Minería de reglas de asociación excepcionales extraidas con algoritmos evolutivos

José María Luna, Francisco Padillo, Sebastián Ventura

Departamento de informática y análisis numérico, Universidad de Córdoba,
Campus de Rabanales, 14071 Córdoba, España.
{jmluna,fpadillo,sventura}@uco.es,

Resumen El creciente interés en el almacenamiento de datos ha provocado que su análisis sea cada vez más variado. En cualquier conjunto de datos es posible identificar un subconjunto cuya distribución es exceptionalmente diferente a la distribución de los datos en todo el conjunto. El descubrimiento de este comportamiento excepcional permite extraer relaciones interesantes entre patrones que permiten describir, de manera diferente, la información contenida en las bases de datos. El objetivo de este artículo es proponer un modelo de extracción de reglas de asociación excepcionales, mostrar la utilidad de este tipo de modelos, así como su aplicación a un conjunto de métricas en el campo de la minería de patrones. Los resultados obtenidos han permitido describir pares de métricas que están positivamente correladas pero que, bajo ciertas condiciones, esta correlación pasa a ser negativa.

Keywords: Minería de patrones, Modelos excepcionales, Asociación

1. Introducción

En la actualidad, el creciente interés en el almacenamiento de datos está acarreando que multitud de compañías almacenen grandes cantidades de datos que luego deben ser procesados y analizados para darle un sentido. Por lo general, los datos en bruto carecen de interés, por lo que su extracción, descripción y análisis es fundamental para descubrir conocimiento que pueda ayudar en la toma de decisiones.

La extracción de patrones [1] y asociaciones entre patrones, así como su descripción y análisis, facilitan la comprensión de la información almacenada en multitud de dominios de aplicación. Un patrón es un término ampliamente utilizado en el análisis de datos, el cual representa cualquier tipo de homogeneidad y regularidad en los datos. Un patrón describe propiedades intrínsecas de los datos, por lo que su extracción juega un papel fundamental en diferentes tareas de análisis de datos. Sin embargo, un patrón *per se* no constituye todas la formas de conocimiento [2], por lo que puede no ser suficiente en algunos dominios de aplicación.

En determinadas ocasiones, puede ser posible la necesidad de extraer modelos excepcionales, es decir, grupos de patrones cuya distribución es excepcionalmente diferente a la distribución sobre el conjunto total de datos [3]. La minería de

Actas de la XVII Conferencia de la Asociación Española para
la Inteligencia Artificial, pp. 127-136

modelos excepcionales se propuso como una tarea de gran interés, descubriendo grupos cuyo comportamiento es substancialmente diferente al mismo modelo (conjunto de items) para toda la base de datos [4]. A modo de ejemplo, consideremos un conjunto de datos con información sobre personas de diferentes países con el fin de analizar la relación entre la edad y la tasa de mortandad. Considerando el conjunto total de datos (todos los países unidos), se obtiene que existe una relación positiva entre la edad y la tasa de mortandad, es decir, a mayor edad, mayor es la tasa de mortandad. Sin embargo, este comportamiento es completamente diferente si consideramos el patrón $\{pobreza,\ enfermedad\ contagiosa\}$. Considerando este patrón, se descubre que la edad no está positivamente correlada con la tasa de mortandad. De hecho, personas que viven bajo pobreza y que poseen algún tipo de enfermedad contagiosa tienen menor probabilidad de sobrevivir durante su primer año de vida, produciendo una correlación negativa entre edad y tasa de mortandad.

El objetivo de este trabajo no es sólo descubrir modelos excepcionales, sino también relaciones entre patrones que permitan describir comportamientos que forman modelos excepcionales. La extracción de este tipo de relaciones excepcionales se formula como una nueva tarea que está definida a medio camino entre la minería de reglas de asociación y la minería de modelos excepcionales. Cabe destacar que esta nueva tarea tiene un coste computacional mucho mayor que la tarea de extracción de modelos excepcionales, pues no sólo tiene que descubrir los modelos, sino que también tiene que describir su comportamiento mediante reglas de asociación. Así, considerando el coste computacional como una desventaja importante de esta tarea de extracción de conocimiento, se considera el uso de un algoritmo evolutivo basado en gramáticas como un paso importante para incluir restricciones sintácticas que permiten reducir el espacio de búsqueda. El uso de un algoritmo de programación genética gramatical permite introducir una serie de restricciones sintácticas que guían el proceso hacia regiones del espacio de búsqueda más próximas a la solución buscada. Además, el uso de gramáticas se considera como una forma de introducir conocimiento subjetivo y externo en el proceso de extracción, por lo que es de gran utilidad para usuarios en determinados ámbitos.

La principal contribución de este trabajo es la propuesta de una nueva tarea, conocida como extracción de reglas de asociación excepcionales, que permite ampliar el modo en el que los datos son tratados, descritos y analizados. Esta tarea permite describir relaciones excepcionales en grupos que, a simple vista, parecen homogéneos. Para demostrar su utilidad, se han realizado una serie de experimentos sobre diferentes métricas en el campo de la minería de patrones. El estudio experimental ha permitido descubrir relaciones interesantes entre pares de métricas, describiendo comportamientos excepcionales. Los resultados son de gran utilidad para expertos en el campo de la minería de patrones, pues pares de métricas que, aparentemente están correladas, se comportan de manera excepcional bajo ciertas circunstancias.

El resto del artículo se organiza como se describe: Sección 2 presenta algunas definiciones importantes y trabajos relacionados; Sección 3 describe la propuesta

presentada en este trabajo; Sección 4 presenta un estudio experimental; por último, la Sección 5 muestra algunas conclusiones alcanzadas con la realización de este trabajo.

2. Preliminares y definición del problema

La minería de reglas de asociación [5] es considerada como una de las tareas descriptivas más importantes en minería de datos. El objetivo de esta tarea es la extracción de relaciones frecuentes y de interés entre patrones para describir comportamientos que, por lo general, son desconocidos en los datos. Este tipo de relaciones entre patrones son conocidas como reglas de asociación, definiéndose como implicaciones de la forma $Antecedente \rightarrow Consecuente$, donde tanto el antecedente como el consecuente son conjuntos disjuntos.

En términos generales, una regla de asociación describe que si el antecedente es satisfecho en un conjunto de datos, entonces es altamente probable que el consecuente de la regla también se satisfaga en dicho conjunto de datos. Por lo general, se considera que la minería de reglas de asociación se diseñó para el análisis de la cesta de la compra, permitiendo el incremento de ventas al tratar de manera diferente determinados productos que están relacionados.

Centrándonos en el problema de minería de modelos excepcionales, *Leman et al.* [3] definió de manera formal el problema como una extensión del problema de *subgroup discovery* [6], donde el objetivo es descubrir subgrupos cuyo modelo es sustancialmente diferente al mismo modelo para los datos no satisfechos por el subgrupo [4]. Esta excepcionalidad del modelo es obtenida al comparar su comportamiento sobre el subgrupo y sobre el complemento.

De manera formal, permitanos considerar una base de datos $\mathcal{D}$ que contiene un conjunto de transacciones $t \in \mathcal{D}$ y donde cada transacción contiene un determinado número de características $f_1, ..., f_k$. Un patrón P en la base de datos $\mathcal{D}$ se define como una función $P \rightarrow \{0, 1\}$. Dicho patrón P tomará el valor $P(t_i) = 1$ para la i-ésima transacción si y sólo si P satisface t_i, mientras que tomará el valor $P(t_i) = 0$ en caso contrario.

De igual forma, un subgrupo G_P para un patrón P se define de manera formal como un conjunto de transacciones $G_P \subseteq \mathcal{D}$ que son satisfechas por P, es decir, $G_P = \{\forall t_i \in \mathcal{D} \ : \ P(t_i) = 1\}$. Por el contrario, el complemento de un subgrupo $\overline{G}_P$, se define como el conjunto de transacciones $\overline{G}_P \subseteq \mathcal{D}$ que no son satisfechas por P, es decir, $\overline{G}_P = \mathcal{D} \setminus G_P$.

Analizando la tarea de extracción de modelos excepcionales, se puede equiparar con la paradoja *Simpson* [7], la cual es una contradicción que puede aparecer en estadística. Esta paradoja ocurre cuando una tendencia específica aparece en diferentes conjuntos de datos pero desaparece cuando todos los datos son combinados: $\frac{A}{B} > \frac{a}{b}, \frac{C}{D} > \frac{c}{d} \not\Rightarrow \frac{A+C}{B+D} > \frac{a+c}{b+d}$

3. Algoritmo EARM

La extracción de reglas de asociación excepcionales puede ser llevada a cabo mediante técnicas de búsqueda exhaustiva. Sin embargo, este tipo de técnicas son computacionalmente costosas cuando la cantidad de datos a analizar es extremadamente grande. A esto hay que añadir la enorme dificultad de este tipo de técnicas cuando los datos están definidos en dominios continuos.

Todos estos inconvenientes, junto con la necesidad añadida de guiar las soluciones hacia zonas específicas del espacio de búsqueda acorde a las necesidades del usuario, hacen que propongamos un modelo de programación genética gramatical para extraer reglas excepcionales. El modelo propuesto, conocido como EARM (*Exceptional Association Rule Mining*), tiene como principal característica el uso de una gramática que permite guiar la búsqueda, definir el tipo de soluciones permitidas, así como incorporar conocimiento externo y subjetivo por parte del usuario final.

Codificación. El algoritmo propuesto está basado en programación genética gramatical [8]. Esta metodología propone el uso de gramáticas para guiar el proceso de búsqueda, de manera que cada solución debe satisfacer las restricciones sintácticas impuestas por la gramática G. La Figura 1 muestra la gramática G de contexto libre usada para el modelo propuesto en este trabajo.

De manera más formal, una gramática de contexto libre puede definirse como una tupla $(\Sigma_N, \Sigma_T, P, S)$ donde Σ_T representa el alfabeto de símbolos terminales, y Σ_N el alfabeto de símbolos no terminales, teniendo en cuenta que ambos alfabetos no poseen ningún elemento en común, es decir, $\Sigma_N \cap \Sigma_T = \emptyset$. Con el fin de producir soluciones válidas en base a la gramática G definida, estas soluciones son producidas mediante la aplicación de un conjunto de reglas de producción del conjunto P, comenzando por el símbolo inicial S.

$G = (\Sigma_N, \Sigma_T, P, S)$ con:

S = Regla
Σ_N = {Regla, Condiciones, Consecuente, Condicion, Condicion_Discreta, Condicion_Continua }
Σ_T = {'Y', 'Atributo', 'Valor', '=', 'ENTRE','Min_valor','Max_valor', 'Atributo_consecuente', 'Valor_consecuente' }
P = {Regla = Condiciones, Consecuente ;
Condiciones = 'Y', Condiciones, Condicion | Condicion ;
Condicion = Condicion_Discreta | Condicion_Continua ;
Condicion_Discreta = 'Atributo', '=', 'Valor' ;
Condicion_Continua = 'Atributo', 'ENTRE', 'Min_valor', 'Max_valor' ;
Consecuente = Condicion_Discreta | Condicion_Continua ; }

Figura 1: Gramática de contexto libre definida para representar reglas de asociación conteniendo tanto atributos discretos como continuos.

Considerando la gramática definida por el algoritmo EARM (ver Figura 1), se obtiene que cualquier solución debe cumplir el siguiente lenguaje $L(G) = \{(Y\ Condicion)^n\ Condicion \rightarrow Consecuente : n \geq 0\}$. Así pues, el lenguaje definido permite obtener reglas con una o más condiciones y un único consecuente. Tanto el consecuente de la regla, como cada una de las condiciones pueden ser de tipo discreta o continua, por lo que las reglas se ajustan a cualquier conjunto de datos, no requiriendo ninguna transformación previa de los datos continuos en un espacio discreto.

Cada uno de los individuos generados por el algoritmo EARM consta de dos partes diferenciadoras. La primera es la regla de asociación obtenida a partir de la gramática G de contexto libre. La segunda permite definir el contexto en el que la regla de asociación tiene un comportamiento excepcional. Cabe destacar que este contexto comprende atributos del conjunto de datos sin que pueda existir repetición entre los atributos del contexto y de la regla de asociación definida por la gramática.

Con el fin de representar el contexto en el que la regla de asociación es excepcional, se utiliza un *bit-set* donde el i-ésimo *bit* representa el i-ésimo atributo. La longitud de este *bit-set* es igual al número de atributos de la base de datos. Considerando de nuevo el problema descrito en la Sección 1 acerca del ratio de mortandad según la edad, se consideran cuatro atributos (*edad*, *tasa_mortandad*, *pobreza*, $enfermedad\_contagiosa$). Partiendo de estos cuatro atributos, es posible definir la regla de asociación $pobreza \rightarrow enfermedad\ contagiosa$ que tiene un comportamiento excepcional cuando se define en el contexto *edad* y *tasa_mortandad*. Así pues, el *bit-set* de este individuo estaría representado como 1100, indicando que los atributos activos en el contexto son *edad* y *tasa_mortandad*.

Esquema evolutivo. La propuesta evolutiva presentada en este artículo (ver Figura 2) sigue un modelo evolutivo clásico, donde las primeras soluciones son generadas, de manera aleatoria, mediante la aplicación de reglas de producción acorde con la gramática de contexto libre definida en la Figura 1. Una vez que estas primeras soluciones se han obtenido, el proceso evolutivo se lleva a cabo a lo largo de un conjunto de generaciones (ver Figura 2, líneas 7 a 22).

El algoritmo EARM posee dos operadores genéticos que permiten generar nuevos individuos en cada generación. El primero de estos operadores genéticos se comporta como un operador clásico, permitiendo seleccionar aleatoriamente una condición de la regla y reemplazándola por otra nueva (ver Figura 2, líneas 9 a 12). Por el contrario, el segundo operador genético es específico para la nueva tarea propuesta en este trabajo, ya que agrupa individuos en base a su contexto, y muta el peor individuo de cada contexto (ver Figura 2, líneas 14 a 19). Esta mutación no se produce sobre la regla de asociación definida por el individuo, sino sobre el *bit-set* en el que se define. Este operador se basa en el hecho de que una regla de asociación no tiene por qué ser excepcional en cualquier contexto, siendo necesario la búsqueda del contexto adecuado. El intercambio de contextos llevado a cabo por este operador genético se representa gráficamente en la Figura 3.

Entrada: $tamPoblacion, maxGeneraciones$ {Número de individuos y generaciones}
Salida: *poblacion*
1: inicialización de todos los conjuntos al valor $\emptyset$
2: $numero\_Generaciones \leftarrow 0$
3: $poblacion \leftarrow$ generar($tamPoblacion$)
4: **for** $\forall elemento \in poblacion$ **do**
5: evaluar($elemento$)
6: **end for**
7: **while** $numero\_Generacioness < maxGeneraciones$ **do**
8: $padres \leftarrow$ seleccionar individuos de $poblacion$ mediante un selector por torneo
9: **for** $\forall elemento \in padres$ **do**
10: reemplazar una condicion aleatoria de $elemento$ por una nueva condicion
11: $hijos \leftarrow hijos \cup elemento$
12: **end for**
13: $poblacionAuxiliar \leftarrow hijos \cup poblacion$
14: **for** $\forall elemento \in poblacionAuxiliar$ **do**
15: **if** $elemento$ es la peor solucion en su contexto **then**
16: seleccionar un nuevo contexto para $elemento$
17: evaluar($elemento$)
18: **end if**
19: **end for**
20: $poblacion \leftarrow$ seleccionar las mejores soluciones de $poblacionAuxiliar$
21: $numero\_Generaciones++$
22: **end while**
23: return *poblacion*

Figura 2: Pseudocódigo del algoritmo evolutivo propuesto.

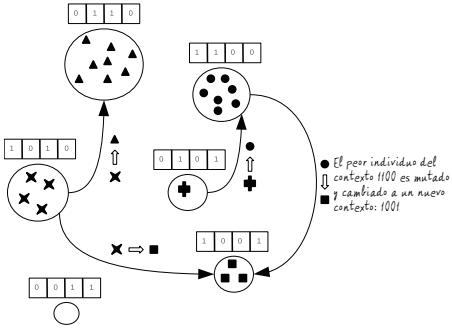

Figura 3: Representación del comportamiento del operador genético para mover de contexto la peor solución de grupo.

Evaluación de las soluciones. Cada una de las soluciones generadas por el algoritmo EARM son evaluadas para determinar su importancia dentro del espacio de soluciones. El proceso de evaluación del algoritmo EARM juega un papel determinante a la hora de esclarecer si una regla de asociación se comporta de manera excepcional en un contexto determinado. Con este fin, el algoritmo incluye un proceso de evaluación basado en el coeficiente ρ de correlación de *Pearson* ($\rho_{X,Y} = COV(X,Y)/(\sigma_x \sigma_Y)$), determinando cómo de excepcional es cada regla respecto al contexto en el que se definen. Este coeficiente ρ toma valores en el rango $[-1, 1]$, donde un valor 1 representa una correlación positiva, un valor -1 determina una correlación negativa, mientras que un valor 0 indica que no existe ninguna correlación.

Volviendo de nuevo al ejemplo en el que se desea saber el ratio de mortandad según la edad. Como muestra la Figura 4, la *edad* y la *tasa_mortandad* están positivamente correladas. Sin embargo, la regla de asociación definida por G_P y representada como $pobreza \rightarrow enfermedad\ contagiosa$ determina un comportamiento excepcional sobre el contexto *edad* y *tasa_mortandad*. De esta manera, el coeficiente ρ para el contexto $\{edad, tasa\_mortandad\}$ es cercano a 1, mientras que dicho coeficiente practicamente cero cuando se considera G_P.

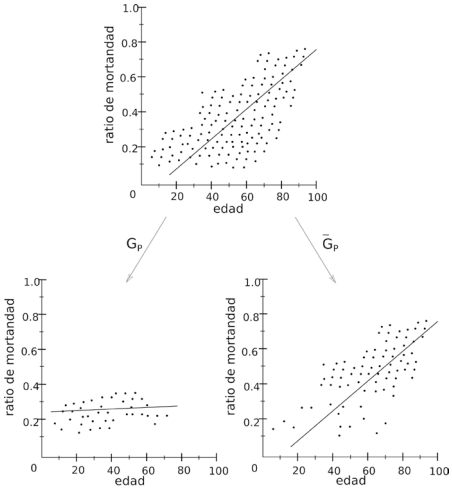

Figura 4: Ejemplo de comportamiento excepcional donde G_P es definida por la regla de asociación $pobreza \rightarrow enfermedad\ contagiosa$.

La función de *fitness* F (ver Ecuación 1) propuesta en este algoritmo está basada en la diferencia existente entre el coeficiente de correlación de las variables contextuales sobre las instancias satisfechas por la regla ($\rho_{Targets}(R)$), y el coeficiente de correlación de las mismas variables contextuales sobre las instancias no satisfechas por la regla ($\rho_{Targets}(\overline{R})$).

$$F(R) = |\rho_{Targets}(R) - \rho_{Targets}(\overline{R})| \quad (1)$$

Esta función de *fitness* es aplicada sólo en aquellos casos en los que la regla de asociación R cumple unas determinados valores mínimos de calidad para el soporte de la regla (frecuencia con la que la regla se cumple en los datos), y la confianza o exactitud de la regla. Si estos umbrales no son satisfechos, entonces la función de *fitness* F toma el valor 0. En caso contrario, tomará el valor obtenido de la ecuación mostrada anteriormente (ver Ecuación 1).

4. Estudio experimental

El objetivo de este estudio experimental es demostrar el interés de la tarea de extracción de reglas de asociación excepcionales sobre un caso de aplicación real. En este estudio, se analizan una serie de métricas utilizadas en minería de patrones a fin de descubrir comportamientos anómalos entre estas medidas. En este estudio experimental, hemos utilizado un *dataset* que incluye todos y cada uno de los posibles valores que puede tener cada una de las medidas utilizadas (Soporte, Soporte del antecedente, Soporte del consecuente, Confianza, *Lift*, *Conviction*, *Leverage*, *Certainty Factor* y *Cosine*).

Al ejecutar el algoritmo propuesto sobre el conjunto de datos descrito, se obtienen una serie de reglas de asociación excepcionales. Una de las reglas excepcionales más interesantes descubiertas es la mostrada en la Figura 5, la cual es definida sobre el contexto *conviction* y soporte del consecuente. Esta regla ha obtenido una valor de *fitness* igual a 0.9631, determinando que si la confianza de una regla está en el rango $[0.2, 0.4]$, entonces el *cosine* tomará un valor en el rango $[0.1, 0.8]$. Esta regla se satisface en un 12.75 % de las transacciones y tiene un exactitud del 97.01 %. Sin embargo, lo más interesante de esta regla es que, definida sobre el contexto de *conviction* y del soporte del consecuente, obtiene un comportamiento completamente diferente que si esta regla no se cumpliese (ver Figura 5). Esta regla de asociación excepcional describe que, a pesar de que el valor de *conviction* y el soporte del consecuente parecen no estar correlados, la correlación entre ellos es negativa si la confianza toma un valor en el rango $[0.2, 0.4]$, y el *cosine* toma un valor en el rango $[0.1, 0.8]$

Analizando las variables contextuales soporte y *lift*, las cuales son dos de las medidas de calidad más ampliamente utilizadas en el campo de la minería de patrones, obtenemos que ambas medidas no están muy correladas. Sin embargo, cuando las medidas de soporte del antecedente y *cosine* entran en juego (ver Figura 6), esta correlación es negativa. Se obtiene así una regla de asociación excepcional, definida como **SI** *soporte antecedente* ENTRE $[0.4, 0.7]$ **ENTONCES** *cosine* ENTRE $[0.2, 0.3]$.

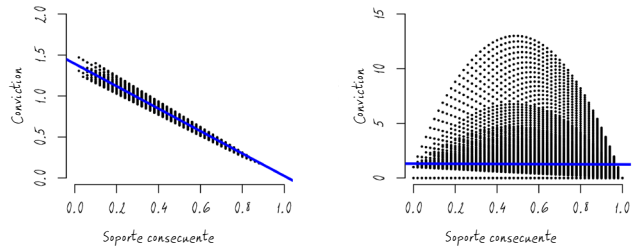

Figura 5: Nube de puntos y recta de regresión de la regla **SI** *confianza* ENTRE $[0.2, 0.4]$ **ENTONCES** *cosine* ENTRE $[0.1, 0.8]$ definido sobre el contexto dado por las variables conviction y soporte del consecuente (gráfico de la izquierda); y su complemento (gráfico de la derecha).

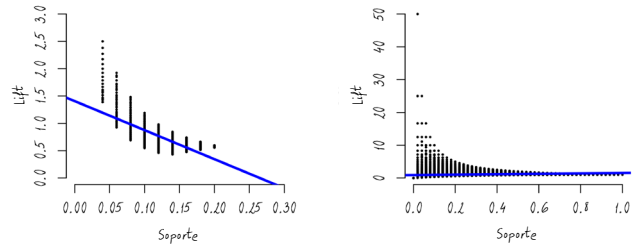

Figura 6: Nube de puntos y recta de regresión de la regla **SI** *soporte antecedente* ENTRE $[0.4, 0.7]$ **ENTONCES** *cosine* ENTRE $[0.2, 0.3]$ definido sobre el contexto dado por las variables soporte y *lift* (gráfico de la izquierda); y su complemento (gráfico de la derecha).

5. Conclusiones

En este trabajo, se ha presentado una nueva forma de describir la información oculta en conjuntos de datos, proponiendo la tarea de minería de reglas de asociación excepcionales. Esta tarea puede considerarse como un híbrido entre minería de asociaciones y minería de modelos excepcionales, el cual tiene como objetivo descubrir subgrupos de datos donde un modelo definido en dichos datos tiene un comportamiento completamente excepcional respecto al mismo modelo definido sobre el complemento del subgrupo.

Con el fin de mostrar la utilidad de esta tarea, se ha presentado un modelo evolutivo basado en programación genética gramatical. El modelo propuesto, llamado EARM, permite guiar el proces de búsqueda hacia soluciones deseadas, restringir el espacio de búsqueda, así como introducir conocimiento externo en el proceso de extracción de conocimiento.

EARM ha sido probado sobre un conjunto de datos para permitir el descubrimiento de comportamientos anómalos entre medidas que, a priori, se comportan de una manera homogénea.

Agradecimientos

El presente trabajo ha sido financiado por el Ministerio de Innovación y Ciencia, y los fondos FEDER, bajo el proyecto TIN-2014-55252-P.

Referencias

1. J. M. Luna, "Pattern mining: current status and emerging topics," *Progress in Artificial Intelligence*, pp. 1–6, 2016. [Online]. Available: http://dx.doi.org/10.1007/s13748-016-0090-4
2. J. M. Luna, A. Cano, V. Sakalauskas, and S. Ventura, "Discovering useful patterns from multiple instance data," *Information Sciences*, vol. 357, pp. 23–38, 2016.
3. D. Leman, A. Feelders, and A. J. Knobbe, "Exceptional model mining," in *Proceedings of the European Conference in Machine Learning and Knowledge Discovery in Databases*, ser. ECML/PKDD 2008, vol. 5212. Antwerp, Belgium: Springer, 2008, pp. 1–16.
4. W. Duivesteijn, A. J. Feelders, and A. Knobbe, "Exceptional model mining," *Data Mining and Knowledge Discovery*, vol. 30, no. 1, pp. 47–98, 2015.
5. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.
6. F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on subgroup discovery: Foundations and applications," *Knowledge and Information Systems*, vol. 29, no. 3, pp. 495–525, 2011.
7. Y. Z. Ma, "Simpson's paradox in GDP and per capita GDP growths," *Empirical Economics*, vol. 49, no. 4, pp. 1301–1315, 2015.
8. R. McKay, N. Hoai, P. Whigham, Y. Shan, and M. O'Neill, "Grammar-based Genetic Programming: a Survey," *Genetic Programming and Evolvable Machines*, vol. 11, pp. 365–396, 2010.