Multiple Task Specification Inspired from Mīmāṁsā for Reinforcement Learning Models

Bama Srinivasan, Ranjani Parthasarathi
Department of Information Science and Technology, CEG Campus, Anna University
bama@auist.net, rp@auist.net

Motivation

- In the current Reinforcement Learning models, agent perceives environment and performs action. The specification of action is predominantly single and rewards are directed from the environment.
- A 3-value logical formalism (MIRA) can be used for representing composite actions and these can be interpreted from the perspective of agent, rather than environment.

Mīmāṁsā

Mīmāṁsā, one of the Indian philosophies provides methods to interpret Vedic texts. One of the methods of interpretation is explained through three types of action performance (*karma*).

- Regular duty (*nitya karma*): Performance at all times.
- Occasional duty (*naimittika karma*): Performance on a specific occasion.
- Desired duty (*kāmya karma*): Performance for attaining an objective.

The performance of *nitya karma* and *naimittika karma* are specified in two ways.

- Performance of karma yields good results (*karanay abhyudhayam*)
- Non performance of karma yields bad results (*akaranay prathyavāya janakam*)

These aspects inspire towards the construction of a 3-valued formalism MIRA and the proposed reinforcement model with rewards directed from agent.

Conclusion

- The formalism of MIRA is introduced for specifying multiple tasks in Reinforcement Learning models.
- Multiple actions can be specified with this approach.
- Rewards marked from the agent leads to a reduced state-space.
- The model can facilitate goal-directed learning.

References

[1] Bama Srinivasan and Ranjani Parthasarathi. A formalism for Action Representation Inspired by Mīmāṁsā. Journal of Intelligent Systems, de Gruyter, 2012.

[2] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[3] Pandurangi. K. T. *Purvamimamsa from an Interdisciplinary Point of View, History of Science, Philosophy and Culture in Indian Civilization, Volume II Part 6*. Motilal Banarsidass, Delhi, India, 2006.

[4] Pujyasri Sri Chandrasekharendra Saraswati. *The Vedas*. Bhavan's Book University, Mumbai - 400 025, India, 2009.

[5] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Teaching Multiple Tasks to an RL Agent using LTL. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Stockholm, 452-461, 2018.

[6] Bama Srinivasan and Ranjani Parthasarathi. A formalism to specify unambiguous instructions inspired by Mīmāṁsā in computational settings. Winner of Bimal Krishna Matilal Logic Prize, To be published in Logica Universalis, 2021.

Mīmāṁsā Inspired Representation of Actions (MIRA)

Syntax: A 3-valued formalism is given by $\mathcal{L}_i = (I, R, P, B)$, where I = $(I^v \cup I^n)$
$I^v = \{i_1^+, i_2^+, \cdots, i_n^+\}$ - Positive Imperatives, $I^n = \{i_1^{'}, i_2^{'}, \cdots, i_n^{'}\}$ - Negative Imperatives
$R = \{r_1, r_2, \cdots, r_m\}$ -Reasons, $P = \{p_1, p_2, \cdots, p_l\}$ - Purposes - these are from proposition logic
$B = \{\wedge, \oplus, \rightarrow_r, \rightarrow_i, \rightarrow_p\}$ - Binary connectives
The formula of imperatives $\mathcal{F}_i$ is given by:

$$\mathcal{F}_i = i|(i \rightarrow_p p)|(i \rightarrow_p p_1) \wedge (j \rightarrow_p p_2)|(i \rightarrow_p \theta) \oplus (j \rightarrow_p \theta)|(\varphi \rightarrow_i \psi)|(\tau \rightarrow_r \varphi)$$

where $i, j \in I$, $p_1, p_2, \theta \in P, \tau \in R, \varphi, \psi \in \mathcal{F}_i$.
Semantics: Indicates action performance. Evaluation is given by:

$$\mathcal{E}(\varphi) = \{S, V, N\}$$
$$\mathcal{E}(\tau) = \mathcal{E}(\theta) = \{\top, \bot\}$$

Use of binary connectives B on $\{S, V, N, \top, \bot\}$ lead to the output values $\{S, V, N\}$.

Task Evaluation in terms of rewards inspired from Mīmāṁsā

The model of agent and environment can be adapted across MIRA formalism, where the environment maps to the sets R and P and the action from agent correspond to I. The agent and environment interaction according to the reinforcement learning model is slightly tweaked to address the formalism of MIRA.

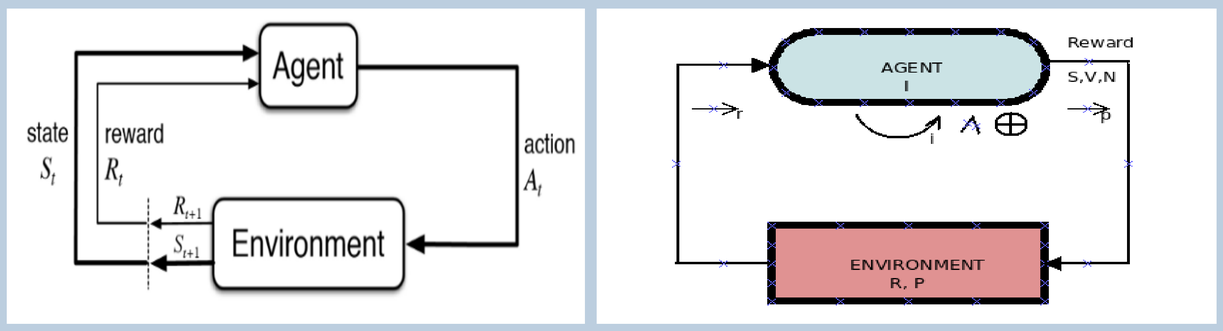

Figure 1: Present RL model; Proposed RL model according to MIRA formalism

Rewards in the prevalent model are from environment. In LTL based approach [5], the reward function is introduced such that if the sequence of states $s_0, s_1, s_2, ...s_n$ attain a *true* value ($\sigma_{0:n} \models \varphi$) then reward is 1. Else, the value of reward is maintained at 0.

$$R_\varphi(< s_0, s_1, ..., s_n >) = \begin{cases} 1 & \sigma_{0:n-1} \not\models \varphi \text{ and } \sigma_{0:n} \models \varphi \\ 0 & \text{otherwise} \end{cases}$$

But through our proposed model, the evaluation of actions in terms of S, V and N can be mapped towards rewards such that if the tasks performed evaluate to S, then it is marked against reward. This evaluation occurs at two levels. In the first level, $\mathcal{F}_i$ is evaluated to $\{S, V, N\}$. With these values, the reward is evaluated to $\{1, 0\}$ in the second level.

$$R_\psi = \begin{cases} 1, & \text{if } \mathcal{E}(\psi) = S \\ 0, & \text{otherwise} \end{cases}$$

where ψ denotes the tasks represented through MIRA formalism.
Example:
The robot can navigate through the doors and move from one room to another. Assuming the agent in room 2, to reach the part5, the specification according to MIRA formalism can be given as:

$$(room2 \rightarrow_r (goto3 \rightarrow_i ((goto1 \rightarrow_p room5) \oplus (goto4 \rightarrow_p part5))))$$

The valuation can be done in such a way that the ultimate goal of the agent is to reach $part5$ from $room2$. If the agent proceeds through room4, then the evaluation results in S.

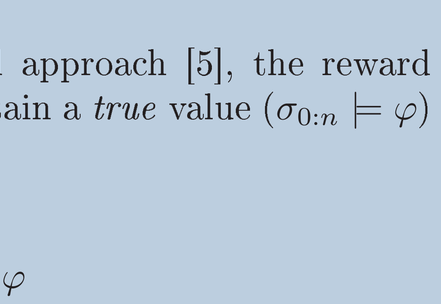

Figure 2: Image of a building

$$(\top \rightarrow_r (S \rightarrow_i ((V \oplus S))) = S$$

Based on the evaluation leading to S, the reward is designated as 1 in the proposed approach.
If the agent picks up the path of $room2 \rightarrow room3 \rightarrow room4 \rightarrow room0$, the evaluation leads to N indicating the agent cannot reach the goal of $part5$).

$$room2 \rightarrow_r (goto3 \rightarrow_i (goto4 \rightarrow_i (goto0 \rightarrow_p room0)))$$
$$evaluates\ to\ \top \rightarrow_r (S \rightarrow_i (S \rightarrow_i (S \rightarrow_p N))) = N$$

Thus the value of N can help the agent to orient towards goal.