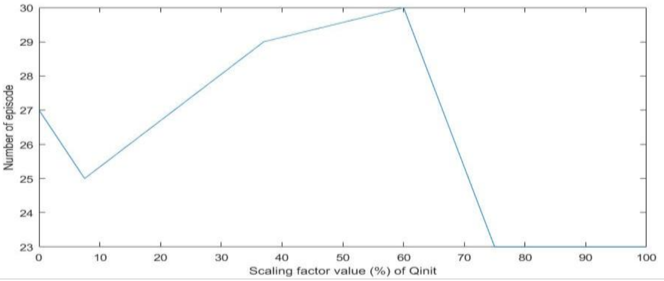

Figure 2

The convergence rate in case of different η discount factor values (γ=0.99, $g_{\max}$ =100)

According to the results (see Figure 2), in this given mountain car example the best convergence rate (23 episodes) can be achieved if the η discount factor is between 0.75 and 1 (γ=0.99, $g_{\max}$ =100).

For checking the performance of the suggested merging of the a priori state-action expert rules with the initial rule-base of the FRIQ-learning, five tests were performed. The first example is the properly defined expert heuristic case, where the initial rule-base of the FRIQ-learning is constructed with all the properly given initial expert rules having the suggested $\tilde{Q}_{\text{init}}$ values (according to Eq. (12)) (Table 5). The second example is a partial lack of knowledge, where the initial rule-base of the FRIQ-learning is constructed from a fragment of the properly given initial expert rules (Table 6). The third example is a partially proper expert knowledge, where some of the expert given initial rules are incorrect (Table 8). The fourth example is a fully incorrect expert knowledge, where all the expert given initial rules are incorrect (Table 10). The fifth example is a full lack of knowledge, where the initial rule-base of the FRIQ-learning is constructed without any expert given initial rules (see the results in Table 1).

The effect of the properly given initial expert rules is demonstrated on Table 5. In this case, the system found the final solution (the car gets out of the valley within 1000 iteration steps) in 10 episodes with 124.3 rules averagely. The expert rule-base contains 17 properly given initial expert rules. This rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 25 rules of the suggested initial rule-base (see Table 3). To reduce the final rule-base size one of the FRIQ-learning reduction strategies [24], [28], [29] could be applied.

Table 5

The results starting with properly given initial expert rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	10	20	17	7	11	10	6	5	6	8	10
Rule-base size	108	125	139	109	135	129	107	124	133	134	124.3

The effect of the partial lack of knowledge, where the initial rule-base of the FRIQ-learning is constructed from a fragment of the properly given initial expert rules, is demonstrated on Table 6. In this case, the system found the final solution in 14.4 episodes with 114.3 rules averagely. The expert rule-base contains 10 rules from the original 17 properly given initial expert rules. This partial expert rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 18 rules of the suggested initial rule-base.

Table 6

The results starting with partial lack of initial expert rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	20	13	10	7	7	15	29	15	22	6	14.4
Rule-base size	107	85	102	85	98	96	111	107	110	98	114.3

The effect of the partially proper expert knowledge, where some of the expert given initial rules are incorrect, is demonstrated on Table 8. In this case, the system found the final solution in 11.7 episodes with 120.1 rules averagely. The expert rule-base contains 11 rules from the original 17 properly given initial expert rules and 6 rules, where the rule consequents are changed to actions (see Table 7, rules no. 1, 2, 3, 15, 16, 17) which have an incorrect conclusion. This partially proper expert rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 25 rules of the suggested initial rule-base.

Table 7

The partially correct expert rule-base

R#	1	2	3	4…14	15	16	17
s_1	-0.5	-0.475	0.475	…	-0.68	-1.09	0.14
s_2	0	-0.014	-0.014	…	0.042	0.042	-0.014
a	0	1	-1	…	0	0	1

Table 8

The results when starting with the partially correct expert rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	8	16	8	13	7	16	10	15	16	7	11.7
Rule-base size	115	134	126	133	135	126	123	135	147	127	120.1

The effect of the fully improper expert knowledge, where all the expert given initial rules are incorrect, is demonstrated in Table 10. In this case, the system found the final solution in 26.6 episodes with 124.4 rules averagely. In this example, the initial “expert” rule-base is 17 randomly generated rules (see Table 9). This improper expert rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 25 rules of the suggested initial rule-base.

Table 9

The randomly generated “expert” rule-base

R#	1	2	3	4	5	6	7	8	9
s_1	-0.475	-0.5	-0.475	-0.475	-0.27	-0.27	-0.27	-0.475	-0.475
s_2	0	0	-0.014	0.014	0	-0.014	0	-0.042	0
a	1	-1	-1	0	-1	0	-1	1	1

R#	10	11	12	13	14	15	16	17
s_1	-0.475	-0.065	0.14	-0.27	-0.885	0.885	-0.065	-1.09
s_2	0	0	-0.014	-0.042	0.042	0.042	0.042	0.042
a	-1	0	1	-1	-1	1	0	-1

Table 10

The results when starting with the randomly generated “expert” rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	29	56	19	16	24	18	37	29	20	17	26.6
Rule-base size	122	127	118	124	131	120	130	124	127	121	124.4

Table 11 summarizes the results of the five different expert rule-base quality cases.

Table 11
The effect of the expert rule-base quality upon the convergence rate and the rule-base size

Expert rule-base type	Convergence rate	Rule-base size
empty	28.3	91.7
properly given	10	124.3
properly given fragment	14.4	114.3
partially incorrect	11.7	120.1
randomly generated	26.6	124.4

Conclusions

In this paper a methodology is suggested, which is suitable for merging an expert heuristic (an a priori state-action fuzzy production rule-base) to the initial state-action-value (Q-function) rule-base of the FRIQ-learning system. The expert-defined a priori rule-base is a preliminary knowledge about the given RL problem. The suggested merging is based on the reformulation of the expert heuristic given in the form of production (state - action) rules to the rule format (state, action - Q value) of the Q-function representation fuzzy rule-base by adding initial Q-values as consequents to them. The proper initial Q-values of the expert rules must be defined by the expert together with the rule definition. With full confidence, these values cannot be determined independently from the corresponding expert rules. For determining the initial Q-values in case if the expert heuristic contains only the worthy production rules, without any additional information related to the initial Q, or reward values, this paper suggest to set the initial Q-value to be the same for all the rules, as a fraction of the estimated maximal achievable Q-value.

For demonstrating the performance of the suggested initial FRIQ-learning rule-base construction methodology, the quality effect of the merged a priori expert rule-base is discussed. The performance of the FRIQ-learning solution of the “mountain car” benchmark example is studied in the case if the a priori state-action expert rules are fully properly defined, partly properly defined, partly improperly defined and fully improperly defined. The results are compared to the lack of a priori knowledge in average convergence rate and in rule-base size (without rule filtering). The best performer in convergence rate was the initial rule-base constructed with the fully properly defined a priori expert rules. The fully improperly defined a priori expert rules has similar convergence performance as the FRIQ-learning starting from an empty initial rule-base. On the other hand because of the unfiltered incremental manner of the rule-base creation, in rule-base size, the FRIQ-learning starting from an empty initial rule-base has the best performance.

The benefits of the suggested expert heuristic injection to the FRIQ-learning are twofold. The first is the improvement of the convergence speed, as it was discussed in this paper. The second is a way for validating the expert heuristic in given situations. Marking the injected expert heuristic rules during the Q-function initialization and fetching them back after the learning phase, the change of the expert production rules can be determined. Small changes can support, large changes or rule disappearance can disapprove the validity of the expert heuristic in the given situation defined by the environment of the learning phase. This kind of validation of the expert heuristic could be beneficial in application areas, where heuristical rule-based models exists, but the collection of vast data has some difficulties, like adaptive affective [30], or ethorobotical [20] models applied for human-machine interaction.

Acknowledgement

The described article was carried out as part of the EFOP-3.6.1-16-00011 ”Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialisation” project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

References

[1] Appl, M.: Model-based Reinforcement Learning in Continuous Environments. Ph.D. thesis, Technical University of München, München, Germany, dissertation.de, Verlag im Internet (2000)

[2] Baranyi, P., Kóczy, L. T., Gedeon, T. D.:A Generalized Concept for Fuzzy Rule Interpolation, IEEE Trans. on Fuzzy Systems, Vol. 12, No. 6, 2004, pp. 820-837

[3] Berenji, H. R.: Fuzzy Q-Learning for Generalization of Reinforcement Learning. Proc. of the 5th IEEE International Conference on Fuzzy Systems (1996) pp. 2208-2214

[4] Berenji, H. R.: Fuzzy Q-Learning for Generalization of Reinforcement Learning. Proc. of the 5th IEEE International Conference on Fuzzy Systems, pp. 2208-2214, 1996

[5] Bianchi, Reinaldo AC, Carlos HC Ribeiro, and Anna HR Costa. "Accelerating autonomous learning by using heuristic selection of actions." *Journal of Heuristics* 14.2 (2008): 135-168

[6] Bianchi, Reinaldo AC, Carlos HC Ribeiro, and Anna HR Costa. "Accelerating autonomous learning by using heuristic selection of actions." *Journal of Heuristics* 14.2 (2008): 135-168

[7] Bonarini, A.: Delayed Reinforcement, Fuzzy Q-Learning and Fuzzy Logic Controllers. In Herrera, F., Verdegay, J. L. (Eds.) Genetic Algorithms and Soft Computing, (Studies in Fuzziness, 8), Physica-Verlag, Berlin, D, (1996), pp. 447-466

[8] Broekens, Joost, Koen Hindriks, and Pascal Wiggers. "Reinforcement learning as heuristic for action-rule preferences." *International Workshop on Programming Multi-Agent Systems*. Springer Berlin Heidelberg, 2010

[9] D. Vincze, Fuzzy Rule Interpolation and Reinforcement Learning. Proceedings of the IEEE 15th International Symposium on Applied Machine Intelligence and Informatics, Herlany, Slovakia (2017) pp. 173-178

[10] Glorennec, Pierre Yves. "Fuzzy Q-learning and dynamical fuzzy Q-learning."*Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the Third IEEE Conference on*. IEEE, 1994

[11] Horiuchi, T., Fujino, A., Katai, O., Sawaragi, T.: Fuzzy Interpolation-Based Q-learning with Continuous States and Actions. Proc. of the 5th IEEE International Conference on Fuzzy Systems, Vol. 1 (1996) pp. 594-600

[12] Klawonn, F.: Fuzzy Sets and Vague Environments, Fuzzy Sets and Systems, 66, 1994, pp. 207-221

[13] Kovács, Sz., Kóczy, L. T.: Approximate Fuzzy Reasoning Based on Interpolation in the Vague Environment of the Fuzzy Rule base as a Practical Alternative of the Classical CRI. Proceedings of the 7th International Fuzzy Systems Association World Congress, Prague, Czech Republic, 1997, pp. 144-149

[14] Kovács, Sz., Kóczy, L. T.: The use of the concept of vague environment in approximate fuzzy reasoning. Fuzzy Set Theory and Applications, Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Bratislava, Slovak Republic, Vol. 12, 1997, pp. 169-181

[15] Kovács, Sz.: New Aspects of Interpolative Reasoning. Proceedings of the 6th. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain, 1996, pp. 477-482

[16] Kovács, Szilveszter. "Extending the Fuzzy Rule Interpolation" FIVE" by Fuzzy Observation." *Computational Intelligence, Theory and Applications* (2006): 485-497

[17] Krizsán, Z., Kovács, Sz.: Gradient based parameter optimisation of FRI "FIVE", Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest, Hungary, November 6-8, pp. 531-538, (2008)

[18] Mamdani, Ebrahim H., and Sedrak Assilian. "An experiment in linguistic synthesis with a fuzzy logic controller." *International journal of man-machine studies* 7.1 (1975): 1-13

[19] Matignon, Laëtitia, Guillaume J. Laurent, and Nadine Le Fort-Piat. "Reward function and initial values: better choices for accelerated goal-directed reinforcement learning." *International Conference on Artificial Neural Networks*. Springer Berlin Heidelberg, 2006

[20] Miklosi, A; Korondi, P; Matellan, V; Gacsi, M: Ethorobotics: A New Approach to Human-Robot Relationship, Frontiers in Psychology, Vol. 8, Paper: 958, 8 p. (2017)

[21] Pourhassan, Mojgan, and Nasser Mozayani. "Incorporating expert knowledge in Q-learning by means of fuzzy rules." *Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA'09. IEEE International Conference on*. IEEE, 2009

[22] Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1, No. 1, Cambridge: MIT press, 1998

[23] The original (discrete state- and action space) cart-pole problem can be found at: http://www.jamh-web.appspot.com/download.htm, last access date: 22.10.2019

[24] Tompa, Tamás, and Szilveszter Kovács. "Clustering-based fuzzy knowledge-base reduction in the FRIQ-learning." *Applied Machine Intelligence and Informatics (SAMI), 2017 IEEE 15th International Symposium on*. IEEE, 2017

[25] Tompa, Tamás, and Szilveszter Kovács. "Determining the minimally allowed rule-distance for the incremental rule-base contruction phase of the FRIQ-learning." 2018 19th International Carpathian Control Conference (ICCC) IEEE, 2018

[26] Vincze, D., Kovács, Sz.: Fuzzy rule interpolation-based Q-learning. *Applied Computational Intelligence and Informatics, 2009, SACI'09, 5th International Symposium on*. IEEE, 2009

[27] Vincze, D., Kovács, Sz.: Incremental Rule Base Creation with Fuzzy Rule Interpolation-Based Q-Learning, I. J. Rudas et al. (Eds.), Computational Intelligence in Engineering, Studies in Computational Intelligence, Volume 313/2010, Springer-Verlag, Berlin Heilderberg, 2010, pp. 191-203

[28] Vincze, D., Kovács, Sz.: Reduced Rule Base in Fuzzy Rule Interpolation-based Q-learning, Proceedings of the 10th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, CINTI 2009, November 12-14, 2009, Budapest Tech, Budapest, pp. 533-544

[29] Vincze, D., Kovács, Sz.: Rule-Base Reduction in Fuzzy Rule Interpolation-Based Q-Learning, Recent Innovations in Mechatronics (RIiM) Vol. 2, (2015) No. 1-2

[30] Vircikova, M., Magyar, G., Sincak, P.: The Affective Loop: A Tool for Autonomous and Adaptive Emotional Human-Robot Interaction. In: Kim JH., Yang W., Jo J., Sincak P., Myung H. (eds) Robot Intelligence Technology and Applications 3. Advances in Intelligent Systems and Computing, Vol. 345, Springer, pp. 247-254 (2015)

[31] Watkins, C. J. C. H.: Learning from Delayed Rewards. Ph.D. thesis, Cambridge University, Cambridge, England (1989)