ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Relevance Determination in Reinforcement Learning

Katharina Tluk v. Toschanowitz[1], Barbara Hammer[2] and Helge Ritter[1]

1- University of Bielefeld - Neuroinformatics Group - Technical Faculty
Universitätsstraße 25, 33615 Bielefeld - Germany

2- Clausthal University of Technology - Institute of Computer Science
Julius Albert Straße 4, 38678 Clausthal-Zellerfeld - Germany

Abstract. We propose relevance determination and minimisation schemes in reinforcement learning which are solely based on the Q-matrix and which can thus be applied during training without prior knowledge about the system dynamics. On the one hand, we judge the relevance of separate state space dimensions based on the variance in the Q-matrix. On the other hand, we perform Q-matrix reduction by means of a combination of Q-learning with neighbourhood cooperation of the state values where the neighbourhood is defined based on the Q-values itself. The effectivity of the methods is shown in a (simple though relevant) gridworld example.

1 Introduction

Reinforcement learning (RL) provides an elegant framework for modeling biological and technical reward-based learning. Since RL algorithms only require a scalar reinforcement signal, they are ideally suited for complex real-world applications where the optimal solution and possibly also the system dynamics are not known beforehand, such as self-exploring robots, complex control, and the synthesis of non-trivial motor behaviour [1, 2]. Q-learning which directly learns a strategy without prior inference of a model has been invented by Watkins and it has been pioneered and mathematically investigated by different researchers, see e.g. [3, 4, 5]. A potential problem of RL and Q-learning is given by the size of the search space. States are often high dimensional and the state variables might contain a large number of qualitatively different, even continuous values. In such cases, RL is only feasible if a good heuristic to shape the search space can be found such as bootstrapping methods, a split into separate modules, an equivalent representation of the search space with lower cardinality or an approximation or shaping of the value function [4, 6, 7, 8, 9].

These solutions are often time consuming and situation dependent, and global convergence is not necessarily guaranteed. Thus, the question of a universal model minimisation algorithm arises. Recently, Ravindran and coworkers proposed an interesting general framework based on work of Dean et al. which allows a minimisation of the model based on general assumptions [10, 11]. Basically, they identify equivalence classes of the state space which behave identically with respect to the model dynamics and the reward gained in these states. This formulation yields a standard model minimisation procedure comparable to the classical minimisation of finite automata. Their method, however, assumes that the model dynamics as well as the rewards are known, and a model minimisation is performed prior to training. Q-learning has been proposed for situations where no prior knowledge about the process and the expected reward is available, and training takes place solely based on the observed rewards. Here, we propose two methods for model minimisation for Q-learning which transfer the idea of equivalence classes to the Q-matrix and which do not require any prior information about the dynamics: relevance determination of state space dimensions by means of the Q-matrix variance and an extension of Q-learning by neighbourhood cooperation induced by the Q-values themselves.

2 Q-matrix learning and state space equivalence

We consider the objective of learning a strategy with maximum overall discounted reward at time t: $\sum_i \gamma^i r_{t+i}$ where r_{t+i} is the reward at time $t+i$ and $\gamma < 1$ is the discount factor. Standard one-step Q-learning [3] finds an optimal strategy by means of the following iterative update:

$$Q_{s_t,a_t} \leftarrow Q_{s_t,a_t} + \alpha \left[r_{t+1} + \gamma \text{max}_a (Q_{s_{t+1},a} - Q_{s_t,a_t}) \right]$$

where s_t and a_t are the state and action at time t and α is the learning rate. In the limit, $Q_{s,a}$ estimates the expected overall reward when taking action a in state s. An optimal strategy can be derived thereof chosing the action $\text{argmax}_a Q_{s_t,a}$ with maximum overall expected reward in state s_t.

The state space often contains a large (or even infinite) number of states s and convergence of Q-learning is slow. A general method of reducing the state space prior to learning without altering the system dynamics has recently been proposed by Ravindran: states are divided into equivalence classes where two states s and s' are equivalent if (i) the immediate reward gained in state s and s' is identical, and (ii) the probability of reaching a specific equivalence class is the same for s and s'. In this scenario, it is sufficient to work on the (smaller) number of equivalence classes instead of the full state space. This approach has the drawback that the immediate reward and the system dynamics need to be known prior to training; it is therefore not applicable in an unknown environment, the standard situation of Q-learning. We now transfer this idea into a formulation in terms of the Q-matrix. The conditions (i) and (ii) ensure that a strategy with the same overall expected reward can be chosen from every two states s and s' in the same equivalence class, as shown in [11]. Note that this formulation ensures that the set of qualities of *all* strategies reachable from s and s' coincides for s and s'. A simpler criterium which only ensures that the same *optimum* value can be reached from s and s', however, can be directly tested by means of the Q-values: s and s' are equivalent if and only if $\max_a Q_{s,a} = \max_{a'} Q_{s',a'}$. As another sufficient condition, one can identify state-action pairs where $Q_{s,a} = Q_{s',a'}$.

3 Reduction based on Q-values

Based on this argument, we propose two different algorithms which reduce the state space during training based on the approximate Q-values. Naturally, we can restrict ourselves to promising regions of the search space which are visited during state space exploration and within an optimum strategy.

Variance based relevance determination (VRL): VRL aims at reducing the dimensionality of the search space by estimating the relevance of the input dimensions. For this purpose, we compute the average variance of the Q-values for each search space dimension. This measure roughly estimates the degree of violation of the equality $Q_{s,a} = Q_{s',a'}$ if elements along a specified dimension are put into the same equivalence class. Assume $\mathbf{s} = (s_1, \ldots, s_M) \in \mathbb{R}^M$. Then the relevance of dimension d for the pair $(\mathbf{s}, a)$ is defined as $\xi^d_{\mathbf{s},a} = \frac{1}{N_d - 1} \sum_{s_d=1}^{N_d} (Q_{\mathbf{s},a} - E_d(Q_{\mathbf{s},a}))^2$ where $\{1, \ldots, N_d\}$ is the enumeration of the dth dimension and $E_d(Q_{\mathbf{s},a}) = \frac{1}{N_d} \sum_{s_d=1}^{N_d} Q_{\mathbf{s},a}$ denotes the expected value within the corresponding dimension. $\xi^d \sim \sum_{\mathbf{s},a} \xi^d_{\mathbf{s},a}$ indicates which dimensions are most irrelevant and can be discarded with the least impact on the solution.

Neighbourhood cooperative Q-learning (NRL): The second approach aims at identifying equivalence classes of a more sophisticated form which are not parallel to state space axes. Since, on the one hand, the equality $Q_{s,a} = Q_{s',a'}$ is only approximately fulfilled for equivalent pairs in early stages of training, and, on the other hand, a false identification of non-equivalent pairs with similar Q-values might be fatal, we use a soft approach which adapts all Q-values within each step based on their similarity compared to the optimum. This neighbourhood cooperation induced by the similarity of Q-values is similar to unsupervised clustering paradigms such as Neural Gas [12] and it boosts a formation of Q-value oriented clusters during training. However, unlike approaches such as [13] it does not use neighbourhood cooperativity induced by the input representation which would make the model unsuitable for discontinuous value functions. The adaptation rule of NRL is given by the standard update rule for the actual state action pair (s_t, a_t) and the adaptation of all other state-action pairs based on their proximity to the actual Q-value

$$\Delta Q_{s,a} = \eta \cdot (Q_{s_t,a_t} - Q_{s,a}) \cdot exp\left(-\frac{|Q_{s_t,a_t} - Q_{s,a}|}{2\sigma^2}\right)$$

with the learning rate $\eta < \alpha$ and standard deviation σ. This update formula enforces the similarity of close Q-values and thus facilitates clustering.

4 Experiments

Since VRL and NRL are able to determine irrelevant parts of the state space with different degrees of detail, we use two different test scenarios. The basic scenario is a simple 10 × 10 gridworld which can be seen as a modified version of rooms gridworld [14] with just one room. The state space consists of a two-dimensional grid with torus boundary conditions and one goal state. The available actions are the four one-square-movements within the *von Neumann*-neighbourhood. One trial consists of initialising the system with the starting state and then choosing and performing actions until the goal is reached (*success*) or until 10,000 steps have been made without reaching the goal (*failure*). We perform a total number of 100 trials where a new starting point is selected randomly every 10 trials. The reward for reaching the goal state is 10, the reward for any other step is -0.1.

For VRL, we deliberately insert irrelevant dimensions by extending the state-space from (x, y) to (x, y, z) with (i) $z = y$, (ii) $z = \textit{random value}$ (iii) $z = y + \textit{noise} \in \{-1, 0, 1\}$ and (iv) $z = y + \textit{random walk}$ with steps $\in \{-1, 1\}$ so that different qualitative causes of irrelevance are tested. As an additional test, we set the extent of the goal state along the y-dimension to N_y so that there are two completely irrelevant dimensions. An exemplary graph of the results of the variance-based approach can be seen in figure 1. The variance of the irrelevant dimension is always considerably lower than the variance along the

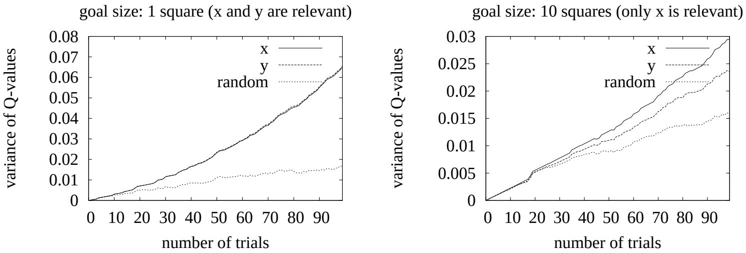

Fig. 1: Development of variances along the dimensions (x, y,$random$)

relevant dimensions even after few trials. This behaviour remains the same for different causes of irrelevance, as reported in Table 1. Even if two of the three dimensions are completely irrelevant, the algorithm results in a noteably lower variance for these dimensions. Problems arise only when the variance in the approximated Q-matrix is not caused by the difference in the real Q-values, but by random initialisation of the Q-values and too few updates of the quantities. As a solution to this problem, the Q-values can be weighed according to the frequency with which they have been visited.

NRL is capable of determining all equivalence classes, not only those along one of the dimensions. We use a simple gridworld with torus boundary conditions and one goal as testing scenario. Even in this simple case, there exist 10 equivalence classes characterised by the distance from the goal state $\in \{0, \ldots, 10\}$. The aim of NRL is to obtain Q-values that facilitate correct clustering of the state space into equivalence classes. The Q-values for the members of the ten equivalence classes for *standard* Q-learning are shown in figure 2 on the left. The variance of the Q-values belonging to each class is rather high with mean value 0.1609 after 50 trials and 0.0336 after 100 trials. In addition, the Q-values of different equivalence classes overlap, making a correct clustering of the state space into equivalence classes based on the Q-values difficult. Exemplary results of NRL, i.e. Q-learning *with neighbourhood cooperation* based on the Q-values are shown in figure 2 on the right. Here, the variance of the Q-values within the equivalence classes is lower with mean values 0.1184 after 50 trials and 0.0089 after 100 trials. In addition, the Q-values of different equivalence classes show no overlap, i.e. clustering of the Q-values into the equivalence classes is possi-

	goal size 1				goal size N_y			
	y=z	rand.	noise	r. walk	y=z	rand.	noise	r. walk
$\xi^x(50)$	*0.7983*	*0.0264*	*0.1580*	*0.0295*	*0.5791*	*0.0374*	*0.0745*	*0.0315*
$\xi^y(50)$	0	*0.0252*	*0.0791*	*0.0255*	0	0.0280	0.0115	0.0321
$\xi^z(50)$	0	0.0072	0.0214	0.0049	0	0.0020	0.0075	0.0065
$\xi^x(100)$	*0.6943*	*0.0680*	*0.454*	*0.0766*	*0.6511*	*0.0420*	*0.1880*	*0.0424*
$\xi^y(100)$	0	*0.0682*	*0.1978*	*0.0707*	0	0.0353	0.0493	0.0358
$\xi^z(100)$	0	0.0161	0.0289	0.0213	0	0.0052	0.0248	0.0078

Table 1: Variances after 50 and 100 trials in different scenarios. Relevant dimensions are emphasized.

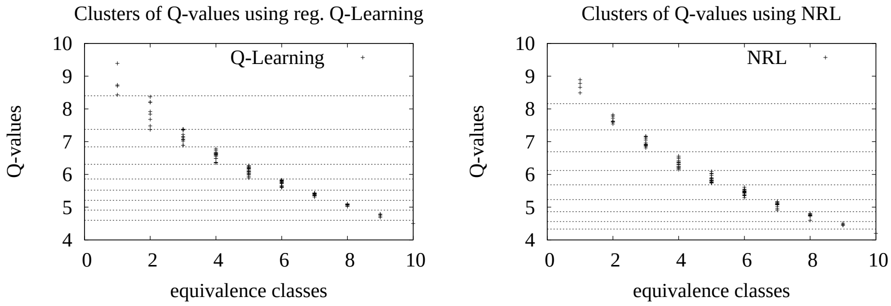

Fig. 2: Q-values within the equivalence classes for regular Q-learning and NRL. Dashed lines represent the borders between the clusters.

ble based on this information. To achieve these results, we applied regular Q-learning during the first 20 trials in order to have partly adapted Q-values, and we used NRL during the last 80 trials. σ was initialised with 1 and multiplicatively decreased by 0.9 per trial up to 0.1. Naturally, for standard Q-learning, similar equivalence classes also arise during training due to the convergence properties of Q-learning; however, a considerably larger number of trials is needed (around 160 trials), i.e. NRL speeds up convergence.

5 Discussion and future work

In this paper, we proposed two approaches for the potential problem of large search spaces and slow convergence of Q-learning: variance based relevance determination (VRL) of input dimensions and neighbourhood cooperation (NRL) of the Q-values. Both methods can be seen as an attempt to transfer the idea of equivalence classes induced by the reward and system dynamics as proposed by Ravindran to solely Q-matrix-based equivalence classes. Thus, our methods can be applied during training with no further knowledge of the reward function or system dynamics. Since VRL focusses on the relevance of input dimensions, it can only determine axes parallel equivalence classes but it nevertheless constitutes a robust and effective approach for the proposed test setting. In addition, it does not increase the necessary amount of training time – it can generalise from the information gathered by visiting only a part of the state-action space to determine the least relevant dimensions. NRL is more sophisticated since it boosts arbitrary cluster formation by means of neighbourhood cooperation. Note that this soft variant allows cluster formation without a potentially fatal exact identification of states in early stages of training. In our experiments, NRL showed much improved convergence speed compared to standard RL, though the training time for single steps is a bit larger. In contrast to Q-learning where each state needs to be visited several times, NRL requires only a few visits to compute feasible Q-values. The NRL-induced reduction of the search space to a few relevant parts could lead to a much faster convergence if combined with additional strategies. Unfortunately, NRL requires a correct parameter choice, including an appropriate choice of the neighbourhood factor σ and the exploration rate. Both

NRL and VRL result in a state space that is reduced to the minimal necessary equivalence classes under the restrictions put on the respective algorithm.

These first results are very promising, however, there remain a number of issues for further research: both algorithms are sensitive to the scaling of the search space and the reward function and scale invariant alternatives might be more robust. One simple possibility which we currently test is the substitution of the Q-values by their rank. So far, we restricted our experiments to determining the irrelevant dimensions without removing states from the system. This will be done in a second set of experiments. In this context, the question of suitable stopping criteria arises since the underlying model and optimum number of equivalence classes are not known. One possibility is a comparison of the measures to the value obtained for definite noise as a default-barrier for reduction. Additionally, there are some apparent variation possibilities for VRL and NRL which remain to be tested such as the entropy as an alternative to the variance measure. Last but not least both VRL and NRL remain to be applied to more complex scenarios within real-life applications.

References

[1] J. Morimoto and K. Doya. Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, 36:37–51, 2001.

[2] M. Riedmiller, A. Merke, D. Meier, A. Hoffmann, A. Sinner, O. Thate, and R. Ehrmann. Karlsruhe brainstormers — a reinforcement learning approach to robotic soccer. In P. Stone, T. Balch, and G. Kraetzschmar, editors, *RoboCup-2000: Robot Soccer World Cup IV*, pages 367–372. Springer, Berlin, 2001.

[3] C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

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

[5] D. P. Bertsekas and J. N. Tsisiklis. *Neuro-dynamic programming*. Athena Scientific, Belmont, MA, 1996.

[6] R. Schoknecht and M. Riedmiller. Reinforcement learning on explicitly specified time scales. *Neural Computing & Applications Journal*, 12(2):61–80, 2003.

[7] C. Atkeson and S. Schaal. Robot learning from demonstration. In *Proc. 14th ICML*, pages 12–20. Morgan Kaufmann, 1997.

[8] K. Tluk von Toschanowitz, B. Hammer, and H. Ritter. Mapping the design space of reinforcement learning problems – a case study. In H.-J. B H.-M. Groß, K. Debes, editor, *SOAVE 2004, 3rd Workshop on SelfOrganization of AdaptiVE Behavior, Fortschritts-Berichte VDI Reihe 10, Nr. 742*. VDI Verlag.

[9] C. Gaskett, D. Wettergreen, and A. Zelinsky. Q-learning in continuous state and action spaces. In *Australian Joint Conference on Artificial Intelligence*, pages 417–428, 1999.

[10] T. Dean and R. Givan. Model minimization in markov decision processes. In *AAAI/IAAI*, pages 106–111, 1997.

[11] B. Ravindran and A. G. Barto. Symmetries and model minimization of markov decision processes. Technical report 01-43, University of Massachusetts, Amherst, 2001.

[12] T. Martinetz and K. Schulten. A 'neural-gas' network learns topologies. *Artificial Neural Networks*, 1:397–402, 1991.

[13] F. Fernandez and D. Borrajo. VQQL. applying vector quantization to reinforcement learning. In *RoboCup*, pages 292–303, 1999.

[14] R.S. Sutton, D. Precup, and S. P. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999.