Autonomous Extraction of a Hierarchical Structure of Tasks in Reinforcement Learning, A Sequential Associate Rule Mining Approach

Behzad Ghazanfari†, Fatemeh Afghah†, Matthew E. Taylor‡

beghazanfari@gmail.com, Fatemeh.Afghah@nau.edu
† School of Informatics, Computing, and Cyber Security, Northern Arizona University
taylorm@eecs.wsu.edu
‡ School of Electrical Engineering and Computer Science, Washington State University

Abstract
Reinforcement learning (RL) techniques, while often powerful, can suffer from slow learning speeds, particularly in high dimensional spaces. Decomposition of tasks into a hierarchical structure holds the potential to significantly speed up learning, generalization, and transfer learning. However, the current task decomposition techniques often rely on high-level knowledge provided by an expert (e.g. using dynamic Bayesian networks) to extract a hierarchical task structure; which is not necessarily available in autonomous systems. In this paper, we propose a novel method based on Sequential Association Rule Mining that can extract Hierarchical Structure of Tasks in Reinforcement Learning (SARM-HSTRL) in an autonomous manner for both Markov decision processes (MDPs) and factored MDPs. The proposed method leverages association rule mining to discover the causal and temporal relationships among states in different trajectories, and extracts a task hierarchy that captures these relationships among sub-goals as termination conditions of different sub-tasks. We prove that the extracted hierarchical policy offers a hierarchically optimal policy in MDPs and factored MDPs. It should be noted that SARM-HSTRL extracts this hierarchical optimal policy without having dynamic Bayesian networks in scenarios with a single task trajectory and also with multiple tasks’ trajectories. Furthermore, it has been theoretically and empirically shown that the extracted hierarchical task structure is consistent with trajectories and provides the most efficient, reliable, and compact structure under appropriate assumptions. The numerical results compare the performance of the proposed SARM-HSTRL method with conventional HRL algorithms in terms of the accuracy in detecting the sub-goals, the validity of the extracted hierarchies, and the speed of learning in several testbeds.

Introduction
Reinforcement learning is known as a commonly used approach for planning and sequential decision making in artificial intelligent (AI) systems, where the agents gradually learn and optimize their actions from delayed rewards through a trial-and-error mechanism. However, one of the main challenges of RL approaches is scalability to high-dimensional state spaces (Barto and Mahadevan 2003). Hierarchical reinforcement learning (HRL) methods are known to reduce the computational complexity of RL approaches by temporal and state abstraction in the form of decomposing the learning problem to a hierarchy of several sub-problems. Sub-goals refer to the local target states that not only provide easy access or high reinforcement gradients, but also must be visited frequently (McGovern and Barto 2002; Stolle 2004). These sub-goals can help an agent to accelerate the learning process, particularly in high dimensional spaces. In (Dietterich 2000), a HRL decomposition method called MAXQ is proposed based on the assumption of having an expert with the knowledge of sub-goals to provide a correct hierarchy, however such assumption can restrict the application of this method in autonomous systems where a limited expert’s understanding is available (Taylor and Stone 2009).

In the absence of an expert, several HRL techniques have been reported for task decomposition, in which a number of sub-goals that are correlated with the successful policies are utilized as the required states to decompose the learning task (Digney, 1998; McGovern and Barto, 2002; Stolle, 2004). However, extracting these states in an autonomous manner is still a challenging problem (Chiu and Soo, 2011). More importantly, in the majority of existing HRL methods, the potential hidden correlations among these sub-goals to achieve the ultimate goal have been overlooked. In general, the current HRL methods with autonomous task decomposition capability can be divided into two groups depending on their domains.

HRL methods in MDPs: The HRL methods based on extracting the sub-goals (McGovern and Barto 2002; Stolle 2004), or the ones based on bottlenecks extraction (Mannor et al. 2004; Simsek and Barto 2004; 2009) can only extract a flat hierarchy, (i.e., one level) which means that these methods only find the sub-goals or the bottlenecks, rather than a hierarchical structure of those. Since these methods often use the paths or the sub-graphs of the agent, or the shortest paths among the nodes of a graph to calculate their required measures such as betweenness (Simsek and Barto 2009), their performance considerably degrades in scenarios with a large state space, or when the number of actions to reach the goal-states increases. They also usually require prior knowledge about the measures that helps to partition state space to parts that are connected densely inside, but sparsely to each other (Ghazanfari and Mozayani 2016).

HRL methods in Factored MDPs (FMDPs): Some of the current HRL methods based on extracting the task-dependent hierarchy in FMDPs include HEX-Q (Hengst 2003), VISA (Jonsson and Barto 2006), and HI-MAT (Mehta et al. 2008; 2011). Since there are implicit struc-
tions of the problems among states variable in factored MDPs, dynamic Bayesian networks (DBNs) as high-level sources of pre-knowledge are often utilized to decompose the tasks in such processes noting their capability to extract the impact of each action on state variables. HI-MAT and VISA algorithms rely on availability of DBNs for each action [Jonsson and Barto 2006; Mehta et al. 2008 2011]. Since VISA considers the impacts of all actions regardless of the domain, it can create unnecessary branches in the extracted hierarchy or unnecessary subtasks. Thus, it may result in an “exponentially sized hierarchy” that limits its application in some domains [Mehta et al. 2008 2011]. To address this problem, HI-MAT was proposed to remove such unsuccessful and redundant actions cycles. This method leverages a single and carefully constructed trajectory to construct a MAXQ hierarchy. It is shown that the constructed hierarchy is compact and comparable to manually engineered ones. However, the main disadvantage of both these methods is that utilizing DBNs require high-level knowledge that should be provided by an expert or needs to be extracted via a large number of computations [Wynkoop and Dietterich 2008]. Among these HRL methods proposed for factored MDPs, HEX-Q is the only one that does not rely on DBNs. However, this method is not capable of identifying the relations among the states variables that can potentially result in divergence of the learning process [Mehta et al. 2008]. Bacon, Harb, and Precup (2017) used a policy gradient method to create temporally extended actions instead of extracting the sub-goals. However, this method can only handle one task at the time and needs to know the number of options in advance; therefore, it may have limited application in multi-task RL, or in cases with a large number of subtasks.

The key contribution of this work is to propose a HRL method based on the idea of sequential association rule mining (SARM) that extracts a hierarchical knowledge from the hidden correlations among the extracted sub-goals and use this knowledge to decompose the tasks to multiple sub-tasks. Conventional subgoal extraction methods that can work in MDPs, do not extract a hierarchical task structure. The few existing hierarchal structure extractor methods in RL including HEX-Q, HI-MAT, and VISA only work in factored MDPs. More importantly, HI-MAT and VISA rely on DBNs knowledge, which is a high-level supplementary knowledge provided by human experts. HI-MAT, the most recent HRL approach in the literature has the following limitations: 1) requiring DBNs knowledge, 2) cannot work based on several trajectories that are a typical situation in RL, 3) cannot support funnel property of a subtask, and 4) cannot be applied in MDPs. However, our proposed SARM-HSTRL method extracts a hierarchal optimum policy task structure for both MDPs and factored MDPs, while it does not rely on DBNs as a preknowledge structure provided by human experts in MDPs. To the best of our knowledge, SARM-HSTRL is also the first method that can extract the hierarchal optimum policy task structure of multiple policies.

An Overview on Association Rule Mining

Association rule mining (ARM) methods use a combination of two key measures of support and confidence in a proven efficient extraction strategy to obtain and evaluate the most efficient and reliable relationships among the variables in a dataset. ARM has been applied in bioinformatics to discover the patterns in datasets that are statistically important [Bebek and Yang 2007], or in retail stores to find the items that are commonly being sold together among millions of transactions [Lin, Alvarez, and Ruiz 2002; Tan, Steinbach, and Kumar 2006]. An ARM problem is defined by a pair of (ITEMSET, Transaction), where ITEMSET = \{i_1, \ldots, i_g\} is the set of all items and Transaction = \{\Omega_1, \ldots, \Omega_N\} is the set of all transactions. Each transaction is a subset of items of ITEMSET. The relationship among the items in the transaction set can be defined by an association rule. An association rule is expressed in the form of \(A \rightarrow B\), where \(A\) and \(B\) are disjoint sets of items; \(A \cap B = \emptyset\). The frequency of the occurrence of \(A\) and \(B\) together in a Transaction is defined as a key factor, also known as support of the association rule. The frequency of occurrence of \(A\) and \(B\), relative to the frequency of the occurrence of \(A\), is known as confidence. The definition of support and confidence are as follows [Tan, Steinbach, and Kumar 2006]:

\[
support(A \rightarrow B) = \frac{\sigma(A \cup B)}{N}
\]

\[
confidence(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}
\]

where \(\sigma(\cdot)\) is the number of observed transactions including the elements inside of the parenthesis, and \(N\) is the total number of transactions. The support factor is often used as a measure to disregard the items that do not occur together so frequently relative to \(N\), and confidence can express the reliability of the extracted rule. The corresponding thresholds for support and confidence, known as \(\text{minsup}\) and \(\text{minconf}\), respectively, can be used to extract important rules [Tan, Steinbach, and Kumar 2006]. ARM algorithms typically consist of two parts: 1) Frequent Itemset Generation. All of the itemsets that satisfy the \(\text{minsup}\) condition are extracted, i.e., frequent item sets. 2) Rule Generation: Building upon the outputs of the Frequent Itemset Generation, this step calculates the confidence of the obtained frequent itemsets and checks their eligibility by comparing their confidences with \(\text{minconf}\) threshold. The frequent pattern growth (FP-growth) algorithm has been proposed for Frequent Itemset Generation by constructing a compact data structure, called a FP-tree. The confidence value is calculated for each of the rules and evaluated based on \(\text{minconf}\). This algorithm outperforms the majority of frequent patterns extraction algorithms in large datasets; for the analysis of time complexity and more details about FP-growth algorithm see [Kosters, Pijs, and Popova 2003; Tan, Steinbach, and Kumar 2006].

Proposed SARM-HSTRL Algorithm

Here, we propose an algorithm to extract a hierarchal structure of tasks in RL named SARM-HSTRL that works in both MDPs and factored MDPs. To the best of our knowledge, despite all of the existing HRL methods in MDPs, SARM-HSTRL extracts a hierarchal abstraction, not a flat abstraction, in MDPs. In continue, we define some of the terms used
throughout the paper. An overview of MDPs and FMDPs, and additional details on notations and definitions are presented in the Background subsection of the “supplementary material” section in the end of document.

**Definition 1:** To assign a unique representation to a set of multiple state variables in FMDPs, here we define a reversible coder-decoder operation. A map function, $MF$, as a coder maps the state variables in FMDPs to one variable. $MF^{-1}$ as a decoder is the reverse process of retrieving the FMDPs’ state variables from just that single value, $L$, as described in Algorithm 1. It means $MF(x_1, x_2, \ldots, x_n) = R_{x_1} \times \prod_{i=2}^{n} numD_i$, should be a surjective, injective, and invertible function, where $D_i$ refers to set of possible values for each state variable, $numD_i$ denotes the number of possible values in $D_i$, and $R_{x_i}$ shows the index of $x_i$ in $numD_i$. Therefore, $\prod_{i=1}^{n} numD_i$ is the total possible number for $X$.

**Definition 2:** Transitions are considered unpredictable when they lead to entering or leaving subgoals. The region and the boundaries among states’ clusters that have unpredictable transitions are considered as exits and defined by a state action pair $G_i = (sT_i, a)$ when taking action $a$, as a primitive action, from state $sT_i$, as a subgoal, leads to the resultant state that is a goal state to complete subtask $T_i$ (Hengst, 2003). This concept has been further explained with an example in the “supplementary material”.

**Definition 3:** A task hierarchy $H$ is generally shown as a tree, or a directed acyclic graph, $(T, E)$, in which the root as the main task, $T_0$, is decomposed to other subtasks $T_1, \ldots, T_n$ and the edges, $E$, represent the relation among them. A subtask, $T_i$, is a semi-MDP (SMDP) that is shown by $(X_i, S_i, G_i, C_i)$ (Mehta et al., 2011), where $X_i$ is the set of variables that their corresponding values change during performing the subtask, $S_i$ denotes the set of admissible states of $T_i$, $G_i$ shows the exits of corresponding subtasks as termination conditions of $T_i$, and $C_i$ is the set of child tasks of $T_i$. Child tasks, $C_i$, can be formed based on different HRL frameworks such as MAXQ or option.

In the task hierarchy graph, leaf nodes correspond to subtasks that interact with the environment directly by applying primitive actions, $A$, to states, $S_i$. Other nodes of the task hierarchy include subtasks as abstract states and their corresponding local policies as abstract actions. We should note that the subtasks are defined over the extracted regions as the policies that lead to leaving these regions via exits. These definitions guarantee that no action can lead to leaving a subtask except via its exits. Each subtask similar to a region includes a set of states, actions, Markov transitions, and reward functions.

The proposed SARM-HSTRL decomposes the tasks into multiple subtasks by extracting the subgoals as subtask’s termination, exits, are the task graph’s nodes and their relations, in the form of association rules, are the edges of the graph. The state space is partitioned recursively in a top-down manner, and the state abstraction and corresponding options exit, temporal abstraction, are formed for these partitions. The state abstraction and temporal abstraction can limit the policy search space that leads to increasing the speed of learning.

Algorithm 1 $MF^{-1}(L)$

$$\text{for } i = n : 1 \text{ do}$$
$$\text{TEMP} = \prod_{i=1}^{n} numD_i$$
$$Rx_i = \frac{L}{\text{TEMP}}, L = \text{mod}(L, \text{TEMP})$$
$$\text{if } L == 0 \text{ then}$$
$$Rx_i = Rx_i - 1, L = \text{TEMP}$$
$$\text{end if}$$
$$\text{end for}$$

The proposed SARM-HSTRL is composed of two phases (see Algorithm 2). In the first phase, several association rules are extracted using an SARM approach following the two steps of i) Frequent Itemsets Generation, and ii) Rule Generation procedure. Then, the proposed HST-construction method converts these association rules to a hierarchical structure tree.

Algorithm 2 SARM-HSTRL

**Input:** Transition, minsup, minconf

**Output:** HST

1. Frequent Itemset = FP-growth (Transition, minsup)
2. Association Rules = Rule Generation (Frequent Itemset, minconf)
3. HST-construction (Association Rules) // See Algorithm 3

In the proposed SARM-HSTRL, each trajectory of visited states, $\Omega_k = \{s_1, \ldots, s_h\}$, is considered as a transaction member of the Transaction set, in which $h$ shows the number of states in that transaction. In FMDPs, the proposed function $MF$ (see Definition 1) is used to map multivariate state variable to a univariate state variable. All visited states in successful trajectories are stored in the ITEMSET. $T_i$ are defined based on sub-goal states. These sub-goals, as exit states, are defined as the states that are frequently visited in successful trajectories (i.e., the trajectories where the agent reaches a goal state). In other words, the problem of finding the sub-goals and the relations among them can be seen as extracting association rules such as $\{s_d, \ldots, s_g \rightarrow s_h\}$, where $\{s_d, \ldots, s_g, s_h\}$ are sub-goal states. It should be noted that there often exists a set of some key subgoals that are common among different tasks, and the proposed SARM-HSTRL method can extract such key subgoals by processing a set of trajectories of tasks with random start and goal states.

Here, we use the FP-growth algorithm to perform the first step of SARM called Frequent Itemset Generation. If the minsup is set to its maximum possible value (i.e., one), the sub-goals must be visited in each trajectory of each transaction. If we set a very small value to the minsup, the performance of FP-growth will be degraded as SARM-HSTRL may provide some false-positive itemsets for the evaluation of Rule Generation. Hence, we face a trade-off in select-

Footnotes:

1 A example of applying SARM-HSTRL on a testbed along with a detailed description of the SARM-HSTRL process notations is presented in the “supplementary material” section.

2 A successful trajectory is defined as a trajectory of states that leads to the goal reward (Mehta et al., 2008).
ing reasonable values for \( \text{minsup} \) and \( \text{minconf} \). On one hand, these values should be small enough to capture different sub-goals and relations in RL domains with multiple types of successful trajectories. On the other hand, if the \( \text{minsup} \) and \( \text{minconf} \) are set to very small values, the extracted hierarchical structure would extract some unnecessary sub-goals and relations. The proper range of these parameters can be set based on the number of trajectories of encountered tasks.

Algorithm 3 HST-construction : Construct a tree, \( T \), with one node that is the root node, \( R \).

1: **Input**: \( \text{AR-set} \) is the set of association rules. \( \text{AR-set} = \{ \text{AR}_1, \ldots, \text{AR}_{\text{NumRules}} \} \)
2: **Output**: \( \text{HST} \)
3: \( \text{num} \) : the number of children of the Parent-Node; \( PN_i \) : the \( t_{th} \) child of the Parent-Node
4: for \( i = 1 \) : \( \text{NumRules} \) do
5: Parent-Node=\( R \)
6: for \( j = 1 \) : \( \text{Len} \) do
7: \( t = 1, \text{FlagM} = 0 \)
8: repeat
9: if \( \text{AR}_{ij} == PN_i \) then
10: Parent-Node=\( PN_i \); \( \text{FlagM} = 1 \)
11: end if
12: \( t = t + 1 \)
13: until \( t <= \text{num} \) and \( \text{FlagM} = 0 \)
14: if \( \text{FlagM} == 0 \) then
15: create a new child Node in the Parent-Node: \( PN_{\text{num} + 1} = \text{AR}_{ij} \)
16: Parent-Node=\( PN_{\text{num} + 1} \)
17: end if
18: end for
19: end for

Next, the Rule Generation procedure is performed on the extracted frequent itemsets as the output of FP-growth algorithm. Recall that a confidence value is the conditional probability of the occurrence of a consequent of a certain rule when its premise has been seen, and are calculated using \( \text{minconf} \) thresholds. The confidence value of each association rule can be used as a priority score to choose among corresponding temporally extended actions of association rules.

Besides extracting a set of sub-goals as the association rules, \( \text{SARM-HSTRL} \) also extracts different possible sequences of these sub-goals for HST construction in a sequential association rule mining procedure. The value of \( t \), time of each sub-goal in each trajectory, can be compared to create a sequence of observed sub-goals. Each sequence shows the relationship among the sub-goals in a flat manner of one association rule. For instance, there are two trajectories of four sub-goals \( a, b, c \rightarrow d \) and \( b, a, c \rightarrow d \), the \( t \)'s values of \( a \) and \( b \) are \( \{2, 1\} \) and \( \{2, 1\} \), respectively, in the trajectories. If the frequencies of those orders are the same, it means that the order of visiting \( a \) and \( b \) is not important to achieve the consequent subgoal although each sequence could have different probability values. Algorithm 3 describes the HST-construction method that makes the hierarchical structure of tasks. The HST helps an agent to choose the correct sub-tasks. Each association rule \( \text{AR}_i \) can be shown in the form of \( \text{AR}_i = s_{t_i}, \ldots, s_{(t_n + n)i} \rightarrow \text{...} \)

Figure 1: (a) The first task hierarchy of the first testbed, Figure 1(c) for experiment 1. (b) The second task hierarchy of the first testbed, Figure 1(c) for experiment 2. (c) The first testbed: the size of the maze is \( 22 \times 22 \) and it has \( 7 \) sub-goals. The subgoals are colored with yellow.

\( s_{(t+n+1)i} \), where \( \{ s_{t_1}, \ldots, s_{(t_n + n)i}\} \) denotes a sequence of sub-goals of the \( \text{AR}_i \). In this algorithm, \( \text{Len}_i \) denotes the number of items in \( \text{AR}_i \). The number of elements of the premise of the \( \text{AR}_i \) is \( n + 1 \), and the number of elements of the consequence of each \( \text{AR} \) is \( 1 \); thus, the \( \text{Len}_i \) is \( n + 2 \). \( \text{AR}_{i,j} \) is the \( j \)th element from the end of \( \text{AR}_{i,j} \). For example, \( \text{AR}_{1,2} \) is \( s_{t+n} \) and \( \text{AR}_{1,\text{len}_i} \) is \( s_1 \). \( \text{NumRules} \) is the number of association rules.

**Theoretical Analysis**

In this section, we provide a theoretical analysis to study the properties of the extracted hierarchical structure of the tasks using the proposed \( \text{SARM-HSTRL} \) method. The problem of extracting a hierarchical structure in RL can be considered as a hierarchical credit assignment problem of MAXQ, and the proposed \( \text{SARM-HSTRL} \) provides a solution to automatically perform such extraction in model free MDPs and FMDPs. Since the required convergence conditions of \( \text{SARM-HSTRL} \) are different in MDPs and FMDPs, its performance has been evaluated in each domain separately.

**Theorem 1**: The proposed \( \text{SARM-HSTRL} \) converges to a hierarchical optimum policy in model-free MDPs. A hierarchical policy is defined as an assignment of a local policy to each subtask. A hierarchical optimum policy is a hierarchical policy that makes the best accumulated reward \( \text{[Mehta et al., 2011]} \).

**Proof**: Using Definition 2 in \( \text{Hengst, 2003} \), the proof that the extracted HST leads to a hierarchical optimal policy is straightforward if we show that only one variable
of X changes as a result of each action; because then the Q function of such hierarchy can be recursively expanded and mapped to a Q function of a flat MDP. In our proposed method in MDPs, each node of HST corresponds to a sub-MDP and based on Definition 1, there exists only one variable in our state variable, X. Therefore, as proved for HEX-Q algorithm, the proposed SARM-HSTRL converges to a hierarchical optimum policy. Hierarchical execution can be applied by using a decomposed value function since the proposed method similar to MAXQ breaks down the MDP to interlinked sub-MDPs directly. Q function for each node, exit, as a sub-MDP of the tree is defined recursively as follows:

\[ Q^*_{T_i}(s^T_i, a) = \sum_{s'} P_{s'n's}^a [R_{s'n's}^a + V^*_{T_i}(s')] \]

where \( s' \) is the hierarchical next state. \( Q^*_{T_i}(s^T_i, a) \) shows the expected value of node \( T_i \) after performing (abstract) action \( a \) in (abstract) state \( s'^T \); and in the continue pursuing the optimal hierarchical policy. \( V^*_{T_i}(s) \) is the decomposition of optimal hierarchical value function that is calculated recursively as follow:

\[ V^*_{T_i}(s) = \max_a [V^*_{C_i(a)}(s) + Q^*_{T_i}(s^T_i, a)] \]

where \( V^*_{C_i(a)} \) shows the child of \( T_i \) implementing action \( a \). \( \Box \)

In continue, we study the convergence of the proposed method in FMDPs. For FMDPs, if the state abstraction and temporally extended actions are constructed based on one state variable in each layer, then proof 1 is still valid (as shown in HEX-Q). However, the assumed condition in HEX-Q of only having one state variable for FMDPs is not a practical assumption; therefore here we evaluate the optimality of the SARM-HSTRL’s solution for a general case. Basically, there is not a straightforward proof for convergence of methods that extract the hierarchical structure of tasks in FMDPs (Jonsson and Barto 2006; Mehta et al. 2011). It is proven in (Dean and Givan 1997) that having the stochastic substitution and reward respecting characteristics preserves optimality for reduced MDPs such as FMDPs. Thus, stochastic substitution and reward substitution can be used to prove optimality by showing that each reduced MDP has the mentioned characteristics. Next we review such characteristics of the proposed method.

**Definition 4:** Transaction, \( \Omega \), a set of extracted trajectories, is called **representative** if \( \Omega \) includes all possible state action pairs that lead to the ultimate goals.

In SARM-HSTRL, the trajectories are used instead of high-level sources of knowledge (e.g. DBNs). Since DBNs show casual relation among state variables for each action; the HRL models based on DBNs can present irrelevant states variables in state abstraction. More importantly, as we mentioned earlier, the assumption of having DBNs in advance is not practical in autonomous settings. Our proposed method solve this problem by extracting the relations among the states and state abstraction in an autonomous manner, where the trajectories are the only source of knowledge to show the effects of actions on state variables.

**Theorem 3:** The hierarchical structure of tasks being extracted by SARM-HSTRL, \( H \), of a representative \( \Omega \) guarantees that “the total expected reward during each trajectory of \( \Omega \) is only a function of the values of \( X \) in the starting state of \( \Omega \)” (Mehta et al. 2011) for any trajectory-task \( \langle \Omega_k, T_i \rangle \) that is consistent with \( H \). Also, there is just one hierarchical structure of tasks that can be extracted based on extracted exit states, subgoals, that is...
safe with respect to $\Omega$.

**Proof:** SARM-HSTRL constructs $T_i = (X_i, S_i, C_i, G_i)$ directly based on subgoals, exit states, of several trajectories not DBNs. The subgoals are used to partition the sequence of states of trajectories, $\Omega$. The actions in any sequence of state-action pairs of each trajectory are primitive and change the values of $X_i$ as their resultant states are in the same partition- except exit states as termination conditions, $G_i$. If it changes the variables outside of current sub-task, $T_i$, that variable, $X_k$, should appear in the sequence of state-actions pairs before exit states’ variables of $T_i$. Thus, it will be placed inside of sub-task $T_j$ what is a contradiction with the assumption that it can have effects on variables more than $X_i$ that are outside of current sub-task. In the same way, we can say that all immediate rewards in the trajectory are functions of the variables in $X_i$. Therefore, the summation of discounted rewards and the probability of transition in each trajectory are just related to $X_i$; thus, the extracted hierarchical structure of tasks is safe with respect to $\Omega$. Since the proposed method form the sub-tasks of subgoals all in once; thus, if there is another hierarchy, $H'$, as it is consistent with $\Omega$, it will violate the safe characteristic with respect to $\Omega$. This completes the prove that the extracted hierarchy by SARM-HSTRL leads to the hierarchical optimal policy. □

**Theorem 4:** The extracted hierarchical structure of tasks using SARM-HSTRL provides the most efficient, reliable, and compact hierarchical structure considering the representative and non-redundant set of trajectories, $\Omega$, when the problem is sparse in both of MDPs and FMDPs. Efficiency is measured by the probability of usage and the resultant performance. Reliability is a function of the accuracy for the certainty of occurrence of next sub-tasks depending on which sub-tasks have been done so far. Resultant performance captures the compactness concept and is defined as how much the extracted structure abstract action space. Thus, efficient subtasks are considered as subtasks that summarized the longest frequent sequence of actions in temporally extended actions.

**Proof:** Sub-tasks are extracted based on support and confidence measures in the form of association rules as the most efficient and reliable sequence of subtasks, exit states, among representative and non-redundant trajectories, $\Omega$. The support measure checks the ratio of witnessing all possible subtasks to all observations in $\Omega$. Thus, it finds the subtasks that happen with the highest probability related to other ones. In other word, these subtasks are the best summarization, longest and most frequent, of what happened in past. Reliability implies providing the highest accuracy of predicting next sub-tasks based on summarization of several trajectories and what have been done so far. The confidence measure evaluates every possible sequence by constructing a tree considering all possible eligible sequences among several trajectories. It preserves their sequences and compacts the extracted subtasks in form of sequential association rules by matching and mapping them from the last task to the first ones. Also, it can be said the required size for value function table is a function of the depth, $l$, and branch, $d$, of the tree. The branch of the tree is the number of sequence of subtasks as they cannot be matched to current nodes of the tree. The depth of the tree is the number of subtasks that in the worst case equals the length of trajectory when they are not subparts of each other. Thus, the space complexity of value function tables of the hierarchy is $O(ld)$.

**Relative Advantages of SARM-HSTRL**

In this section, a summary of key advantages of the proposed SARM-HSTRL related to other methods is provided.

One key contribution of this method is that despite other methods in the literature that are restricted to only MDPs or FMDPs, SARM-HSTRL can be applied in both MDPs and FMDPs since it does not need an advance knowledge such as the state transitions, or some knowledge or constraints about the size of abstraction or reversible state transitions. Our proposed method works from scratch based on trajectories and without the need for the state transition graph or DBNs, and considers both topological and value intrinsic relationships and structures in trajectories to extract hierarchical structure of tasks.

The proposed SARM-HSTRL can also outperform the HI-MAT algorithm in the sense that HI-MAT only works on a single successful trajectory, while in many RL settings, there are several optimal or near-optimal trajectories that cannot be represented in HI-MAT, unless it is generalized by using another function (i.e., action generalization). However, our proposed method does not require a single, carefully formed, trajectory, and it can efficiently handle the funnel property of subtasks, while HI-MAT cannot be generalized from many different starting places in a few terminal states (i.e., it does not have the funnel property (Mehta et al., 2008)).

In last, the proposed SARM-HSTRL method can be easily scaled up to high dimensional discrete action space and even continuous action space as it considers all paths together at once. The complexity of SARM-HSTRL is a function of complexity of FP-growth algorithm as its main component to extract the associate rules, which is proven to be very practical in terms of time complexity for real usages (Kosters, Pijls, and Popova 2003; Tan, Steinbach, and Kumar 2006).

**Time Complexity**

In this section, we discuss the time complexity of the proposed SARM-HSTRL. Associate rule mining method has a considerably better performance compared to conventional correlation extraction methods such as mutual information or statistical hypothesis testing, since they are often not able to precisely extract the intrinsic correlations among these variables (Tan, Steinbach, and Kumar 2006). However, ARM improves upon such simple methods by using a combination of two key measures of support and confidence in a proven efficient extraction strategy to obtain and evaluate the most efficient and reliable relationships among the variables in datasets. As shown by Tan, Steinbach, and Kumar (2006), for a dataset with $d$ number of items, the number of rules can be calculated as $R = 3^d - 2d + 1$. Therefore, it is impractical to enumerate all possible rules in
large datasets in a naive manner. In Rule Generation, each frequent $k$-itemset has $2^k - 1$ rules, where $k$ is the number of items of the corresponding itemset (Tan, Steinbach, and Kumar [2006]).

As mentioned in [Tan, Steinbach, and Kumar [2006], “the size of a FP-tree typically is smaller than the size of the uncompressed data,” and in the worst-case scenario, the size of a FP-tree is effectively equal to the size of the data. The performance of the FP-growth algorithm is related to the compactness factor of the trajectories and the value of $\minsup$. In the worst-case scenario, the support values of all combinations of items are bigger than $\minsup$, and $2^{d+1}$ itemsets will be generated, where $d$ is the number of items. However, SARM-HSTRL is looking for sub-goals, and the number of sub-goals in an RL task is much less than the size of state space. Thus, using the FP-growth algorithm is efficient and practical in SARM-HSTRL when the state space is large, and the number of sub-goals is relatively low. Such sparsity is a very common assumption in HRL methods [Jonsson and Barto [2006], Mehta et al. [2011]].

As mentioned above, SARM-HSTRL by using FP-growth algorithm method provides a promising solution in practical applications where the state space is large and sparse. If the state space is small, or the successful trajectories have many similarities to each other, many states will be visited frequently, and hence detected by ARM as sub-goals. Clearly, the concept of sub-goals becomes meaningless in such conditions. Another possible scenario to consider is when the adjacent states around the sub-goals are visited frequently. For both these conditions, one efficient solution is to cluster the adjacent sub-goals as one entity and create one corresponding temporally extended action for that entity. $t$, order of occurrence, for each state in each trajectory is already stored by SARM-HSTRL as they are used in HST for possible orderings of sub-goals. They can be also used to find the close sub-goals for clustering purposes.

**Experimental Results**

In this section, several experimental results are presented to evaluate the performance of SARM-HSTRL on four different testbeds. In the first two experiments, the agent has 5 actions, press-key and 4 movement primitive actions. The press-key does not change the place of the agent. The agent can move with its primitive actions in four directions: up, right, down, left. If there is a wall in the way, the agent stays in its current state. In all of the experiments, if the agent does the press-key action, it will receive a reward of 0 in the sub-goal places and a reward of −10 in other states. The reward of other actions is −1. The agent movement with probability 0.8 is according to an intended action and is randomly in one of the directions with probability 0.2. The discount factor is set to $\gamma = 0.9$.

In constructing the HST, 10 start and goal places are chosen randomly. A goal state is defined as an important, task-specific state that ends an episode once visited. A start state, $s_0$, is a state from which an agent begins an episode. For each of them, the agent starts the learning using a common learning mechanism such as Q-learning; the learning is finished after 5000 episodes. They are ordered based on the accumulated reward, and the best five ones are selected. They are given to the SARM-HSTRL and the HST helps the agent to choose their phase of learning. If they are expanded as primitive actions, the number of steps to reach a goal is equal to the number of action selection calls.

The performance of SARM-HSTRL in HRL is evaluated in Figure 1(c) for two different hierarchical structure of tasks, experiment 1 and experiment 2. In these figures, for the sake of comparison between Q-learning, Cascading Decomposition (Chiu and Soo [2010]), HI-MAT (Mehta et al. [2008]) and SARM-HSTRL, 10 runs are considered where in each of them, a start state and a goal state are chosen randomly. The maximum number of actions for each episode is 4000, and the total number of episodes is 8000. SARM-HSTRL is compared to Cascading Decomposition as a representative approach in MDPs, which as discussed in [Chiu and Soo [2010], Ghazanfari and Mozayani [2016], is the latest and considerable improvement for methods proposed in [Simsek and Barto [2009], Mannor et al. [2004], Stolle [2004]]. As seen in this figure, the proposed SARM-HSTRL method results in a hierarchical optimum policy task structure, as does HI-MAT, while our method does not rely on any prior knowledge (e.g. DBNs). It has been proven in [Mehta et al. [2008]] that HI-MAT leads to better results compared to VISA, and this concludes that SARM-HSTRL outperforms the VISA method too. It is worth mentioning that HI-MAT cannot be implemented in experiment 2 including multiple successful trajectories, as it can only work with one successful trajectory that interprets the tasks.

In experiment 1, Figure 1(a), the task hierarchy has 7 levels – it has $(484 \times (7 + 1)) = 3872$ states. If the agent enters in sub-goals states in the following order 1, 2, 3, 4, 5, 6 and 7 and does the press-key action in each of them, and then enters in the goal state of the run and performs the press-key action again, the agent receives a reward of +10, and the episode will be finished. The value of $\minsup$ is 0.9 and the value of $\minconf$ is 0.9.

In experiment 2, Figure 1(b), the task hierarchy has 4 levels, but with a more complicated structure – it has $(484 \times (4 + 1)) = 2420$ states. If the agent enters in one of the sub-goals states from the leaves of tree 1 or 2 or 3, then enters in one of their parent 4 or 5, then in 6 and 7 in order and does the press-key action in each of them, and finally enters in the goal state of the run and performs the press-key action, the agent receives a reward of +10 and the episode will be finished. The value of $\minsup$ is 0.3 and the value of $\minconf$ is 0.9. There is a significant difference in speed of learning between the proposed method with Cascading Decomposition and Q-learning as shown in Figures 2(b) and 2(c). The most important attribute of SMDP framework is using temporally extended actions to decrease the number of steps. As it is shown in HRL in Figures 2(a) and 2(c), the temporally extended actions considerably decrease the number of steps. p-values have been calculated between the proposed method with Q-learning and Cascading Decomposition in each diagram by using the t-test for $\alpha = 0.01$; the significant change is validated – p-values are much smaller than $1 \times 10^{-3}$.

In experiment 3, the accuracy of SARM-HSTRL is evalu-
Figure 2: Performance comparison of SARM-HSTRL with Q-learning, HI-MAT, and Cascading Decomposition in experiment 1 (Fig. 1(a)) and experiment 2 (Fig. 1(b)) of the first testbed described in Figure 1(c). Since experiment 2 including multiple successful trajectories, HI-MAT cannot be implemented. HI-MAT can only work with one successful trajectory that interprets the tasks. (a) Represents the number of steps along episodes in experiment 1. (b) Comparison of receiving rewards along episodes in experiment 1. (c) Represents the number of steps along episodes in experiment 2. (d) Comparison of received rewards along episodes in experiment 2.

Figure 3: (a) A maze world. (b) The frequency of visiting detected subgoals by SARM-HSTRL in transitions. (c) Taxi driver problem as an example in FMDPs. (d) The frequency of visiting the detected subgoals by SARM-HSTRL in transitions in a 4 times scale in places’ dimensions of Taxi driver problem, 16 times larger state space. The states near to wall states of three places are more probable to visit because of they experience less influence of the stochastic rate and the place of pick up places. Also, the four places as the SARM-HSTRL are detected correctly that have the most observing, the brightest ones.
ated thorough all possible subgoals, where 10 random states for start and goal states are selected. The minsup and min-conf are set to 0.6 and 0.9, respectively for Figure 3(a). The agent has four actions up, right, down, and left. Both the stochastic rate and learning rates are 0.1, and the discount factor and the e-greedy are the same as the previous experiments. The agent receives a reward of zero for each action, unless enters to the goal state where it receives 10. The number of trial is 500 for each pair of start and goal states that 5 of the best trajectories are used. As it can be seen in Figure 3(b) SARM-HSTRL detects the subgoals properly. SARM-HSTRL with the given threshold did not consider all the possible subgoals in the right side of Figure 3(a) since the middle ones are placed in better policies, they can reach the possible goals with more probability and less actions.

In experiment 4, we aim to show the accuracy of SARM-HSTRL in FMDPs, using Taxi driver problem as a known testbed (Fig. 3(c)). We scale up both place dimensions of Taxi driver problem for 4 times to reach 20 × 20. The taxi domain is composed of a 5 × 5 grid world, a taxi, and a passenger, where the taxi starts from a random place and pick-up the passenger from one of those places (B, G, R, and Y) and put-down the passenger in one of these places. The place of pick-up and put-down are chosen randomly. The taxi has six primitive actions, north, south, east, west, pick-up, and put-down. The agent receives a reward −1 for movement actions, a reward −10 for wrongly doing the action pick-up or put-down, and a reward +20 for successfully completing the mission. Each action succeeds in its job with the probability of 0.8 in each state and it has a random effect in that state with the probability of 0.2. The number of trials is 2000, and 16 random start and goal states to capture all possible combinations of pick-up and put-down. The maximum number of action is 1000 in each trial. minsup is set to 0.0625 and minconf is set to 0.7. The minsup value is selected as 0.0625 noting that there are 16 combinations for pick-up, and put-down. Discount factor, $\epsilon$-greedy, and learning rate have been initialized similar to experiment 1. As is shown in Figure 3(d) the number of observing detected subgoals for pick-up is correct (the brightest ones). Also, some states in the paths to subgoals are visited more frequently, therefore are being detected as the subgoals. For example, when these states are in the optimal paths of several subgoals, or they are adjacent to the wall states, they will be visited more because of the stochastic rate. They can be easily pruned by considering the sequence and their adjacency to states with biggest support. There is another way in such condition, where the adjacent extracted sub-goal states can be considered as a cluster to define just one temporally extended actions for them.

**Conclusion**

A HRL method called SARM-HSTRL is proposed to autonomously extract a task hierarchy for RL by utilizing a sequential associate role mining approach, where multiple subgoals are extracted as frequently visited states from successful trajectories in the form of association rules. These subgoals are used to define exits as termination conditions to form temporal and state abstractions. Despite the majority of the previously proposed HRL methods (e.g., HI-MAT and VISA) that rely on DBNs model to use prior knowl-

edge about the effects of actions on state variables, our proposed method independently extracts the relations among states and state abstraction. Moreover, since DBNs show the causal relations among the state variables for each action, it can determine irrelevant states variables for state abstraction. However, the proposed method only extracts the relevant correlations. The convergence of the proposed method to a hierarchical optimal solution is proven for both MDPs and FMDPs. The experimental results show a considerable improvement in the speed and quality of the learning process for the analyzed experiments. It is expected that the extracted hierarchical structure in the form of sub-tasks provides a supplementary, and a more robust and higher level of knowledge to be transfered among the sub-tasks rather than sharing value functions, which are highly sensitive to the type and the amount of similarity between the source and target domains. Therefore, the decomposed structure of tasks based on SARM-HSTRL provides an abstraction that an agent can reuse, generalize, and transfer to new domains.

**References**

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In AAAI, 1726–1734.

Barto, A. G., and Mahadevan, S. 2003. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems* 13(4):341–379.

Bebek, G., and Yang, J. 2007. Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC bioinformatics* 8(1):335.

Chiu, C.-C., and Soo, V.-W. 2010. Automatic complexity reduction in reinforcement learning. *Computational Intelligence* 26(1):1–25.

Chiu, C.-C., and Soo, V.-W. 2011. Subgoal identifications in reinforcement learning: A survey. INTECH Open Access Publisher.

Dean, T., and Givan, R. 1997. Model minimization in markov decision processes. In *AAAI/IJCAI*, 106–111.

Dietterich, T. G. 2000. Hierarchical reinforcement learning with the maxq value function decomposition. *J. Artif. Intell. Res.(JAIR)* 13:227–303.

Digney, B. L. 1998. Learning hierarchical control structures for multiple tasks and changing environments. In *Proceedings of the fifth international conference on simulation of adaptive behavior on From animals to animats*, volume 5, 321–330.

Ghazanfari, B., and Mozayani, N. 2016. Extracting bottlenecks for reinforcement learning agent by holonic concept clustering and attentional functions. *Expert Systems with Applications* 54:61–77.

Hengst, B. 2003. *Discovering hierarchy in reinforcement learning*. University of New South Wales.

Jonsson, A., and Barto, A. 2006. Causal graph based decomposition of factored mdps. *Journal of Machine Learning Research* 7(Nov):2259–2301.

Kosters, W. A.; Pijls, W.; and Popova, V. 2003. Complexity analysis of depth first and fp-growth implementations.
of apriori. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 284–292. Springer.

Lin, W.; Alvarez, S. A.; and Ruiz, C. 2002. Efficient adaptive-support association rule mining for recommender systems. *Data mining and knowledge discovery* 6(1):83–105.

Mannor, S.; Menache, I.; Hoze, A.; and Klein, U. 2004. Dynamic abstraction in reinforcement learning via clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 71. ACM.

McGovern, A., and Barto, A. G. 2002. *Autonomous discovery of temporal abstractions from interaction with an environment*. Ph.D. Dissertation, PhD thesis, University of Massachusetts.

Mehta, N.; Ray, S.; Tadepalli, P.; and Dietterich, T. 2008. Automatic discovery and transfer of maxq hierarchies. In *Proceedings of the 25th international conference on Machine learning*, 648–655. ACM.

Mehta, N.; Ray, S.; Tadepalli, P.; and Dietterich, T. 2011. Automatic discovery and transfer of task hierarchies in reinforcement learning. *AI Magazine* 32(1):35–50.

Sigaud, O., and Buffet, O. 2013. *Markov decision processes in artificial intelligence*. John Wiley & Sons.

Şimşek, Ö., and Barto, A. G. 2004. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 95. ACM.

Şimşek, Ö., and Barto, A. G. 2009. Skill characterization based on betweenness. In *Advances in neural information processing systems*, 1497–1504.

Stolle, M. 2004. *Automated discovery of options in reinforcement learning*. Ph.D. Dissertation, McGill University.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tan, P.; Steinbach, M.; and Kumar, V. 2006. *Introduction to Data Mining*. Always learning. Pearson Addison Wesley.

Taylor, M. E., and Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10(Jul):1633–1685.

Wynkoop, M., and Dietterich, T. 2008. Learning mdp action models via discrete mixture trees. *Machine Learning and Knowledge Discovery in Databases* 597–612.
Supplementary Material

Background
In this section, a brief introduction on Markov decision process (MDP) and factored MDP (FMDP) and the corresponding notations in the main paper is provided.

MDPs and FMDPs: RL tasks are typically defined in a Markov Decision Process (MDP) framework as a 5-tuple: \((S, A, P, R, \gamma)\). In this paper, we focus on finite MDPs, where \(S=\{s_1, \ldots, s_n\}\) is a finite set of states, \(A=\{a_1, \ldots, a_m\}\) is a finite set of primitive actions, \(P : S \times A \times S \to [0,1]\) is a one-step probabilistic state transition function, \(R : S \times A \to \mathbb{R}\) is a reward function, and \(\gamma \in (0,1]\) denotes the discount rate. The agent’s goal is to find a policy (a mapping from states to actions), \(\Pi : S \times A \to [0,1]\) that maximizes the accumulated discounted reward \(R = \sum_{t=0}^{\infty} \gamma^t r_t\), for each state in \(S\). FMDPs are known as an extension of MDPs that contain structured representation of problems, where \(T\) and \(R\) are represented in a compact way. In factorized MDPs, the states are described by a set of state variables. To have a unified definition for both MDPs and FMDPs, each state in a MDP can be described by a random variable \(X\), which contains one variable \(X_1, X= (X_1)\), that takes different values. In FMDPs, \(X\) is a multivariate random variable, \(X = (X_1, X_2, \ldots, X_n)\). Each state \(x\) is an instantiation of \(X\), and it can be shown as a vector of \((x_1, x_2, \ldots, x_n)\) such that \(\forall x_i \in \text{Dom}(X_i)\), in which \(\text{DOM}(X) = \{D_1, D_2, \ldots, D_n\}\) refers to the set of possible values for \(X\) as a multivariate variable (Singard and Buffet, 2013).

The value of a state \(s\) based on a policy \(\pi\) is defined as follow: there is always at least one policy that its expected return is equal or greater than any other policies for all states (Sutton and Barto, 1998). Such policy or policies are known as optimal policies and shown with \(\pi^*\). Hence, the corresponding state-value function, \(V\), and action-value function, \(Q\), are optimal and shown as follows: \(V^*(s) = \max_a V_a(s)\) for all \(s \in S\), and \(Q^*(s,a) = \max_a Q_a(s,a)\) for all \(s \in S\) and for \(a \in A(S)\), respectively.

Definitions
Here, we present the definitions of terms used throughout the main paper. Also, all of the following terms has been further explained with some examples in “An Example of SARM-HSTRL” section.

Exit: Transitions are considered unpredictable when they lead to entering or leaving subgoals. The region and the boundaries among states’ clusters that have unpredictable transitions are considered as exits and defined by a state action pair \(G_i = (s_i^E, a)\) when taking action \(a\), as a primitive action, from state \(s_i^E\), as a subgoal, leads to the resultant state that is a goal state to complete subtask \(T_i\) (Hengst, 2003).

Subtask: A subtask, \(T_i\), is a semi-MDP (SMDP) that is shown by \((X_i, S_i, G_i, C_i)\) (Mehta et al., 2011), where \(X_i\) is the set of variables that their corresponding values change during performing the subtask, \(S_i\) denotes the set of admissible states of \(T_i\), \(G_i\) shows the exits of corresponding subtasks as termination conditions of \(T_i\), and \(C_i\) is the set of child tasks of \(T_i\). Child tasks, \(C_i\), can be formed based on different HRL frameworks such as MAXQ or option.

Successful Trajectories: A successful trajectory is defined as a trajectory of states that leads to the goal reward (Mehta et al., 2008).

Abstract State and Abstract Action: In the task hierarchy, subtasks are considered as abstract states and their corresponding local policies as abstract actions (Hengst, 2003).

Region: A region is a set of states that are reachable of each other “such that any exit state in a region can be reached from those state with probability 1 (Hengst, 2003).

An Example of SARM-HSTRL
In this section, we provide a detailed example to describe the proposed SARM-HSTRL on a testbed described in Figure 4. To define a association rule, a pair of \(\langle ITEMSET, Transaction \rangle\) needs to be defined. \(ITEMSET\) is equivalent with \(S = \{s_1, \ldots, s_n\}\) in which \(n\) denotes the size of state space. Therefore, \(ITEMSET = \{s_1, \ldots, s_{60}\}\) in the following example. \(Transaction = \{\Omega_1, \Omega_2, \ldots, \Omega_N\}\) is the set of all transactions. As mentioned earlier, each transaction is defined as a successful trajectory. In other words, each transaction is a trajectory of states from a start state to a goal state. Since start states and goal states are chosen randomly, the first elements as start states and last elements as goal states of these trajectories can be different.

Let us define an experiment to describe the different steps and terms of the SARM-HSTRL in the following maze depicted in Figure 4. In this experiment, there are 3 phases in the system and the agent has five primitive actions: up, right, down, left, and enter. The goal states are in the third phase. The agent starts from the first phase and can move in the second phase if the agent enters state \(s_7\) and takes the action enter. The phase three activates if the agent is in the second phase and enters \(s_{34}\) and does the action enter. There are 60 states, where the first four actions are movement ones (i.e., up, right, down, left) and action enter can take the agent to the next phase. The device starts from a random place, and should pass through states \(s_{27}\) and \(s_{54}\), and go to the goal states, which are selected randomly to receive the goal reward. There is a positive reward to reach the goal state by passing these phases in the right order and a negative smaller reward for taking each action. If we run the Q-learning method for the agent on this maze for different start and goal states, it learns policies gradually during different episodes. An episode is a trajectory of sequence of states and actions and it leads to a goal reward if it reaches the goal state that is in the third phase. Clearly, many episodes in the first of running will not lead to the goal reward. But, Q-learning learns gradually policies to reach the goal state that goes through \(s_{27}\) and \(s_{54}\). We consider three runs, each run corresponds to a different start and goal state, and 200 episodes of learning for each run.
Among 200 episodes of each run, we select the two ones that have the biggest accumulated rewards as follows:

The first run: the start state is $s_1$ and the goal state is $s_{57}$.
$$\Omega_1 = \{s_1, s_2, s_3, s_7, s_7, s_{27}, s_{31}, s_{32}, s_{54}, s_{55}, s_{58}, s_{57}\}.$$ $$\Omega_2 = \{s_1, s_5, s_6, s_7, s_{27}, s_{31}, s_{30}, s_{34}, s_{54}, s_{53}, s_{57}\}.$$ The second run: the start state is $s_3$ and the goal state is $s_{59}$.
$$\Omega_3 = \{s_3, s_7, s_{27}, s_{28}, s_{30}, s_{34}, s_{54}, s_{55}, s_{56}, s_{60}, s_{59}\}.$$ $$\Omega_4 = \{s_3, s_7, s_{27}, s_{31}, s_{35}, s_{34}, s_{54}, s_{58}, s_{59}\}.$$ The third run: the start state is $s_{12}$ and the goal state is $s_{59}$.
$$\Omega_5 = \{s_{12}, s_8, s_7, s_{27}, s_{26}, s_{30}, s_{34}, s_{54}, s_{58}, s_{59}\}.$$ $$\Omega_6 = \{s_{12}, s_{11}, s_7, s_{27}, s_{26}, s_{30}, s_{31}, s_{35}, s_{34}, s_{54}, s_{58}, s_{59}\}.$$ In the above example, $ITEMSET = \{s_1, \ldots, s_{60}\}$, $Transaction = \{\Omega_1, \ldots, \Omega_6\}$, and the number of transaction is 6 ($N=6$). All of the mentioned transactions lead to the goal states of their runs, it means they are successful trajectories. To calculate the association rules in the $ITEMSET$, we need to calculate the support and confidence factors for each possibility of extracted association rules in form of $A \rightarrow B$ that is performed using FP-growth algorithm noting its efficient performance. As mentioned, support and confidence are calculated as follow:

$$support(A \rightarrow B) = \frac{\sigma(A \cup B)}{N}$$ $$confidence(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

For instance, if $A=s_1$ and $B=s_{58}$, $support(s_1 \rightarrow s_{58}) = \frac{1}{6}$. Since $s_1$ and $s_{58}$ are only simultaneously observed in $\Omega_1$; thus, $\sigma(s_1 \cup s_{58}) = 1$. The confidence $(s_1 \rightarrow s_{58})= \frac{1}{6}$. Here, we use the sequential ARM (SARM) technique rather than conventional ARM technique due to its capability to consider the order of items in addition to their occurrence frequency. In this case, the confidence $(s_{58} \rightarrow s_1)= 0$.

As another example, if we consider $A=s_7$ and $B=s_{34}$, $support(s_7 \rightarrow s_{34}) = \frac{1}{6}$. Since $s_7$ and $s_{34}$ are visited in all $\{\Omega_1, \ldots, \Omega_6\}$; thus, $\sigma(s_7 \cup s_{34}) = 6$. The confidence $(s_7 \rightarrow s_{34}) = \frac{1}{6}$. If we set $minsup=0.9$ and $minconf=0.9$, $s_7$, $s_{27}$, $s_{34}$, and $s_{54}$ validates. Since $s_7$ and $s_{27}$ are consecutive by one action, $enter$, the exit is defined as $(s_7, enter)$. In the same way for $s_{34}$ and $s_{54}$, $(s_{34}, enter)$ is considered as the exit. Therefore, only states $s_7$ and $s_{34}$ are used for HST-construction method and the result structure is presented in Figure 5.

The proposed SARM-HSTRL method extracts the exits $G_0 = (s_7, enter)$ and $G_1 = (s_{34}, enter)$ to form subtasks $T_0$ and $T_1$. The edge between subtasks $T_0$ and $T_1$, denoted by $E$, in the extracted graph shows the relation between these subtasks. Based on the extracted exits, the transactions are partitioned into three regions as follows: $\{s_1, \ldots, s_{20}\}, \{s_{21}, \ldots, s_{40}\}$, and $\{s_{41}, \ldots, s_{60}\}$. $T_0$ as the subtask is formed of $S_0=\{s_1, \ldots, s_{20}\}$ and $G_0 = (s_7, enter)$. It does not have any child. $T_1$ as the subtask is formed of $S_1=\{s_{21}, \ldots, s_{40}\}$, $G_1 = (s_{34}, enter)$, and its child is $C_1=I_0$. $T_0$ and $T_1$ nodes can be considered as abstract states and their corresponding policies as abstract actions. The

Figure 4: The example testbed: the size of the testbed is $4 \times 5 \times 3 = 60$ states.
Figure 5: The extracted task hierarchy of the Figure 4. An example of a sequential association rule mining and HST-construction of one rule.

Figure 6: An example of a HST-construction of several rules.

An Example of HST Construction of Several Rules

As an example, let us consider $AR_1 = bcde$, $AR_2 = dbce$, $AR_3 = acde$ (see Figure 6). The proposed algorithm, first constructs a tree with the reverse of $AR_1$, creating one branch with values $edcb$. Then, the reverse of $AR_2$ is added to the tree, making a new branch from $c$ since $AR_{2,2} = c$ cannot be matched in the tree from that point. Thus, a new branch from $c$ is created to contain the remaining values of $AR_2$. Finally, the reverse of $AR_3$ is added to the tree, where a mismatch happens in $AR_{1,4}$ that results in a new branch created at node $e$.

References

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In AAAI, 1726–1734.

Barto, A. G., and Mahadevan, S. 2003. Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems 13(4):341–379.

Bebek, G., and Yang, J. 2007. Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC bioinformatics 8(1):335.

Chiu, C.-C., and Soo, V.-W. 2010. Automatic complexity reduction in reinforcement learning. Computational Intelligence 26(1):1–25.

Chiu, C.-C., and Soo, V.-W. 2011. Subgoal identifications in reinforcement learning: A survey. INTECH Open Access Publisher.

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

Dieterich, T. G. 2000. Hierarchical reinforcement learning with the maxq value function decomposition. J. Artif. Intell. Res.(JAIR) 13:227–303.

Digney, B. L. 1998. Learning hierarchical control structures for multiple tasks and changing environments. In Proceedings of the fifth international conference on simulation of adaptive behavior on From animals to animats, volume 5, 321–330.

Ghazanfari, B., and Mozayani, N. 2016. Extracting bottlenecks for reinforcement learning agent by holonic concept clustering and attentional functions. Expert Systems with Applications 54:61–77.

Hengst, B. 2003. Discovering hierarchy in reinforcement learning. University of New South Wales.

Jonsson, A., and Barto, A. 2006. Causal graph based decomposition of factored mdps. Journal of Machine Learning Research 7(Nov):2259–2301.

Kosters, W. A.; Pijls, W.; and Popova, V. 2003. Complexity analysis of depth first and fp-growth implementations of apriori. In International Workshop on Machine Learning and Data Mining in Pattern Recognition, 284–292. Springer.

Lin, W.; Alvarez, S. A.; and Ruiz, C. 2002. Efficient adaptive-support association rule mining for recommender systems. Data mining and knowledge discovery 6(1):83–105.

Mannor, S.; Menache, I.; Hoze, A.; and Klein, U. 2004. Dynamic abstraction in reinforcement learning via clustering. In Proceedings of the twenty-first international conference on Machine learning, 71. ACM.

McGovern, A., and Barto, A. G. 2002. Autonomous discovery of temporal abstractions from interaction with an environment. Ph.D. Dissertation, PhD thesis, University of Massachusetts.

Mehta, N.; Ray, S.; Tadepalli, P.; and Dieterich, T. 2008. Automatic discovery and transfer of maxq hierarchies. In Proceedings of the 25th international conference on Machine learning, 648–655. ACM.
Mehta, N.; Ray, S.; Tadepalli, P.; and Dietterich, T. 2011. Automatic discovery and transfer of task hierarchies in reinforcement learning. *AI Magazine* 32(1):35–50.

Sigaud, O., and Buffet, O. 2013. *Markov decision processes in artificial intelligence*. John Wiley & Sons.

Şimşek, Ö., and Barto, A. G. 2004. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 95. ACM.

Şimşek, Ö., and Barto, A. G. 2009. Skill characterization based on betweenness. In *Advances in neural information processing systems*, 1497–1504.

Stolle, M. 2004. *Automated discovery of options in reinforcement learning*. Ph.D. Dissertation, McGill University.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tan, P.; Steinbach, M.; and Kumar, V. 2006. *Introduction to Data Mining*. Always learning. Pearson Addison Wesley.

Taylor, M. E., and Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10(Jul):1633–1685.

Wynkoop, M., and Dietterich, T. 2008. Learning mdp action models via discrete mixture trees. *Machine Learning and Knowledge Discovery in Databases* 597–612.