Marcos Teóricos del Aprendizaje por Refuerzo Multiagente
Limitaciones y perspectivas

Marcelo Luis Errecalde
merreca@unsl.edu.ar
Universidad Nacional de San Luis
San Luis. Argentina

1. Introducción

El Aprendizaje por Refuerzo (en inglés Reinforcement Learning y de ahora en más AR) ataca el problema de aprender a controlar agentes autónomos, mediante interacciones por prueba y error con un ambiente dinámico desconocido, el cual le provee señales de refuerzo por cada acción que realiza.
Si los objetivos del agente están definidos por la señal de refuerzo inmediata, la tarea del agente se reduce a aprender una estrategia de control (o política) que permita maximizar la recompensa acumulada a lo largo del tiempo (ver [11] para una formalización de esta tarea).
El AR ha demostrado una considerable eficacia en la resolución de problemas prácticos como robótica y manufacturación industrial, permitiendo encontrar políticas de control óptimas en escenarios de aprendizaje en línea con un único agente. Otro aspecto relevante que ha suscitado el creciente interés en este área, es el hecho de que el AR está basado en un modelo matemático formal conocido como *Proceso de Decisión Markoviano (MDP)* que ha permitido no sólo una formalización del problema a resolver y de su solución, sino también la integración con otras áreas de Inteligencia Artificial que toman a los MDP´s como uno de sus modelos formales subyacente como por ejemplo decision-theoretic planning.
Por otra parte, en el área de Sistemas Multiagentes (SMA), gran parte de los esfuerzos estuvieron dirigidos a resolver problemas de coordinación entre agentes asumiendo que existía un conocimiento adecuado del dominio e información compartida entre los agentes. Dado que existen muchos dominios en SMA en que los agentes conocen poco sobre los otros agentes y el ambiente cambia en forma dinámica, el AR recibió un importante interés como nueva de técnica de coordinación para este tipo de situaciones [10], debido a que no necesita un modelo del ambiente y puede ser utilizado en línea.
En este sentido, este trabajo analiza las limitaciones del marco teórico del AR con un único agente (MDP) cuando aplicado en SMA´s. Se describen además algunas extensiones a este formalismo surgidas del área de teoría de juegos y finalmente se presenta nuestra línea actual de investigación, orientada a aplicar estos nuevos formalismos en dominios no explorados aún por el AR en SMA, como por ejemplo el uso del AR para coordinar agentes que defienden sus intereses personales.

2. Limitaciones del uso de MDP´s como marco teórico para sistemas Multiagentes

Un MDP es un modelo matemático que explícitamente considera la incertidumbre en las acciones del agente (considerando que los efectos de las acciones están representadas por probabilidades de transición entre estados) y asume que los efectos de estas acciones son perfectamente observables (cada estado del problema contiene toda la información relevante para acciones subsecuentes).
Un MDP, implícitamente asume que las probabilidades de transición permanecen inalteradas durante el transcurso del tiempo (el ambiente del agente es estacionario).
Por otra parte, para cualquier MDP siempre hay una política π: $S \rightarrow A$ óptima, que permite decidir en cada estado que acción tomar de manera tal de maximizar la suma esperada de refuerzos descontados. Esta política π es *estacionaria* (no cambia en función del tiempo) y *determinística* (siempre se elige la misma acción cuando se está en el mismo estado).
Con respecto a la asumción de ambiente estacionario, podemos observar que este marco matemático es inapropiado para SMA´s, en particular en aquellas situaciones donde el ambiente contiene otros agentes adaptativos.
A nuestro criterio sin embargo, la principal limitación de los MDP`s como modelo de decisión

subyacente para SMA´s, surge de las características de lo que es considerado como solución para MDP`s. En estos casos, siempre existe una política óptima determinística que *no es dominada* por ninguna otra política. Este criterio es útil cuando las acciones de un agente no son influenciadas por las acciones de los otros agentes. Sin embargo, cuando la utilidad de las acciones de un agente depende directamente de las acciones de los agentes restantes, uno debe considerar qué acciones constituyen la *mejor respuesta* ante las acciones de los otros agentes. En este sentido, el principal aporte a este problema proviene del área de teoría de juegos, en lo que se conoce como *equilibrio de mejor respuesta* o *equilibrio Nash.*. Este enfoque plantea como posible solución a un juego, a la colección de estrategias para cada uno de los jugadores, tal que la estrategia de cada jugador es la mejor respuesta a las estrategias de los otros jugadores. De esta manera, ningún jugador tiene incentivo para desviarse de su estrategia en la medida que los otros jugadores tampoco se desvían.
Uno podría pensar, que el problema de AR se reduce en estos casos a lograr que el agente aprenda la política π: $S \rightarrow A$, tal que las políticas aprendidas en su conjunto constituyan uno de los posibles equilibrios Nash. Desde ese punto de vista, no deberíamos alejarnos demasiado del concepto de solución para un MDP. El problema surge de que en muchos juegos y SMA´s no existen equilibrios Nash si nos restringimos a que los agentes selecciones sus acciones en forma determinística. En otras palabras, si un agente sigue una política determinística, esto puede ser explotado por otros agentes que defiendan sus intereses personales.
Una solución a este problema, es que las políticas de los agentes sean *estocásticas*, tal que ahora una política para el agente *i* se define como $\rho : S \rightarrow \mathrm{PD}(A_i)$, que mapea estados a *estrategias mixtas*, las cuales son distribuciones de probabilidad sobre las acciones del agente.
La idea de que las políticas óptimas puedan ser estocásticas tal vez resulte extraño a las personas familiarizadas con MDP´s o algunos juegos con movimientos alternados ya que en estos casos siempre hay una política determinística superior a la mejor política probabilística. Es a partir de la incertidumbre del movimiento actual del oponente que surge la necesidad de una elección de acción probabilística que evite ser "adivinado" en una segunda instancia.

3. Extensiones de MDP´s para Aprendizaje por Refuerzo Multiagente

La aplicación más sencilla del AR en SMA consiste en trasladar sin modificación los algoritmos de AR para un único agente (ver [7] para un survey de estas técnicas) y utilizarlos en cada agente individual del SMA. En este enfoque los agentes, denominados *aprendices independientes* [4], aprenden la utilidad de sus propias acciones individuales ignorando la existencia de los otros agentes, considerándolos implícitamente como parte integrante del ambiente. En este caso, se asume que los refuerzos y las transiciones son Markovianas y estacionarias, hipótesis que como se explicó previamente no se satisface en los SMA.
Por otra parte, estos algoritmos resuelven el supuesto MDP subyacente aprendiendo una política óptima determinística, por lo que se torna evidentemente inválido en aquellas situaciones donde el equilibrio es una política mixta.
Otros enfoques más realistas, parten de reconocer las limitaciones de los MDP´s como modelo formal subyacente para SMA´s y plantean modelos que explícitamente extienden las ideas de MDP´s cuando existen varios agentes que interactúan en un mismo ambiente. Sin lugar a dudas el formalismo que mayor adhesión ha logrado en este sentido, es el denominado en la comunidad de teoría de juegos como *Juegos Estocásticos* (JE).
Los JE son una extensión muy natural de MDP´s a múltiples agentes, ya que explícitamente consideran a todos los agentes del SMA, la función de transición de estados se define en función del espacio de acciones conjuntas de los agentes y existe una función de refuerzo para cada uno de los integrantes del SMA.
Dos ejemplos relevantes de este enfoque son los trabajos de Boutellier en [1] y de Littmann en [8]. En ambos casos, se consideran dos especializaciones de JE denominadas *multiagent Markov decision process (MMDP)* por Boutellier y *Juegos Markov (JM)* por Littmann. En este sentido los MMDP se restringen a juegos completamente colaborativos (igual matriz de pagos para todos los agentes) y los JM a juegos con agentes con intereses diametralmente opuestos (Juegos de suma cero).
Recientemente se han propuesto algoritmos de AR para JE que aprenden equilibrios mixtos en juegos

de suma general [2,6]. En estos casos, si bien los agentes juegan estrategias mixtas, no existen demostraciones de convergencia a un equilibrio, a menos que se cumplan ciertas restricciones en el juego.

4. Conclusiones, estado de avance y trabajo futuro

Reconocer las limitaciones de los MDP´s como marco teórico para el AR en SMA fue un paso importante que condujo al análisis de otros formalismos que consideraran explícitamente a los otros agentes de un SMA.
En este sentido, importantes aportes se han obtenido del área de teoría de juegos en general y Juegos Estocásticos en particular. Esto ha redundado en una serie de trabajos que han comenzando ha considerar la solución a los problemas en un SMA basándose en la convergencia a equilibrios Nash mixtos como una alternativa a políticas óptimas determinísticas. Sin embargo, salvo contadas excepciones, estos formalismos se han restringido a escenarios completamente cooperativos o completamente antagónicos, dejando fuera de consideración aquellos casos en que los agentes defienden sus intereses personales pero necesitan cooperar con otros agentes en la ejecución de una tarea compartida o la resolución de un conflicto.
Nuestro trabajo de investigación se propone analizar el uso del AR en estos casos, pero en lugar de considerar a los estados de un JE como juegos de suma general, visualizarlos como una situación bargaining [9] donde los agentes pueden coordinar sus acciones para obtener un beneficio mutuo pero tienen un conflicto de intereses sobre cual de los posibles acuerdos elegir.
La idea consiste en determinar de que manera estos agentes pueden coordinar sus acciones en forma dinámica sin los requerimientos de racionalidad completa y conocimiento completo de las preferencias de los otros agentes que es asumido en teoría de juegos clásica. Para ello, estamos estudiando cómo, a partir de una adecuada definición del protocolo de interacción entre los agentes y de la función de recompensa para los agentes, se puede converger a una solución que satisfaga los axiomas de Nash en situaciones bargaining que, como ha sido demostrado [9] maximiza el producto de las utilidades de los agentes.
Dado que en estos casos, el conjunto de posibles soluciones es convexo y contínuo, lo cual implica trabajar con políticas mixtas, estamos actualmente analizando técnicas de aprendizaje para jugar estrategias mixtas utilizadas en el área de aprendizaje en juegos, como por ejemplo *stochastic ficitious play* [5]

5. Referencias

[1] C. Boutellier. "Planning, learning and coordination in multiagent decision processes". Proceedings of the Sixth Conference on the Theoretical Aspects of Rationality and Knowledge, 195-210. Amsterdam, 1996.
[2] M. Bowling y M. Veloso. "Rational Learning of Mixed Equilibriain in Stochastic Games". Tech. Report.
[3] M. Bowling y M. Veloso. "An analysis of Stochastic Game Theory for Multiagent Reinforcement Learning". Tech. Report.
[4] C. Claus y C. Boutellier. "The dynamics of reinforcement learning in cooperative multiagent systems". In Proceedings of the Fifteenth National Conference on Artificial Intelligence, 1998.
[5] D. Fudenberg y D. K. Levine. "The theory of learning in games". The MIT Press.
[6] J. Hu y M. P. Wellman. "Multiagent reinforcement learning: Theoretical framework and an algorithm. In In Proceedings of the Fifteenth International Conference on Machine Learning, 242-250. San Francisco, 1998.
[7] L. P.Kaelbling, M. Littman y A. Moore. "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research 4 (1996) - 237-285 - Mayo 1996.
[8] M. L. Littmann. "Markov games as a framework for multi-agent reinforcement learning". In Proceedings of the Eleventh International Conference on Machine Learning, 157-163. New Brunswick, 1994.
[9] M. Osborne y A. Rubinstein. "Bargaining and Markets". Academic Press. 1990
[10] S. Sen y G. Weiss. "Learning in Multiagent Systems". Chapt. 6 in [12]
[11] R.Sutton y A. Barto. "Reinforcement Learning: an introduction". The MIT Press, 1998.
[12] G. Weiss (Ed.). "Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence". The MIT Press