Análisis de partidas de ajedrez mediante Data Mining

Rubén Márquez
Minería de datos
Universidad de Castilla-La Mancha
Ruben.Marquez@alu.uclm.com

Alberto Velasco
Minería de datos
Universidad de Castilla-La Mancha
Alberto.Velasco1@alu.uclm.com

Diego Pedregal
Minería de datos
Universidad de Castilla-La Mancha
Diego.Pedregal@alu.uclm.com

Antonio Manjavacas
Minería de datos
Universidad de Castilla-La Mancha
Antonio.Manjavacas@alu.uclm.com

1 Introducción

La minería de datos se define como un proceso utilizado para extraer, analizar e identificar patrones a partir de conjuntos masivos de datos. Un área donde esta disciplina ha alcanzado una gran relevancia en las últimas décadas ha sido en el contexto del ajedrez, donde, desde *DeepBlue* [1], se han elaborado múltiples sistemas de inteligencia artificial basados en el análisis de partidas con distintos fines: jugar contra humanos, jugar contra otras máquinas, predecir resultados, analizar partidas, etc.

En base a la utilidad de la minería de datos en el contexto del ajedrez, hemos decidido aplicar técnicas de extracción de conocimiento sobre datos de partidas disponibles en la web.

En este documento se definirán los aspectos iniciales del problema a abordar: fuente de datos, descripción de estos y objetivos que se pretenden conseguir.

2 Fuente de datos

Para realizar este proyecto se ha elegido como fuente de datos el conjunto de partidas de ajedrez de la web *lichess.org* [2] correspondientes al mes de **septiembre de 2019**. De este conjunto de partidas se han obtenido las partidas correspondientes a los 200 mejores jugadores de la clasificación en modo clásico.

La extracción selectiva de los datos por fecha y mejores jugadores se ha llevado a cabo haciendo uso de la API `berserk` [3]. Todos los datos recopilados son de dominio público sin derechos reservados.

3 Descripción de los datos

El formato de cada uno de los registros del dataset es el siguiente:

- *Event*: evento en el cual se enmarca la partida.
- *Site*: enlace a la partida.
- *White*: jugador con blancas.
- *Black*: jugador con negras.
- *Result*: resultado de la partida.
- *UTCDate*: fecha UTC.
- *UTCTime*: hora UTC.
- *WhiteElo*: puntuación ELO del jugador con blancas.
- *BlackElo*: puntuación ELO del jugador con negras.
- *WhiteRatingDiff*: puntos obtenidos por blancas.
- *BlackRatingDiff*: puntos obtenidos por negras.
- *WhiteTitle*, *BlackTitle*: abreviación de la titulación de los jugadores (según estándar FIDE).
- *ECO*: código estándar de la apertura utilizada.
- *Opening*: apertura utilizada.
- *TimeControl*: control de tiempo utilizado.
- *Termination*: motivo de victoria.
- Lista de *jugadas* incluyendo evaluación y tiempo de reloj para cada movimiento.

Aunque estos son los campos que se recogen en la mayoría de los datasets en ajedrez, existen campos adicionales permitidos por el formato PGN (Portable Game Notation) utilizado. Información más detallada sobre cada uno de los campos puede consultarse en http://www.saremba.de/chessgml/standards/pgn/pgn-complete.htm#c9.1.1.

4 Formulación de hipótesis

Partiendo del conjunto de datos elegido, se pretenden contrastar las siguientes hipótesis:

H1) Es posible pronosticar el resultado de un torneo de ajedrez en base a la información disponible acerca de los jugadores involucrados.

H2) Es posible definir los factores de éxito que influyen en el devenir de una partida de ajedrez entre dos jugadores.

H3) Es posible definir el estilo de juego de un jugador en base a su histórico de partidas jugadas.

5 Objetivos

En base a este conjunto de datos, se pretenden abordar los siguientes objetivos aplicando las diferentes fases del **proceso KDD** [4]:

1. Caracterizar a los jugadores en base su histórico de partidas estableciendo un perfil de juego para cada uno de ellos (en base a su agresividad, aperturas, etc.).
2. Tratar de pronosticar el resultado de las partidas individuales que conforman un torneo entre unos jugadores predefinidos.

6 Target data

De cara a llevar a cabo el contraste de nuestras hipótesis, se procedió con la definición del conjunto de datos objetivo ("tarjeta de datos" o *target data*) compuesto por las siguientes características:

- *USER_ID*: identificador de usuario (`String`).
- *GAME_LINK*: enlace a la partida (`String`).
- *ELO*: puntuación ELO del jugador (`int`).
- *COLOUR*: color de piezas del usuario (`String`).
- *OPENING*: apertura (y variación) utilizada en la partida (`String`).
- *TIME_TOTAL*: duración de la partida en segundos (`double`).
- *MOVEMENTS*: número de movimientos de la partida (`int`).
- *TERMINATION*: modo de finalización de la partida (`int`):
    a. 0 = gana_blanco
    b. 1 = gana_negro
    c. 2 = tablas
- *MEAN_TIME_PER_MOVEMENT_EARLY*: tiempo medio entre movimientos en el *early game* (`double`).
- *MEDIAN_TIME_PER_MOVEMENT_EARLY:* mediana de los tiempos entre movimientos en el *early game* (`double`).
- *VAR_TIME_PER_MOVEMENT_EARLY:* varianza de tiempos entre movimientos en el *early game* (`double`).
- *MAX_TIME_PER_MOVEMENT_EARLY*: máximo tiempo empleado entre movimientos en el *early game* (`double`).
- *MIN_TIME_PER_MOVEMENT_EARLY*: mínimo tiempo empleado entre movimiento en el *early game* (`double`).
- *MEAN_TIME_PER_MOVEMENT_MID*: tiempo medio entre movimientos en el *mid game* (`double`).
- *MEDIAN_TIME_PER_MOVEMENT_MID:* mediana de los tiempos entre movimientos en el *mid game* (`double`).
- *VAR_TIME_PER_MOVEMENT_MID:* varianza de tiempos entre movimientos en el *mid game* (`double`).
- *MAX_TIME_PER_MOVEMENT_MID*: máximo tiempo empleado entre movimientos en el *mid game* (`double`).
- *MIN_TIME_PER_MOVEMENT_MID*: mínimo tiempo empleado entre movimiento en el *mid game* (`double`).
- *MEAN_TIME_PER_MOVEMENT_END*: tiempo medio entre movimientos en el *end game* (`double`).
- *MEDIAN_TIME_PER_MOVEMENT_END:* mediana de los tiempos entre movimientos en el *end game* (`double`).
- *VAR_TIME_PER_MOVEMENT_END:* varianza de tiempos entre movimientos en el *end game* (`double`).
- *MAX_TIME_PER_MOVEMENT_END*: máximo tiempo empleado entre movimientos en el *end game* (`double`).
- *MIN_TIME_PER_MOVEMENT_END*: mínimo tiempo empleado entre movimiento en el *end game* (`double`).
- *POINTS_BALANCE*: balance de puntos al final de la partida (`int`).
- *TAKEN_BALANCE*: balance de piezas al final de la partida (`int`).
- *AGGRESSIVENESS*: nivel de agresividad del jugador en el rango [0,6]. Se mide a partir de los siguientes factores (`double`):
    a. *EARLY_TAKEN*: número de piezas comidas en el *early game* (primer tercio de la partida).
        - Alto = +1
        - Medio = +0.5
        - Bajo = +0
    b. *SACRIFICES*: número de sacrificios a largo plazo realizados (más de 5 turnos):
        - Alto = +1
        - Medio = +0.5
        - Bajo = +0
    c. *AGRESSIVE_OPENING*: la apertura utilizada es agresiva para el color de piezas del jugador:

- True = +2
- False = +0

d. *CASTLING*: el jugador se enroca a lo largo de la partida:
    - True = +0
    - False = +2

REFERENCIAS

[1] Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. Artificial Intelligence, 134(1-2), 57–83. doi:10.1016/s0004-3702(01)00129-1

[2] https://database.lichess.org/

[3] https://berserk.readthedocs.io/en/master/

[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27–34. doi:10.1145/240455.240464