Adversarial Risk Analysis for Bi-Agent Influence Diagrams

Jorge González Ortega

David Ríos Insua, Javier Cano

Resumen (SP)

En su artículo de referencia, Shachter (1986) propuso la extensión al caso multi-agente de la computación de decisiones óptimas en diagramas de influencia como un problema importante. Hasta la fecha, esta sugerencia se ha afrontado desde la perspectiva de la teoría de juegos (no-cooperativos) a raíz del trabajo de Koller y Milch (2003). En el mismo, los autores introdujeron los Diagramas de Influencia Multi-Agente (MAIDs por sus siglas en inglés) y elaboraron algoritmos para la búsqueda de equilibrios de Nash en problemas generales de la teoría de juegos no-cooperativos.

Un inconveniente clave de esta metodología es la fuerte hipótesis de conocimiento común subyacente, criticada por ejemplo en Raiffa et al. (2002) o Lippman y McCardle (2012). Dicha hipótesis permite un análisis normativo conjunto de carácter simétrico en el que cada jugador maximiza su utilidad esperada (y espera que el resto de jugadores haga lo mismo). De esta forma, las decisiones se pueden anticipar y son capturadas por equilibrios de Nash y conceptos relacionados. No obstante, en muchos contextos incluyendo la lucha contra el terrorismo o la ciberseguridad, los jugadores no suelen tener tamaño conocimiento acerca de sus oponentes. Es más, el problema se agrava dado que los agentes implicados tratan de ocultar información.

El Análisis de Riesgos de Adversarios (ARA) ofrece una vía alternativa en la que no es necesaria una hipótesis de conocimiento común. Apoyando a uno de los agentes al que se denomina Defensora (Ella), se aborda su problema como un análisis clásico de decisión, pero utilizando procedimientos que emplean la estructura del juego teórico, y otra información disponible, para estimar las probabilidades de cada posible acción del oponente (llamado Atacante, Él). El caso descrito es bi-agente, pero de forma natural se extiende la metodología al caso multi-agente general.

Una de las principales motivaciones para el desarrollo del enfoque ARA surge en estudios en las áreas de seguridad y lucha contra el terrorismo. Casos relacionados con la protección ante amenazas inteligentes incluyen la prevención de barcos frente a los riesgos de la piratería (Sevillano et al., 2012) o la defensa contra artefactos explosivos improvisados en problemas de enrutamiento (Wang y Banks, 2011). Para una revisión más amplia, véase Banks et al. (2015).

Éstas y otras aplicaciones se han tratado con modelos ARA relativamente simples, con secuencias básicas de movimientos de ataque y defensa. De hecho, se pueden identificar una serie de plantillas que pueden considerarse como bloques de construcción básicos para problemas generales de análisis de riesgos de seguridad, ver Ríos y Ríos Insua (2012). Dichas plantillas difieren en la forma y el orden en que los movimientos de ataque y defensa se llevan a cabo dentro de la secuencia global de decisiones y eventos, así como en la información disponible para cada agente. En particular, cinco plantillas básicas, con nombres auto-explicativos, están analizadas al detalle en Ríos y Ríos Insua (2012) y Ríos Insua et al. (2013): el modelo secuencial de Defensa-Ataque; el modelo simultáneo de Defensa-Ataque; el modelo secuencial de Ataque-Defensa; el modelo secuencial de Defensa-Ataque-Defensa y, finalmente, el modelo secuencial de Defensa-Ataque con información privada.

Más allá de estas plantillas, en nuestro trabajo tenemos en cuenta problemas generales de confrontación entre dos agentes en los que se permiten interacciones más complejas entre ellos, típicamente consistentes en movimientos secuenciales y simultáneos entremezclados que abarcan diferentes períodos de planificación. Nuestro objetivo es apoyar a uno de los agentes, la Defensora, en su toma de decisiones, para lo cual debe prever las intenciones del Atacante. Suponiendo que este agente maximiza su utilidad esperada, podemos predecir sus acciones encontrando aquéllas de máxima utilidad esperada. La incertidumbre en nuestras apreciaciones sobre las probabilidades y utilidades del Atacante se propaga a la decisión óptima aleatoria que constituye nuestro pronóstico acerca del ataque. Para resolver estos problemas generales de confrontación recurrimos a la capacidad para modelar interacciones complejas de los MAIDs aprovechando el concepto de relevancia estratégica definido por Koller y Milch (2003), pero relajando la hipótesis de conocimiento común a través de la metodología ARA.

El documento está estructurado de la siguiente manera. En la sección 2, se presentan los problemas a tratar y desarrollamos el concepto de Diagrama de Influencia Bi-Agente (BAID por sus siglas en inglés) basado en los MAIDs introducidos por Koller y Milch (2003). Así mismo, se muestra la representación gráfica de los BAIDs, así como su descomposición en diagramas de influencia propios para ambos agentes (Defensora y Atacante). Para ilustrar los conceptos, hacemos uso de un ejemplo asociado con la Protección de Infraestructuras Críticas (PIC).

La sección 3 presenta las principales características computacionales de nuestra propuesta aplicadas al ejemplo PIC. En primer lugar, generamos una solución al problema de la forma más natural. Ésta se corresponde con resolver el problema de la Defensora tanto como es posible hasta que se requiere algún tipo de evaluación del problema del Atacante, cambiando así entre ambos problemas en términos de los requerimientos del problema de la Defensora sobre la base de una estrategia anticipativa de nivel 2 de profundidad. Después, se describe un esquema de resolución más general que ejemplifica el procedimiento final propuesto.

La metodología general se aborda en la sección 4. Se comienza con una revisión de las operaciones de reducción de los BAIDs incorporando la incertidumbre sobre las probabilidades y utilidades del Atacante a las reducciones de su problema. A continuación, se explican brevemente los conceptos de relevancia. Por último, se proporciona un algoritmo ARA para el apoyo a un agente en un problema general

modelado como un BAID distinguiendo los casos en que el grafo de relevancia es acíclico o cíclico.

Finalmente, en la sección 5 aplicamos el desarrollo metodológico al conocido problema del envenenador de árboles establecido por Koller y Milch (2003) y que adaptamos a un escenario PIC. Concluimos con una discusión.

Palabras Clave:

Juegos no-cooperativos, Análisis de la decisión, Análisis de riesgos adversarios, Diagramas de influencia bi-agente, Relevancia, Protección de infraestructuras críticas.

Abstract (EN)

In his landmark paper, Shachter (1986) proposed extending the computation of optimal decision policies in influence diagrams to the multi-agent case as an important problem. So far, this suggestion has been faced from a (non-cooperative) game theoretic perspective, stemming from Koller and Milch (2003) who introduced Multi-Agent Influence Diagrams (MAIDs) and provided algorithms for finding Nash equilibria in general non-cooperative game-theoretic problems.

A main drawback of such methodology is its underlying common knowledge assumption, criticised in e.g. Raiffa et al. (2002) or Lippman and McCardle (2012). Such assumption allows for a symmetric joint normative analysis in which players maximise their expected utilities (and expect other players to do the same). Their decisions can be anticipated and are predated by Nash equilibria and related concepts. However, in many contexts, including counter-terrorism or cybersecurity, players will not typically have such knowledge about their opponents. This is aggravated as participants try to conceal information.

Adversarial Risk Analysis (ARA) provides a way forward, as common knowledge is not required. In supporting one of the participants, which we call the Defender (She), we view her problem as a decision analytic one, but procedures which employ the game-theoretical structure, and other information available, are used to estimate the probabilities of the opponent's (called the Attacker, He) actions. The described case is the bi-agent one, but the methodology extends itself naturally to the general multi-agent case.

A main motivation for ARA developments arises from security and counter-terrorism studies. Cases dealing with protection from intelligent threats include preventing ships from piracy risks (Sevillano et al., 2012), or anti-IED defence in routing problems (Wang and Banks, 2011). For a broad review, see Banks et al. (2015).

These and other applications have been dealt with relatively simple ARA models, with basic sequences of defence and attack movements. Indeed, we can identify a number of templates which may be viewed as basic building blocks for general security risk analysis problems, see Ríos and Ríos Insua (2012). They differ in the way and order in which attack and defence movements take place within the global sequence of decisions and events, as well as in the information revealed. In particular, five basic templates,

with self-explanatory names, are covered in full detail in Ríos and Ríos Insua (2012) and Ríos Insua et al. (2013): the sequential Defend-Attack model; the simultaneous Defend-Attack model; the sequential Attack-Defend model; the sequential Defend-Attack-Defend model and, finally, the sequential Defend-Attack model with private information.

Beyond these templates, we consider here general adversarial problems between two agents in which we allow for more complex interactions between them, typically consisting of intermingled sequential and simultaneous movements, spanning across different planning periods. Our aim is to support one of the agents, the Defender, in her decision making. For that, she needs to forecast the Attacker's intentions. Assuming that this agent is an expected utility maximiser, we can predict his actions by finding his maximum expected utility action. The uncertainty in our assessments about the Attacker's probabilities and utilities propagates to his random optimal decision which constitutes our required attack forecast. We solve general adversarial problems using the MAIDs ability to model complex interactions, taking advantage of the concept of strategic relevance (Koller and Milch, 2003), but relaxing the common knowledge assumption through the ARA methodology.

The paper is structured as follows. In Section 2, we present the problems we shall be dealing with and develop the concept of Bi-Agent Influence Diagram (BAID) based on MAIDs introduced by Koller and Milch (2003). Graphic representation of BAIDs is shown as well as their decomposition into proper ID's for both agents (Defender and Attacker). To illustrate BAIDs, we make use of a driving example associated with Critical Infrastructure Protection (CIP).

Section 3 presents the key computational features of our proposal applied to the CIP example. We produce a solution to the problem in the natural way, which corresponds to solving as much of the Defender's problem as possible until some assessment from the Attacker's problem is required, thus switching between both of them in terms of the Defender's problem's requirements based on a level-2 thinking strategy. At the end of the section, we describe a more general solution scheme which exemplifies the proposed procedure.

The general methodology is approached in Section 4. It begins with a review of BAID reduction operations incorporating the uncertainty about the Attacker's probabilities and utilities to the reductions in his problem. Relevance concepts are also briefly explained. Finally, an algorithm for ARA support to an agent in a general problem modeled as a BAID is provided distinguishing the cases in which the relevance graph is acyclic or cyclic.

Lastly, Section 5 applies the developed methodology to the well known Tree Killer Problem established in Koller and Milch (2003) adapted to a CIP scenario. We end up with a discussion.

Keywords:

Non-cooperative games, Decision analysis, Adversarial risk analysis, Bi-agent influence diagram, Relevance, Critical Infrastructure Protection.

Referencias

D. Banks, J. Ríos and D. Ríos Insua. *Adversarial Risk Analysis*. CRC Press, 2015.

D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1): 181-221, 2003.

S. Lippman and K. McCardle. Embedded Nash bargaining: Risk aversion and impatience. *Decision Analysis*, 9(1): 31-40, 2012.

H. Raiffa, J. Richardson and D. Metcalfe. Negotiation *Analysis: the Science and Art of Collaborative Decision Making*. Harvard University Press, 2002.

J. Ríos and D. Ríos Insua. Adversarial risk analysis for counterterrorism modeling. *Risk Analysis*, 32(5): 894-915, 2012.

D. Ríos Insua, J. Cano, W. Shim, F. Massacci and A. Schmitz. SECONOMICS "Socio-Economics meets Security". Deliverable 5.1. Basic Models for Security Risk Analysis. *Technical report, European Union*, 2013.

J. C. Sevillano, D. Ríos Insua and J. Ríos. Adversarial risk analysis: The Somali pirates case. *Decision Analysis*, 9(2): 86-95, 2012.

R. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6): 871-882, 1986.

S. Wang and D. Banks. Network routing for insurgency: An adversarial risk analysis framework. *Naval Research Logistics*, 58(6): 595-607, 2011.