R14/Sarthak Mittal/200050129

March 29, 2023

This paper introduces the concept of **Rule Transfer**, a "transfer algorithm" where the **rules** for a **source** policy are learnt and then applied to improve learning in a **target** task. In past works, the learnt advice has been used for initialization in discrete domains or as soft constraints. There have also been attempts to learn the task relationships. Using three metrics, the authors link a source task (**gridworld** domain) to the 3 v/s 2 Keepaway task (**robot soccer** domain).

The main goal of Transfer Learning (TL) in a Reinforcement Learning (RL) setting is to **reduce training time** using knowledge from a source task. The past work that explored TL in RL dealt with source and target tasks which were **closely related** (differing in reward, actions, and/or state variables, but within the same domain). The authors claim that a more powerful way to simplify a task would be to **abstract** it into a different domain. If the relationship between the source and target tasks is known then, Cross-Domain Transfer (CDT) can improve the learning speed.

Rule Transfer consists of the following steps: (1) learn a **policy** in the **source** task (2) learn a **decision list** summarizing the source policy (3) **modify** decision list for the target task (4) use decision list to learn a **policy** in the **target** task. The rules serve as an **intermediate representation** which are abstracted from the source and leveraged by the target, thus allowing them to have different internal representations. The **translation procedure** maps the actions and state variables of the source to the **most likely** action and state variables of the target. The decision list is considered as **advice** and the agent **refines** its decisions using **experience** and **utilization schemes**. The schemes are: (1) **Value Bonus** (depending on the decision list, the **Q-value** of recommended actions gets a **"bump"**) (2) **Extra Action** (the agent executes the recommended action on choosing this **pseudo-action**, but over time the **bias** can be reduced) (3) **Extra Variable** (the agent learns the **importance** of the recommended action, but could learn to **ignore** the state if sub-optimal).

The main tasks considered in the paper are: **Keepaway (K)** (stochastic actions, noisy observations, continuous state space, 3 keepers v/s 2 takers, keepers get a reward for retaining the ball, keepers learn using Sarsa) with radial basis function approximator, **Ringworld (R)** (similar state variables, similar distances, grid field, player gets a reward as long as the opponent has not caught it, player can choose to stay put or run to a target, randomness is the proba-

bility of capture, player learns using Sarsa along with a Q-table) and **Knight Joust (J)** (similar to Ringworld, players alternate moves, player has to reach the opposite end, player can move or jump, the opponent has a fixed stochastic policy, player uses Sarsa along with a Q-table).

To determine the effective utilization of rules, the authors make use of three metrics: (1) **initial performance** (average hold time at $t = 0$) (2) **asymptotic performance** (average hold time after learning plateau) (3) **accumulated reward** (sum of reward accumulated at every hour). They run the analysis using the pairs **(K, K)**, **(R, K)** and **(J, K)** as the source and target tasks and obtained significant **improvements** for all metrics in the second analysis and one of the metrics in the third. The results of the second analysis indicated that all three utilization methods could significantly increase all three transfer metrics, and the **Extra Action** method was slightly superior. A few experiments were also conducted to analyze the **robustness** of RT with variations in the amount of data and the parameters/settings. From the third analysis, they concluded that **Knight Joust** significantly improved the **initial performance**.

The authors also discuss a method using which **cross-domain mappings** could be learnt using some knowledge about **qualitative** characteristics of the source domain. For the Ringworld task, they make the following **observations**: (1) an opponent approaches the player (**distance** to opponent state variable) (2) the opponent will continue to move **towards** the player (3) run(near) takes **longer** to complete than run(far) (4) a sense of **"identity"** which can help the learner distinguish objects. Keeping these in mind, the following **procedure** assists in the construction of a mapping: (1) identify the state variables that are **near 0 at the end** (2) identify the action that causes target task variables to decrease **most consistently** (3) identify state and action **corresponding** to distance and run respectively (the **observations** crucially help in this) (4) identify the **distances** and **angles**.

The authors claim that the observations about the source task are as important as the methods that use them. When the mapping is impossible due to **noise**, the **violation** of observations can be detected. The paper thus shows that **inter-domain transfer** could also prove beneficial. Though the paper shows the **flexibility** of transfer, the transfer between more distinct internal representations could be explored. When both the source and target have a **finite** number of states, the transfer could be executed without using rules as an intermediary.