Multi-Skill Agents Coalition Formation Under Skill Uncertainty

Seyyed Mohammad Sayyadi Kenari
Software Engineering Department
Mashhad Branch - Islamic Azad University
Mashhad, Iran
sm_sayyadi@yahoo.com

Majid Vafaei Jahan
Software Engineering Department
Mashhad Branch - Islamic Azad University
Mashhad, Iran
VafaeiJahan@mshdiau.ac.ir

Mehrdad Jalali
Software Engineering Department
Mashhad Branch - Islamic Azad University
Mashhad, Iran
Jalali@mshdiau.ac.ir

Abstract— **In a multi-agent system, there are situations in which agents are unable to do their tasks individually. Therefore forming coalitions is inevitable. In natural settings, an agent decides to form coalition based on beliefs it has regarding the capabilities of other agents. In the previous works, it was assumed that a single type can reflects all capabilities an agent has. We introduce multi-skill agents which have a value per skill. This helps us to solve more problems and to reason about the results, more exactly. We use Bayesian Reinforcement Learning (BRL) as the learning mechanism. Through the repeated use of BRL, agents can form more rewarding coalitions. We extend existing algorithms of the repeated coalition formation under type uncertainty to the skill uncertainty and exploit them in experimental studies that type uncertainty couldn't do or reason about. Average long term discounted expected reward that agents accumulate in the learning process, is the criteria we test our methods based on. We test the algorithms on a sample soccer sub-team formation problem. To have a notion of the best performance, we solve the problem in the absence of uncertainty. Results show that the VPI method does approximately 85% of the best performance.**

Keywords: Bayesian Reinforcement Learning, Coalition Formation, Multi-Skill Agent.

I. INTRODUCTION

The problem of coalition formation has received much attention these years [7, 9]. The problem of coalition formation arises in situations in which an individual agent is incapable of doing a special task. The rational autonomous agents act together to achieve some notion of reward the environment prepares for them as a result of doing a task that they couldn't perform individually.

To be complete, a task requires coalition's members have at least specific amounts of different capabilities. But what combination of agents is good for an agent to join, for a typical task. The criterion which an agent makes decision upon it is the capabilities of other agents. Most existing models of coalition formation assume that agents have knowledge of their potential partners' capabilities, or at least that this knowledge can be reached via communication [7] [6]. However in many natural settings this hypothesis is not the case. In other words, in these settings agents don't know the actual capabilities of other agents; however, they have beliefs about those. Chalkiadakis and Boutilier in [2] introduce this uncertainty inherent in the coalition formation process. The uncertainty they deal with is Type Uncertainty, assuming each agent has a type that reflects its capabilities. The type of each agent is unknown for other agents. On the other hand, agents have beliefs regarding the types of other agents. Using Bayesian multi-agent reinforcement learning algorithm repeatedly in the formation process, agents can update their beliefs about the types of other agents and consequently learn what is good for them to do.

Unfortunately there are problems that type uncertainty can't solve it. In other words, a single type can't necessarily reflect all the capabilities an agent has. For more illustration consider the problem of forming a football team. To be successful, the team requires goalkeepers, defenders, midfielders and forwards. These four, are the available skills in the environment. Intuitively we can assume that each agent can play in more than one of these roles. (i.e. the agents can have multiple types at the same time). Additionally it is natural to assume that two players which have defender and midfielder skills are not the same. For example suppose the defense skill of agent1 is more than that of agent2 and agent2's midfielder skill overcomes that of agent1. This is what type uncertainty can't describes. In other words, each agent have some value of each skill exists in the environment. A forward that doesn't have defender skill is modeled by having defender skill at its lowest possible value.

In this paper we introduce Multi-Skill agents that each of them has a value regarding each skill exists in the environment, to handle the situations in which the value of each skill is important. The uncertainty inherent in the environment can be modeled as the agents' beliefs about the skill-values of other agents. This Skill Uncertainty can be handled through the use of RL. Like Chalkiadakis and Boutilier in [7] we cast the model of single agent BRL introduced in [5] as a solution of exploration vs. exploitation trade-off. We extend the Myopic, VPI and MAP algorithm introduced in [4] to match multi-skill circumstances and compare their result to fully-informed algorithm's result as a performance metric. We show that those algorithm's tractability remain while full POMDP solutions are intractable.

978-1-4244-9834-5/11/$26.00 ©2011 IEEE

The rest of the paper is structured as follows: Section 2 provides background on coalition formation, multi-skill agents and BRL; Section 3 describes the BRL framework for optimal repeated coalition formation process under skill uncertainty. In section 4 we present our solutions of coalition formation process. It consists of the detailed scenario of multi-agent coalition formation process from beginning to the end. In section 5, we present our algorithms to evaluate the prescribed solution and show compared results of the algorithms to the fully-informed algorithm's result, and Section 6 provides a discussion of related work.

II. BACKGROUND

In this section we provide some background on Reinforcement Learning and Coalition Formation.

Coalition formation is the process by which individual agents form such coalitions, generally in order to solve a problem by combining their efforts [4]. Suppose there are N $(N > 2)$ agents in the environment. Any subset $C \subseteq N$ is called a *Coalition* and we assume that they act in a way that achieves a common goal. A coalition structure is a partition of the set of agents containing exhaustive and disjoint coalitions [2]. In other words, the coalition structure at time step t, consists of all coalitions already exist in the environment. It is worth mentioning that single agent coalitions are possible. We denote by $V(C)$ the value that coalition C's members can achieve through acting together. Consequently, in a repeated coalition formation process the agents try to learn that joining to what coalitions offers them a greater $V(C)$.

The method we used for learning is reinforcement learning (RL). A single-agent RL problem is the problem of an agent acting in an environment which its transition and reward model is unknown. Through a trial and error process the agent learn to do best in the environment [12]. Formally a RL problem can be modeled as a Markov Decision Process (MDP). A MDP can be formally defined by a tuple $\langle S, A, T, R\rangle$. S is the set of states of environment; the set of actions is denoted by A. The transition model T describes the probability $T(s, a, s')$ of reaching state s' by doing action a at s. The reward $R(s, a, s')$ agent receives by doing in this manner is defined by reward function R. The aim of agent is to maximize its future discounted expected reward through finding the best policy π^*. A policy $\pi: S \to A$ maps states to actions. The value of optimal policy V^* can be computed using a number of standard classic RL methods such as policy iteration or value iteration.

In partially observable state environments things differ a little bit. In these environments we deal with a Partially Observable Markov Decision Process (POMDP) which can formally be defined by a tuple $\langle S_P, A_P, O_P, T_P, Z_P, R_P\rangle$ [11]. $S_P = S \times \theta_a^{s,s\prime}$ is the product of basic states' set S and the unknown continuous parameters $\theta_a^{s,s\prime}$. θ is the transition model. $A_P = A$ is as same as underlying set of actions. $O_P = S$ is the set of observable states. The transition function $T_P(s, \theta, a, s', \theta') = \Pr(s'|s, \theta_a^{s,s'}, a) \times \Pr(\theta'|\theta)$, can be rewritten $T_P = \theta_a^{s,s'} \times \delta_\theta(\theta')$ where δ is Kronecker delta. The probability of observing o when the state s', θ' is reached is defined by observation function Z_P. Since the reward the environment produces, doesn't depend on the transition model θ, R_P doesn't differ from MDP.

Based on this POMDP formulation, we can learn the transition model θ by belief monitoring. At each time step, the belief $b(\theta) = \Pr(\theta)$ over all unknown parameters $\theta_a^{s,s'}$ is updated based on the observed transition s, a, s' using Bayes' theorem: $b_a^{s,s'}(\theta) = zb(\theta)\Pr(s'|s, a, \theta) = zb(\theta)\theta_a^{s,s'}$. Note that z is a normalizing constant. As mentioned above, $b(\theta)$ is a probability and its sum over all possible value must be akin to one. This can be done through the use of a normalizing constant. Belief monitoring can perform easily when prior and posterior belong to the same family of distributions. We use product of Dirichlets because it has the described property. A Dirichlet distribution is defined as $D(p; n) = z\prod_i p_i^{n_i - 1}$ where p is a multinomial and parameters n_i can be interpreted as the number of times that the p_i-probability event has been observed plus one. Since the unknown transition model θ is made up of one unknown distribution θ_a^s per s, a pair, let the prior be $b(\theta) = \prod_{s,a} D(\theta_a^s; n_a^s)$ such that n_a^s is a vector of hyperparameters $n_a^{s,s\prime}$. The posterior obtained after transition $\hat{s}, \hat{a}, \hat{s}'$ is: $b_a^{s,s'}(\theta) = z\theta_a^{s,s'}\prod_{s,a} D(\theta_a^s; n_a^s) = D(\theta_a^s; n_a^s + \delta_{\hat{s},\hat{a},\hat{s}'}(s, a, s'))$.

Exploration and exploitation trade-off is a major challenge in RL studies. Through the use of this BRL method we get the best possible trade-off.

III. MULTI-SKILL AGENTS BRL

In this section we deal with the problem of multi-skill agents' coalition formation in an uncertain environment.

Coalition formation problem can be translated to trying some cooperative agents to achieve as more reward as possible in a multi-agent environment through joining together in the manner that is sequentially rational. Therefore, the first challenge an agent faces to, is joining to what coalition can be more rewarding. Naturally a coalition that the capabilities of its members satisfy the coalition's facing task's requirements is good enough to join. So the agents decide upon the capabilities of other agents. But in natural settings the capabilities of agents are unknown to each other. Instead, each agent has beliefs regarding the capabilities of other agents.

The assumption of having a single type per agent can't be the case in some settings. There are settings in which agents have *multiple skill*s instead of a single type. For example consider the problem selecting a defender player to incorporate in a free kick task in a football game at the state of attack. Intuitively an individual single player can't necessarily be successful. Therefore a sub-team (coalition) of players is needed to the team be successful with the kick. This coalition requires some capabilities. Suppose the plan is

to achieve the goal through the use of head kick. Having a single type defender causes the other players don't form coalition with him. But introducing multi-skill agents in the way that the player has two skills, both forward and defender, can change decisions. One can argue that defining a single type as a mixture of these two skills can probably fix the problem. However, there remains the problem of choosing between players which have this new type but at different *value*s. In fact, the value of defender skill of player1 can be less than that of player2 and be for forward skill vice versa. Intuitively the others prefer player2 in the prescribed situation. But with the new defined type they can't reason about the success with player2 and failure with player1.

Multi-Attribute (Multi-skill) agents was introduced in [6] but in a certain environment. Additionally the focus of that work was upon computational complexity of reaching to a stable coalition structure, rather than formation process. Solving the repeated coalition formation problem of agents with multiple skills and a value per skill in an uncertain environment is our main contribution in this paper.

To be sequentially rational, the agents must update their beliefs in repeated coalition formation. This can be done via the BRL method introduced in section 2. The reason we choose the BRL is that it solves the exploration and exploitation dilemma through focusing on achieving as more reward as possible; so it gets the best performance ever possible.

A. Multi-Skill Agents Coalition Formation Model

Assume there are $N = \{1, .., n\}$ agents in the environment. We denote by S the set of available skills in the environment and each skill $s \in S$ has a set V_s, which contains the possible values for that skill. Each agent i has a possible value of all skills available in the environment. We briefly show the skill-values of agent i by the set $SV_i = \times_{v \in V_s} v, \forall s \in S$ W. Consequently $SV_C = \times_{i \in C} SV_i$ denotes the skill value profile of coalition C's members. The beliefs B_i of agent i construct a joint distribution over skill-values of other agents. The belief agent i has about the skill-values that other members of a coalition $C \subseteq N$ have, is denoted by $B_i(SV_C)$. Each agent knows its SV_i and assigns the probability of one to its true skill-value and zero to all other possible skill-values. The set of coalitional actions which a typical coalition faces to, to choose from is denoted by the set A.

To use BRL method defined in section 2, we should specify its elements. A state s defined by a *Coalitional Agreement (CA)* which is a <Coalition, agreed upon [1] Action> pair. The transition model is the set SV of skill-values agents have and is given a priori. The set of actions is A. The Transition function is defined by the probability $\Pr(s'|SV_C, \alpha)$ of observing outcome state s' (reaching state s') after doing action α at state (coalitional agreement) s when the skill-values of the coalition's members be SV_C. As mentioned in section 2, we use Dirichlets priors to calculate the observation function. $R(s')$ is the reward that environment assigns to the coalition's members once they reach to state s'. The value of each coalition is the maximum reward that its members expect to gain through acting jointly. Since agents have beliefs about the skill-values of each others, the expected value of a coalition differs from an agent to another. This can be calculated via formulation of (1):

$$V_i(C) = \max_{\alpha \in A} \sum_{sv_C \in SV_C} B_i(SV_C) \Pr(s'|SV_C, \alpha) R(s') = \max_{\alpha \in A} Q_i(SV_C, \alpha) \quad (1)$$

B. Optimal Coalition Formation

We now describe the scenario of optimal repeated coalition formation.

Agents repeatedly form coalitions; see the outcome and update their beliefs. At each stage, an agent is chosen randomly as a proposer. The proposer is able to propose one of the following proposals:

1. Stay in its current CA and optionally propose a new coalitional action.
2. Leave current coalition to join another CA.
3. Leave current coalition and form a singleton coalition.

In the entire aforementioned scenarios, all affected agents must be agreed upon to the change. Affected agents include all of the source and destination coalitions. After forming a coalition and performing the coalitional action, at time step t, members will update their belief in the way introduced in section 2 via formulation of (2):

$$B_i^{t+1}(SV_C) = z \Pr(s'|s, SV_C, \alpha) B_i^t(SV_C) \quad (2)$$

Where z is a normalizing constant, as described before, in section 2.

When deciding to choose between possible coalitions, agents should not consider solely the immediate reward they can receive. They should decide upon the expected discounted future reward they can achieve via forming a coalition and continue the process afterward. They can use the transition model to predict the future coalitions that will formed. But through the use of updated beliefs in prediction of the next coalitions, agents' decisions will be more rational and the sequential rationality will be guaranteed.

Therefore the optimality equations for the POMDP, can be formulated as an *Infinite Horizon* [2] problem with the

[1] In fact after forming a coalition, its members must all agree upon an action to perform.

[2] Infinite horizon is a common term in RL literatures and point to infinite chain of state then action then next state and so on.

discount factor $\gamma, 0 \leq \gamma < 1$. Let $Q_i(C, \alpha, B_i^t)$ be the long term value that agent i places on being a member of coalition C which its members agreed upon action α while $i's$ current belief state is B_i^t. This can be defined through Bellman like equations:

$$\begin{aligned} & Q_i(C, \alpha, B_i^t) \\ &= \sum_{s'} \Pr(s'|C, \alpha, B_i^t) \left(R(s') + \gamma V_i(B_i^{t+1})\right) \\ &= \sum_{SV_C} B_i^t(SV_C) \sum_{s'} \Pr(s'|SV_C, \alpha) \left(R(s') + \gamma V_i(B_i^{t+1})\right) \end{aligned} \quad (3)$$

$$V_i(B_i^t) = \sum_{C|C \in VC_i^t} \Pr(C, \alpha|B_i^t)\, Q_i(C, \alpha, B_i^t) \quad (4)$$

Where valid coalition's set for agent i at time step t, VC_i^t, is the set of next coalitions that agent i may be member of, through proposal of itself or the other agents. For more illustration, we prepare a more detailed description for formulation of (3) here. Intuitively, a good way to calculate a function over an infinite horizon is to solve it recursively. In this formulation, the sigma indicates all possible next CAs (states) which the agent i's coalition can go to after going to evaluating coalition C. The next term, $\Pr(s'|C, \alpha, B_i^t)$, indicates the probability of this transition. The term $R(s')$ shows the amount of reward that environment assigns to i for this transition. The last term, $\gamma V_i(B_i^{t+1})$, is the recursive term. With a discount factor, it incorporates the value of the updated belief, to the calculation of long term discounted expected reward. While agent i is in belief state B_i^t, it may find itself participating in any number of such valid CAs, so $V_i(B_i^t)$ specifies the value of belief state B_i^t to agent i. Note that unlike classic Bellman equations, the value of belief state B_i^t can't be defined by maximizing the Q-value function. Because the coalitions that will be formed don't depend solely on agent $i's$ decision but will be on all affected agents' decisions[3].

The difficulty with the aforementioned formulation is prediction of the term $\Pr(C, \alpha|B_i^t)$ in (. The term $\Pr(C, \alpha|B_i^t)$ can be estimated in a variety of ways. Nevertheless, in realistic environments it can be challenging due to the lack of knowledge of other agents' strategies or the limit on set of observable states of the environments. However, by assuming this value at hand, the agent can incorporate the new more refined beliefs regarding the skill-values of other agents in predicting the future coalitions. This way the prediction becomes more accurate.

Assuming the Q-value of a CA to be at hand, the repeated coalition formation process will be straightforward. At each stage, an agent is chosen randomly from the set of not affected agents as proposer. This proposer agent proposes the CA with the highest Q-value. The other agents in the current coalition of proposer and the agents already exist in the proposed coalition are the affected agents. All the affected agents must agree to the change. They decide upon the Q-value of new coalition structure they belief, too. In fact, if an agent believes that the proposed CA is the best coalitional agreement possible, it agrees. Also with a little probability $\epsilon > 0$, an agent may agree to a non-best change, hoping that may be the proposed change will lead the coalitional structure to a better unseen state. This is as same as the *Best Reply with Experimentation (BRE)* method introduced in [2]. If agreement be revealed, all the affected agents will be removed from not-affected agents' set. The process then continues until the set of not-affected agents becomes empty. Then, the formed coalitions do their agreed upon action and the environment reward them. The reward of each coalition is divided between its members. The agents update their beliefs in the way that described before and the process continues.

[3] Note that we didn't describe the difficulties which may be added to the problem due to modeling the strategies of other agents.

IV. Computational Approximation

The exact solution to the POMDP formulation of (3) is computationally infeasible. On the other hand, the number of states and actions will grow exponentially with the number of agents. So some algorithms should be developed to approximately compute the formulation.

In this paper we introduced three algorithms to approximate the formulation of (3). These algorithms can be seen as extensions of those algorithms introduced in [4] to the multi skill-value state. The algorithms include: *Myopic BRL Algorithm, Value of Perfect Information Algorithm* and *Maximum A Posteriori Skill Assignment Algorithm.*

A. Myopic BRL

In this algorithm, agents myopically see the future and take it into account. This algorithm doesn't gain from incorporating the refined belief state in predicting possible future CAs. However, since at the end of each stage, the agents update their beliefs regarding the skill-values of other agents, this update affect the decisions the agent will make in the next stage. In fact this algorithm consider only the immediate reward, the agents achieve and can be formulated as formulation of (5):

$$Q_i(C, \alpha, B_i^t) = \sum_{SV_C} B_i^t(SV_C) \sum_{s'} \Pr(s'|SV_C, \alpha) R(s') \quad (5)$$

B. Value of Perfect Information

Value of perfect information algorithm (VPI) tries to gain from choosing CAs which have greater *value of information*. Value of information of a CA, is the amount of new information it accompanies which consequently affect the agent's beliefs. Since it tries better CAs between all valid CAs, we should call it *VPI Exploration*. In fact, VPI Exploration predicts the myopic value of information of a CA.

VPI has special properties. It samples over skill-value configurations of other agents instead of its belief state. Also, it calculates VPI for CAs instead of actions of a single agent. The calculated VPI then is combined with the true Q-value

of the CA to bias proposing mechanism to choose states with more value of information. Intuitively, states which have proposed more, have less value of information. Note that knowing the true skill-values of the other agents is not the goal of agents. The single criterion for agents is to achieve to the best discounted expected long term reward possible.

The formulation of VPI is as follows. We want to choose between available CAs. Therefore assume a typical coalitional agreement $\sigma = \langle C, \alpha \rangle$ among the available CAs. Suppose σ is adopted and its coalitional action is executed. Let s' is an exact observation the agent sees with the assumption of the *true* skill-values of the other agents are available. What can be learned if this assumption be valid? Under these assumptions the true Q-value of σ will be achieved. Let q^* be the true Q-value of σ:

$$q_\sigma^* = q_{\langle C,\alpha \rangle}^* = Q_i(C, \alpha | SV_C^*) = \sum_{s'} \Pr(s' | SV_C, \alpha) R(s') \quad (6)$$

This formula myopically calculates the true Q-value of typical coalitional agreement σ under the assumption that the true skill-values of coalition's members are available.

The new information achieved, is useful only if it affects the future decisions made by agent. Two scenarios are possible which the new knowledge will affect the future decision making.

Before going into detail, suppose the Q-value of first and second best CAs regarding the current belief state to be $q_1 = Q_i(\sigma_1 | B_i^t) = Q_i(C_1, \alpha_1 | B_i^t)$ and $q_2 = Q_i(\sigma_2 | B_i^t) = Q_i(C_2, \alpha_2 | B_i^t)$ respectively. If the new knowledge indicates that there is a coalitional agreement σ that isn't as same as previously considered best coalitional agreement σ_1 and has higher Q-value than q_1, the agent must prefer σ to σ_1 gaining $q_\sigma^* - q_1$.

Also if the CA with the best Q-value be equal to the typical coalitional agreement σ ($\sigma = \sigma_1$) and myopically calculated value for σ indicates that the second best coalitional agreement σ_2 has a higher Q-value than q_σ^*, the agent should prefer σ_2 to σ_1; gaining $q_2 - q_\sigma^*$.

Therefore the gain from learning the true value q_σ^* of coalitional agreement σ can be formulated as:

$$gain(q_\sigma^* | SV_C^*) = \begin{cases} q_\sigma^* - q_1, & if\ \sigma \neq \sigma_1\ and\ q_\sigma^* > q_1 \\ q_2 - q_\sigma^*, & if\ \sigma = \sigma_1\ and\ q_\sigma^* < q_2 \\ 0, & othewise \end{cases} \quad (7)$$

However, the agent doesn't know a priori the true skill-values other agents have. So the agent should calculate the expected gain, it can achieve, of all possible skill-value configuration of the other agents based on its current belief state. The formulation of expected gain of perfect information is:

$$EVPI(\sigma | B_i^t) = \sum_{SV_C^*} gain(q_\sigma^* | SV_C^*) B_i^t(SV_C^*) \quad (8)$$

This is the expected gain from learning the true value of σ. We use $EVPI(\sigma | B_i^t)$ as the representative value of choosing σ instead of choosing solely based on the current belief state. By this value at hand, the value of σ can be formulated as follows:

$$QV_i(\sigma | B_i^t) = Q_i(\sigma | B_i^t) + EVPI(\sigma | B_i^t) \quad (9)$$

Using QV_i values instead of Q_i, agents can choose more informative CAs and this way the learning speed increases considerably. We summarize the steps to use VPI exploration below:

1. Calculate the true value of each possible CA σ using (6).
2. Compute the gain of learning the true value of σ using (7).
3. Since the true value of skill-values, the coalition's members have, is unknown in advance; compute the expected gain of perfect information via the formulation of (8).
4. Make the Q-value computed by formulation of (9) as the criterion by which agents decide upon changes in coalition structure.

One can argue that VPI exploration is also a myopic method. The answer that this it is not correct is that although it calculates the myopic Q-value of the CA, but it takes the current belief state into account in calculation of expected gain of perfect information. This way, an agent exploits the true q-value of the CA and explores the gain it expects based on the current belief state.

C. Maximum a Posteriori Skill Assignment

In this algorithm, we estimate the true Q-value of a CA through assuming the true skill-values of coalition's members to be same as the most probable skill-values based on the current belief state. In other words, at time step t, the skill-values of agent j based on the current belief state of agent i in this algorithm is: $MAPSV_j^t = \arg\max_{SV_j^t} B_i(SV_j^t)$. This way the vector of all possible skill-values of coalition's members reduces to a single element. Finally the agent myopically calculates the Q-value of the CA:

$$Q_i(C, \alpha, B_i^t) = \sum_{s'} \Pr(s' | MAPSV_C^t, \alpha) R(s') \quad (10)$$

Notice that this is a myopic algorithm, too. Beliefs show what fractions of model are not well explored; so this algorithm focuses on exploration and hopes that the most probable skill-values configuration at time step t is the true skill-values that can consequently approaches agents to the true Q-value of the CA.

V. EXPERIMENTAL RESULTS

We conducted several types of experiments to evaluate our method but we present two in this paper. To show the performance of each algorithm, in the first test scenario we compare the algorithms introduced in the previous section to an optimal behavior method ever possible, called "*Fully Informed*". Also, as mentioned above, the main contribution of our method is to solve problems which the type uncertainty method is unable to solve. Therefore we present a test scenario that matches this claim as the second test case.

In the first test scenario we compare the performance of introduced algorithms in a 10-agent environment. Each agent has a possible value of all three skills available in the environment. The number of skill-values for a special skill may vary. For simplicity reasons we assume two skill-values per skill (for example: low and high). The environment is homogeneous, i.e., all agents in the environment use the same algorithm. Each possible coalition faces two coalitional actions and the discount factor, we selected is 0.9.

Our metric to compare the performance of the algorithms is the average discounted accumulated reward agents gained in the repeated coalition formation process. This is as same as the metric used in all BRL methods. As mentioned above using this metric the best exploration vs. exploitation tradeoff is established. For the sake of simplicity, we translate the skill values to the numeric one. The low skill value translates to zero and one is equivalent to high skill value. The sample reward function gives more reward to coalitional agreements with more sum of Skill2 or Action1 as the chosen action. However this is a sample reward function and there may be no equivalent to it in realistic settings.

To show the difference between performances of our algorithms to an optimal algorithm, we introduce an algorithm called "*Fully Informed*". Agents using this algorithm have no uncertainty regarding the skill values of the other agents. In other words, all agents know the skill values of the other agents a priori. Therefore they can do their best, without wasting their time to sample over all possible POMDPs.

The result of our experiments is depicted in Fig. 1. As shown in this figure, the Fully Informed algorithm's performance is the best. It accumulates 4.72 reward units after 500 RL steps. It's not wonderful, because there is no uncertainty regarding the skill values of partners for an agent. The best performance algorithm introduced in this paper is VPI. VPI pays more attention to states that have greater impact on future decision making through calculation of Expected Value of Perfect Information. In VPI, a CA is selected not only for its long term discounted expected reward but also for its expected change of the best next state. Using this heuristic, VPI can form more rewarding CAs against other algorithms in the same number of iterations. The final accumulated discounted expected reward of VPI after 500 RL steps is 4.1 (approximately 87% of the best performance). Maximum a posteriori skill assignment algorithm is the second best performance algorithm. MAP myopically calculates the long term discounted expected reward and assumes that the true skill values of other agents already are at hand, based on the current belief state. Beliefs show what partitions of model are not well explored. Therefore, focusing to explore these partitions, MAP algorithm gathers more information and can do somewhat good against the fully informed algorithm. As shown in Fig. 1, MAP algorithm accumulates 3.7 reward units which is approximately 78% of the best performance. The Myopic algorithm accumulates only 2.7 reward which is 57% of the best performance. Myopically calculation of expected reward means not accounting for sequential decision making. Additionally, the lack of any heuristic, leads to poor performance of Myopic algorithm against other algorithms introduced in this paper. Although, as mentioned in previous section, in this algorithm, the agents update their beliefs based on the observed outcomes at each stage and this way they can form more and more rewarding coalitions as the number of iterations goes up.

We test the algorithms under various sample reward functions and discount factors in this environment. All of them report approximately the same results proportionally.

The most important contribution of this paper is to show that assuming multiple skills for an agent; we can solve more problems that match the repeated coalition formation process. So we consider an example of soccer world for more illustration as the second test case.

Assume there are three skills available for an agent in a soccer environment: {Pass, Shoot, Head}. Each player (agent) has all of these skills but at one of these values: {Low, High}. These agents intend to form optimal sub-teams to do one of these actions: {Defense, Midfielder, Attack}. The reward function of the environment changes proportionate to soccer's requirements. For example, for midfielder action, having the head skill is not so rewarding unlike other two skills. Additionally to encourage agents to form coalitions versus acting individually, the reward function added some reward when the number of coalition's members increases. At the end of repeated coalition formation process each agent can reason what skill values of its partners leads to success for each action; which is what type uncertainty can't do due to lack of multiple skills and a value per skill. The density of total values of each skill in the final[4] coalition structure can be served as a metric of the weights that the reward function of environment places on the values of each skill for each action.

The result of repeated coalition formation process of test case two is depicted in Fig. 2. Four agents {Agent0, Agent1, Agent2 and Agent3} with skill values {Agent0: Pass (High), Shoot (High) and Head (High)}, {Agent1: Pass (High), Shoot (High) and Head (Low)}, {Agent2: Pass (Low), Shoot (High) and Head (High)} and {Agent3: Pass (High), Shoot (Low) and Head (High)} engage in this process. According

[4] It was proved in that BRE process will converge to a stable state if there be. Therefore, the final coalitional structure can be used only if a stable state is reached. According to Fig. 2 a stable state exists at the end of iterations.

to the reward function the optimal coalition structure is {{Attack: Agent0, Agent3}, {Midfielder: Agent1}, {Defense: Agent2}}.

Again the VPI algorithm does best among other algorithm. It accumulates 85% of the optimal behavior's reward. The final coalition structure of this algorithm is as same as the best coalition structure but the VPI gathers less reward in these 500 steps against the optimal behavior. The resulting coalition structure shows that, for example, having Pass and Shoot skills at their High values is good for doing midfielder tasks. Having a single type Midfielder, this reasoning could not be done.

MAP algorithm could not form the best coalition structure. Agent0 has formed a coalition with agent1 rather than Agent3. The reason can be of its persistence on exploring more informational states. However its performance was 74% of the best performance.

Additionally, because this algorithm reduces the complexity and the time needed to solve the problem, through assuming the true values based on the current belief state, the gathered reward and formed coalitions are worth

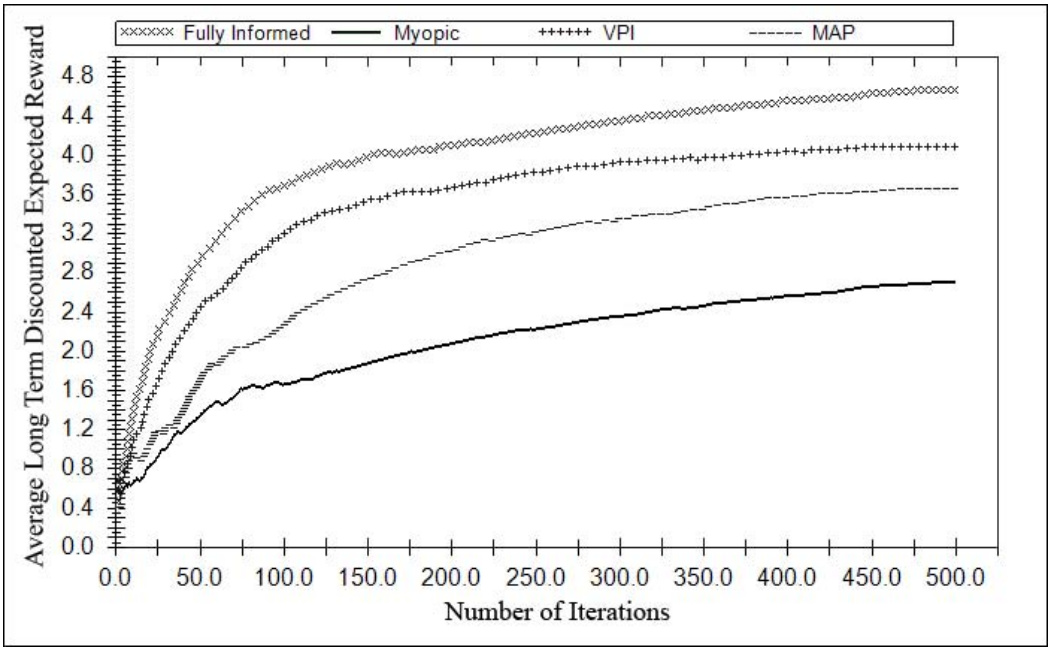

Figure 1. Average Total Discounted Accumulated Reward of Introduced Algorithms. The performance of VPI algorithm outweighs the others.

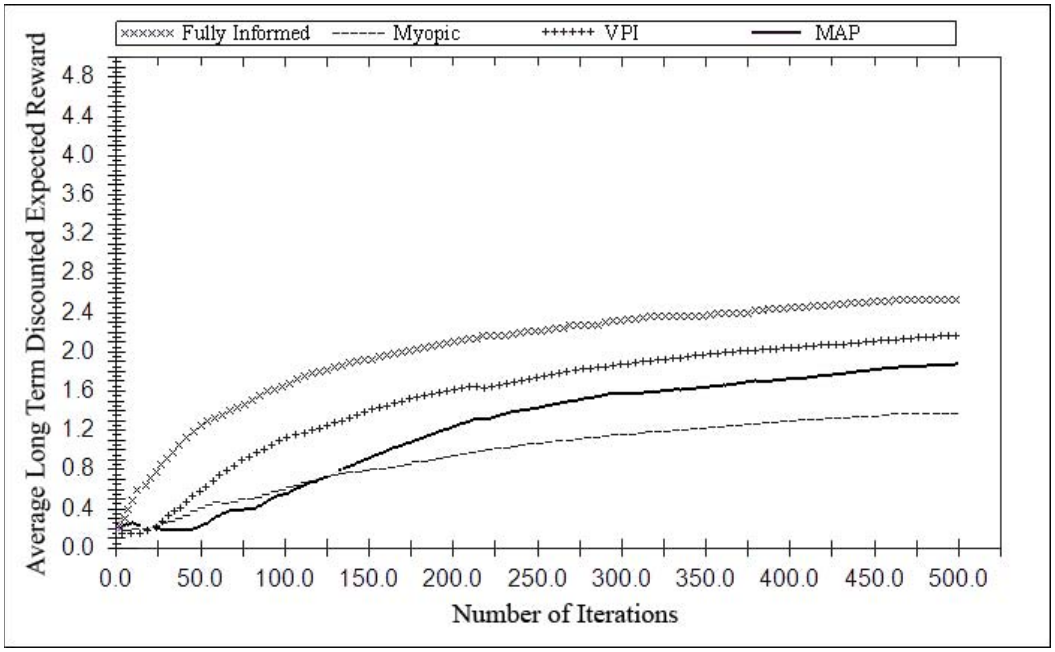

Figure 2. Learning to play soccer by multi-skill agents. VPI method does the best again.

enough. Myopic algorithm didn't do well and its performance was 54% of the best performance. The final structure of this algorithm is as same as the best structure, except Agent0 and Agent3 formed separated coalitions. Myopic agents in this algorithm didn't know joining together at the future states can help them to gather more reward.

VI. Related Works And Discussion

There are so much attention to the problem of coalition formation these days [7, 9]. This problem arises in circumstances in which there are agents that are unable to do their task individually and need to form coalitions with other agents to do those tasks.

In this situation agents need to learn who is good enough to form coalition with. In fact, there is uncertainty about the capabilities of possible partners for a typical agent [1-4, 8]. Despite the existence of this uncertainty, agents have to learn to act optimal in the environment through a repeated coalition formation process. Among all learning approach, Reinforcement Learning [12] is the best approach that matches the requirement of repeated coalition formation process.

Through the repeated use of RL algorithms, agents can form more rewarding coalitions [5]. Because forming coalitions are not under full control of a single agent, a more advanced RL method is needed that works over belief state MDPs [10-11].

Assuming multi-skills for an agent [6] can leads to solve more problems. Additionally we can reason about the obtained results more exactly. This is the most important contribution of this paper.

VII. Conclusions

Through the use of multi-skill agents we can model more problems to be solved in a repeated coalition formation process. We extend the repeated coalition formation under type uncertainty problem, and present a framework in which there are multiple skills. All agents have a value of all existing skills. Extending Myopic, VPI and MAP algorithms, we could test our framework. The results show that the VPI algorithm can do near optimal through the use of its internal heuristic.

Being so much time consuming and designing new heuristics that reduce the need to explore unrewarding states can be two areas of study for future works.

References

[1] G. Chalkiadakis, "A bayesian approach to multiagent reinforcement learning and coalition formation under uncertainty," PhD Thesis, University of Toronto, 2007.

[2] G. Chalkiadakis and C. Boutilier, "Bayesian Reinforcement Learning for Coalition Formation under Uncertainty," in Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, New York, New York, 2004, pp. 1090-1097.

[3] G. Chalkiadakis and C. Boutilier, "Coordination in multiagent reinforcement learning: a Bayesian approach," in Proceedings of the second international joint conference on Autonomous agents and multiagent systems, Melbourne, Australia, 2003, pp. 709-716.

[4] G. Chalkiadakis and C. Boutilier, "Sequential decision making in repeated coalition formation under uncertainty," in Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, Estoril, Portugal, 2008, pp. 347-354.

[5] R. Dearden, et al., "Model-Based Bayesian Exploration," in Proceedings of the Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99), San Francisco, CA, 1999, pp. 150-15.

[6] S. Ieong and Y. Shoham, "Multi-attribute coalitional games," in Proceedings of the 7th ACM conference on Electronic commerce, Ann Arbor, Michigan, USA, 2006, pp. 170-179.

[7] J.-G. Jiang, et al., "Multi-task Coalition Parallel Formation Strategy Based on Reinforcement Learning," Acta Automatica Sinica, vol. 34, pp. 349-352, 2008.

[8] S. Kraus, et al., "Coalition formation with uncertain heterogeneous information," in Proceedings of the second international joint conference on Autonomous agents and multiagent systems, Melbourne, Australia, 2003, pp. 1-8.

[9] R. A. Krzysztof and W. Andreas, "A Generic Approach To Coalition Formation," International Game Theory Review (IGTR), vol. 11, pp. 347-367, 2009.

[10] P. Poupart and N. Vlassis, "Model-based Bayesian Reinforcement Learning in Partially Observable Domains," in Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM), Fort Lauderdale, Florida, 2008.

[11] P. Poupart, et al., "An analytic solution to discrete Bayesian reinforcement learning," in Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, 2006.

[12] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction: MIT Press, 1998.