Análisis de Reglas de Asociación en Partidas de Ajedrez Online: Aplicación del Algoritmo Apriori en Datos de Lichess

Mario Marín Hinojosa, Alberto Bartolomé Iruela

No Institute Given

Resumen Este trabajo presenta un análisis exhaustivo de reglas de asociación aplicado a partidas de ajedrez online utilizando el algoritmo Apriori. Se analizaron 121,332 partidas del portal lichess.org correspondientes a enero de 2013, con especial énfasis en el subconjunto de ajedrez relámpago (600+0). El estudio implementó un sistema de categorización de variables que incluye niveles de Elo, duración de partidas y características de juego. Los resultados revelan patrones significativos como la fuerte correlación entre partidas muy largas y resultados en tablas (lift=7.39), y la alta predictibilidad de victoria cuando existe una diferencia de 2 categorías de Elo entre jugadores (confianza=93.8 %). La metodología desarrollada demuestra la efectividad del algoritmo Apriori para descubrir conocimiento actionable en dominios deportivos complejos.

Keywords: Reglas de asociación Algoritmo Apriori Ajedrez online Minería de datos Análisis deportivo

1. Introducción

El ajedrez, como uno de los juegos estratégicos más estudiados en la historia, genera enormes volúmenes de datos estructurados que proporcionan oportunidades únicas para el análisis mediante técnicas de minería de datos. Con el auge de las plataformas de ajedrez online como lichess.org, se ha creado un ecosistema digital que registra millones de partidas con información detallada sobre jugadores, movimientos y resultados.

Las reglas de asociación, introducidas originalmente por Agrawal et al. [1], han demostrado ser especialmente efectivas para descubrir patrones frecuentes en datasets transaccionales. El algoritmo Apriori, basado en el principio de que todos los subconjuntos de un conjunto frecuente también son frecuentes, permite identificar relaciones significativas entre variables categóricas.

Este estudio tiene como objetivo principal aplicar el algoritmo Apriori para descubrir patrones de juego en partidas de ajedrez online, con especial foco en:

- Identificación de reglas de asociación significativas entre características de partida
- Verificación estadística de hipótesis específicas sobre patrones de juego

- Análisis comparativo entre diferentes modalidades temporales
- Evaluación de la predictibilidad de resultados basada en características de jugadores

2. Metodología

2.1. Dataset y Preprocesamiento

El dataset utilizado corresponde a partidas de ajedrez de lichess.org del mes de enero de 2013, conteniendo 121,332 registros con 11 variables originales. Las variables incluyen información sobre jugadores (nombres, puntuaciones Elo), características de partida (resultado, control de tiempo, número de movimientos) e información técnica (código ECO, apertura, tipo de finalización).

El preprocesamiento incluyó los siguientes pasos:

1. **Tratamiento de valores especiales**: Los valores ¿.$^{\text{en}}$ las puntuaciones Elo fueron reemplazados por 900 puntos, siguiendo las especificaciones del proyecto.
2. **Categorización de Elo**: Se implementó un sistema de categorización basado en rangos estándar internacionales:
   - Principiante: 0-1199
   - Intermedio: 1200-1599
   - Avanzado: 1600-1999
   - Experto: 2000-2399
   - Maestro: 2400-2799
   - Gran Maestro: 2800+
3. **Categorización de duración**: Las partidas fueron clasificadas según número de movimientos:
   - Corta: ¡20 movimientos
   - Media: 20-39 movimientos
   - Larga: 40-59 movimientos
   - Muy larga: 60+ movimientos
4. **Variables derivadas**: Se calcularon variables adicionales como diferencia absoluta de Elo, diferencia en categorías de Elo y identificación del jugador más fuerte.

2.2. Selección del Subconjunto de Análisis

Para el análisis principal se seleccionó el subconjunto de partidas con control de tiempo 600+0 (ajedrez relámpago), que representa 2,452 partidas (2.02 % del dataset total). Esta modalidad fue elegida por:

- Equilibrio entre tiempo de reflexión y presión temporal
- Representatividad significativa en el dataset
- Relevancia en el ajedrez online moderno

Las características del subconjunto muestran:

- Distribución de resultados: 49.2 % victorias blancas, 47.9 % victorias negras, 2.9 % tablas
- Elo promedio: 1553 (blancas), 1552 (negras)
- Promedio de movimientos: 32.9
- Diferencia Elo promedio: 145 puntos

2.3. Implementación del Algoritmo Apriori

La implementación siguió los siguientes pasos:

1. **Preparación transaccional**: Conversión de variables categóricas a formato de transacciones, donde cada partida representa una transacción conteniendo ítems como Result_1-0", "WhiteElo_Cat_Intermedio", etc.
2. **Codificación binaria**: Utilización de TransactionEncoder para crear una matriz binaria de 2,452 × 20.
3. **Extracción de conjuntos frecuentes**: Aplicación de Apriori con soporte mínimo de 0.01, generando 873 conjuntos frecuentes.
4. **Generación de reglas**: Extracción de 9,346 reglas de asociación con confianza mínima de 0.1.

3. Resultados

3.1. Reglas de Asociación Principales

El análisis reveló reglas altamente significativas, destacando las 10 más relevantes por valor de lift:

Cuadro 1. Top 5 Reglas de Asociación por Lift

Regla	Soporte	Confianza	Lift
Partidas muy largas + Terminación normal → Tablas	0.0114	0.2171	7.39
Tablas → Partidas muy largas + Terminación normal	0.0114	0.3889	7.39
Tablas → Partidas muy largas	0.0143	0.4861	7.01
Partidas muy largas → Tablas	0.0143	0.2059	7.01
Intermedio vs Principiante → Victoria del más fuerte	0.0171	0.1567	4.42

3.2. Verificación de Hipótesis

Se verificaron sistemáticamente varias hipótesis específicas:

H1: Diferencia 1 categoría de Elo Para partidas con diferencia de al menos 1 categoría de Elo (1,014 casos de 2,452 total):

- Jugador más fuerte con blancas: Confianza = 68.7 % (349/508 casos)
- Jugador más fuerte con negras: Confianza = 68.6 % (347/506 casos)
- Clasificación: Regla moderada

H2: Diferencia 2 categorías de Elo Para diferencias más significativas (31 casos):

- Jugador más fuerte con blancas: Confianza = 93.8 % (15/16 casos)
- Jugador más fuerte con negras: Confianza = 93.3 % (14/15 casos)
- Clasificación: Regla fuerte

H3 y H4: Partidas entre Grandes Maestros No se encontraron partidas entre Grandes Maestros en el subconjunto analizado, lo que refleja la rareza de estos encuentros en modalidades de tiempo rápido durante el período estudiado.

3.3. Patrones Significativos Identificados

1. **Correlación duración-resultado**: Las partidas muy largas (60 movimientos) muestran una fuerte tendencia hacia resultados en tablas, con un lift de 7.39, indicando una correlación 7.39 veces mayor que la esperada por azar.
2. **Predictibilidad por diferencia de nivel**: Cuando existe una diferencia significativa de nivel (2 categorías), la probabilidad de victoria del jugador más fuerte supera el 93 %, estableciendo una regla altamente predictiva.
3. **Ventaja de las blancas**: En el subconjunto analizado, las blancas muestran una ligera ventaja (49.2 % vs 47.9 %), consistente con la teoría ajedrecística clásica.
4. **Terminaciones por tiempo**: Las reglas revelan patrones específicos entre el tipo de terminación y las características de los jugadores, especialmente en casos de diferencias de nivel significativas.

4. Análisis y Discusión

4.1. Interpretación de Resultados

Los resultados obtenidos confirman varios principios fundamentales del ajedrez mientras revelan patrones específicos del juego online:

- **Efecto de la diferencia de nivel**: La alta confianza (93.8 %) para diferencias de 2 categorías valida la importancia del nivel relativo de los jugadores como predictor de resultado.
- **Características de partidas largas**: El lift de 7.39 para la regla "partidas muy largas → tablas"sugiere que las partidas equilibradas tienden a prolongarse y finalizar en empate, reflejando la complejidad estratégica del ajedrez.

- **Modalidad relámpago**: Los patrones identificados son específicos de esta modalidad temporal, donde la presión de tiempo influye significativamente en las decisiones y resultados.

4.2. Limitaciones del Estudio

1. **Temporalidad de datos**: El dataset corresponde únicamente a enero de 2013, lo que puede no reflejar patrones actuales.
2. **Representatividad del subconjunto**: El análisis se focalizó en una modalidad específica (600+0), limitando la generalización a otros controles de tiempo.
3. **Ausencia de Grandes Maestros**: La falta de partidas entre GM en el subconjunto impidió verificar hipótesis relacionadas con el más alto nivel.
4. **Variables no consideradas**: Factores como apertura jugada o momento del día no fueron incluidos en el análisis principal.

4.3. Aplicaciones Prácticas

Los resultados tienen aplicaciones potenciales en:

- **Sistemas de emparejamiento**: Optimización de algoritmos de matching basados en predictibilidad de resultados.
- **Entrenamiento de jugadores**: Identificación de patrones para mejorar estrategias según el nivel del oponente.
- **Análisis de plataformas**: Comprensión de dinámicas de juego para mejorar experiencia de usuario.
- **Investigación ajedrecística**: Base empírica para estudios sobre teoría del juego y psicología competitiva.

5. Conclusiones

Este estudio demuestra la efectividad del algoritmo Apriori para descubrir patrones significativos en partidas de ajedrez online. Los principales hallazgos incluyen:

1. **Reglas altamente predictivas**: Se identificaron reglas con lift superior a 7, indicando relaciones muy fuertes entre variables.
2. **Validación de hipótesis**: La verificación sistemática confirmó la importancia de la diferencia de nivel como predictor, especialmente para diferencias 2 categorías (93.8 % de confianza).
3. **Patrones específicos de modalidad**: Los resultados revelan características únicas del ajedrez relámpago, diferenciándolo de otras modalidades temporales.
4. **Metodología escalable**: El framework desarrollado puede aplicarse a otros subconjuntos y modalidades de juego.

La investigación contribuye al campo de la analítica deportiva aplicada al ajedrez, proporcionando una base metodológica sólida para futuros estudios. Las reglas identificadas no solo tienen valor descriptivo sino también predictivo, abriendo oportunidades para aplicaciones en sistemas inteligentes de ajedrez.

5.1. Trabajo Futuro

Las direcciones de investigación futura incluyen:

- Análisis comparativo entre diferentes modalidades temporales
- Incorporación de variables temporales y secuenciales
- Aplicación de algoritmos de reglas de asociación más avanzados
- Validación de patrones en datasets contemporáneos
- Desarrollo de sistemas de recomendación basados en los patrones identificados

Referencias

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD Record, vol. 22, no. 2, pp. 207-216 (1993)
2. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco (2011)
3. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
4. Lichess.org: Chess Games Database. https://database.lichess.org/ (2013)
5. Zaki, M.J., Meira Jr, W.: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, Cambridge (2014)
6. Raschka, S.: MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Journal of Open Source Software, 3(24), 638 (2018)