Reglas de Asociación Difusas para la Detección de Anomalías

M.D. Ruiz, D. Sánchez, M.J. Martin-Bautista,
M.A. Vila, M. Delgado

6 de Febrero de 2014

Motivación

- Las Reglas de Asociación permiten identificar conocimiento potencialmente nuevo, útil y comprensible para el usuario.
- Representan la aparición conjunta de un conjunto de ítems en la mayoría de las transacciones de una base de datos.
- Anteriores propuestas usan técnicas de minería de datos para:
  - Descubrir los perfiles usuales del comportamiento de los clientes
  - Después, buscan anomalías asociadas a estos comportamientos con técnicas como el clustering.

Motivación

- Las **Reglas de Asociación Anómalas** permiten la extracción de conocimiento mediante reglas de asociación, obteniendo:
  - el patrón usual/normal
  - el patrón anómalo asociado al usual/normal.
- Mucha de la información disponible es imprecisa, ambigua o incierta $\rightsquigarrow$ Uso de la Teoría de Subconjuntos Difusos
- Propuesta: Reglas Difusas Anómalas
- Ventajas:
  1. Conocimiento comprensible y semánticamente significante para el usuario.
  2. Comportamiento usual y el anómalo asociado a éste.
  3. Menos cantidad de reglas obtenidas.

Campos de Aplicación

- Seguridad, detección de fraude
- Análisis de datos financieros
- Análisis de flujos en redes de datos
- Procesos biológicos y químicos...

Resumen

Trabajos relacionados

Existen otros tipos de reglas infrecuentes que capturan conocimiento que puede ser de interés para el usuario:

- Reglas Peculiares
- Reglas Infrecuentes
- Reglas de Excepción
- Reglas Anómalas

Reglas de Asociación Crisp

- Una base de datos D estará compuesta por transacciones t_i (filas) y atributos (columnas).
- Llamaremos *ítem* a un par del tipo $\langle atributo, valor \rangle$ or $\langle atributo, intervalo \rangle$.

D	i_1	i_2	$\ldots$	i_j	i_{j+1}	$\ldots$	i_m
t_1	1	0	$\ldots$	0	1	$\ldots$	0
t_2	0	1	$\ldots$	1	1	$\ldots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_n	1	1	$\ldots$	0	1	$\ldots$	1

- Una Regla de Asociación es una expresión de la forma $A \rightarrow B$ donde A, B son conjuntos no vacíos de ítems con intersección vacía.
- Una Regla de asociación representa la aparición conjunta de A y B.

Reglas de Asociación Crisp

- El soporte de un itemset A se define como la probabilidad de que una transacción contenga a A

$$supp(A) = \frac{|t \in D \ : \ A \subseteq t|}{|D|}$$

- Para validar la regla, se suele utilizar el $soporte$ (probabilidad conjunta $P(A \cup B)$) y la $confianza$ (probabilidad condicionada $P(B|A)$)

$$\text{Sop}(A \to B) = \frac{\text{sop}(A \cup B)}{|D|}; \quad \text{Conf}(A \to B) = \frac{\text{sop}(A \cup B)}{\text{sop}\,(A)}$$

que deben ser $\geq minsop$ y $\geq minconf$ resp. (umbrales impuestos por el usuario), i.e, la regla es frecuente y confidente.

Reglas de Asociación Crisp

- Debido a algunos problemas al usar la confianza, usaremos como alternativa el uso del *factor de certeza*, $FC(A \to B)$

$$\begin{cases} \dfrac{\text{Conf}(A \to B) - \text{sop}(B)}{1 - \text{sop}(B)} & \text{if } \text{Conf}(A \to B) > \text{sop}(B) \\ \dfrac{\text{Conf}(A \to B) - \text{sop}(B)}{\text{sop}(B)} & \text{if } \text{Conf}(A \to B) < \text{sop}(B) \\ 0 & \text{en otro caso.} \end{cases}$$

- El FC mide cómo nuestra creencia de que B está en una transacción cambia cuando sabemos que A está en dicha transacción.
- El Factor de Certeza posee mejores propiedades que la confianza y que otras medidas, en particular, reduce el número de reglas obtenidas filtrando aquellas que corresponden a dependencias negativas o independencia estadística.
- Si una regla es $\geq minFC$ diremos que es fiable.

Reglas de Asociación Difusas

- I un conjunto finito de ítems.
- Una transacción difusa es un subconjunto difuso no vacío $\tilde{\tau} \subseteq I$.
- Un ítem $i \in I$ pertenecerá a $\tilde{\tau}$ con grado $\tilde{\tau}(i) \in [0, 1]$.
- Un itemset $A \subset I$ pertenecerá a $\tilde{\tau}$ con grado

$$\tilde{\tau}(A) = \min_{i \in A} \tilde{\tau}(i)$$

- Una regla de asociación difusa $A \rightarrow B$ se cumplirá en $\tilde{D} \Leftrightarrow$

$$\tilde{\tau}(A) \leq \tilde{\tau}(B) \quad \forall\, \tilde{\tau} \in \tilde{D}$$

- Esta definición preserva el significado original de regla de asociación crisp.

Reglas de Asociación Difusas

Las medidas de soporte, confianza y factor de certeza se generalizan al caso difuso mediante la evaluación de sentencias cuantificadas con el cuantificador $Q(x) = x$ de la forma:

- Dado A conjunto de ítems, $\tilde{\Gamma}_A(\tilde{\tau}) = \tilde{\tau}(A)$.
- Soporte de A es la evaluación de la sentencia cuantificada "Q de los $\tilde{D}$ son $\tilde{\Gamma}_A$".
- $\mathrm{Fsop}(A \to B)$ en $\tilde{D}$ es la evaluación de la sentencia cuantificada "Q de los $\tilde{D}$ son $(\tilde{\Gamma}_A \cap \tilde{\Gamma}_B)$".
- $\mathrm{FConf}(A \to B)$, es la evaluación de la sentencia cuantificada "Q de los $\tilde{\Gamma}_A$ son $\tilde{\Gamma}_B$" .
- $\mathrm{FFC}(A \to B)$ se obtiene usando Fsop y FConf usando la definición anterior.

Reglas de Asociación Difusas

Para evaluar la sentencia cuantificada "Q de los F son G" usaremos el método GD:

$$GD_Q(G/F) = \sum_{\alpha_i \in \Lambda(G/F)} (\alpha_i - \alpha_{i+1})\, Q\left(\frac{|(G \cap F)_{\alpha_i}|}{|F_{\alpha_i}|}\right)$$

donde $\Lambda(G/F) = \Lambda(G \cap F) \cup \Lambda(F)$, siendo $\Lambda(F)$ el conjunto de niveles de F, y $\Lambda(G/F) = \{\alpha_1, \ldots, \alpha_p\}$ con $\alpha_i > \alpha_{i+1}$ para cualquier $i \in \{1, \ldots, p-1\}$ donde $\alpha_{p+1} = 0$. El conjunto F deberá estar normalizado. Si no, F se normalizará y el mismo factor de normalización se le aplicará a $G \cap F$.

Reglas de Asociación Difusas

Ejemplo:

- Conjunto de ítems $I = \{i_1, i_2, i_3, i_4\}$
- Conjunto de transacciones difusas

	i_1	i_2	i_3	i_4	i_5
$\tilde{\tau}_1$	1	0.2	1	0.9	0.9
$\tilde{\tau}_2$	1	1	0.8	0	0
$\tilde{\tau}_3$	0.5	0.1	0.7	0.6	0
$\tilde{\tau}_4$	0.6	0	0	0.5	0.5
$\tilde{\tau}_5$	0.4	0.1	0.6	0	0
$\tilde{\tau}_6$	0	1	0	0	0

- $\tilde{\tau}_6$ es una transacción crisp.
- Algunos grados de pertenencia son: $\tilde{\tau}_1(\{i_3, i_4\}) = 0.9$, $\tilde{\tau}_1(\{i_2, i_3, i_4\}) = 0.2$ and $\tilde{\tau}_2(\{i_1, i_2\}) = 1$.

Reglas de Asociación Difusas

Algunas reglas difusas que pueden encontrarse:

Regla	Fsop	FConf	FFC
$\{i_1, i_2\} \rightarrow \{i_3\}$	0.167	0.8	0.6
$\{i_4\} \rightarrow \{i_5\}$	0.2	0.767	0.68

Reglas de Asociación Anómalas

Idea: Las Reglas Anómalas surgen cuando se elimina el efecto dominante producido por la regla frecuente (csr) [Berzal et al., 2004]

- Interpretación:

  Cuando se da X, entonces tendremos Y (usual) o A (inusual)

- Esta semántica se captura con el conjunto de reglas:

  X implica de forma fuerte a Y,
  pero en los casos donde X implica $\neg Y$,
  entonces X implica confidentemente A

- Ejemplo:

  "SI *un paciente tiene síntomas X*,
  ENTONCES *normalmente tendrá la enfermedad Y*,
  SI NO, *tendrá la enfermedad A*",

- A será el comportamiento alternativo cuando el "usual" o "normal" falla (representado por $X \rightarrow Y$).

Reglas Anómalas Difusas

- Formalmente, let $\tilde{D}_X = \{t \in \tilde{D} : X \subset t\}$.
- Una *regla anómala difusa* será una terna $(Fcsr, Fanom, Fref)$ cumpliendo:
  - $X \to Y$ (Fcsr) es frecuente y fiable en $\tilde{D}$.
  - $\neg Y \to A$ (Fanom) es fiable en $\tilde{D}_X$.
  - $A \to \neg Y$ (Fref) es fiable in $\tilde{D}_X$.
- Ventajas:
  - La cantidad de ternas $(csr, anom, ref)$ se reduce al usar el Factor de Certeza.
  - El conjunto de reglas obtenidas son más fiables.

Experimentos y Resultados

Algunos problemas de implementación:

- La extracción de reglas difusas es algorítmicamente más costoso que extraer reglas crisp.
- Las reglas anómalas son infrecuentes $\rightsquigarrow$ Las técnicas de poda no pueden usarse.

Solución:

- Extraer reglas difusas mediante una paralelización por niveles.
- Adaptarlo para la extracción de reglas anómalas difusas.

Experimentos y Resultados

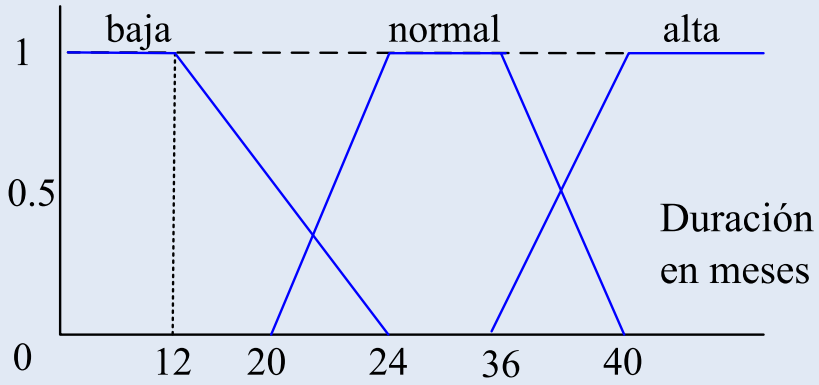
Bases de datos: `Auto-mpg` y `German-statLog` del repositorio UCI Machine Learning.

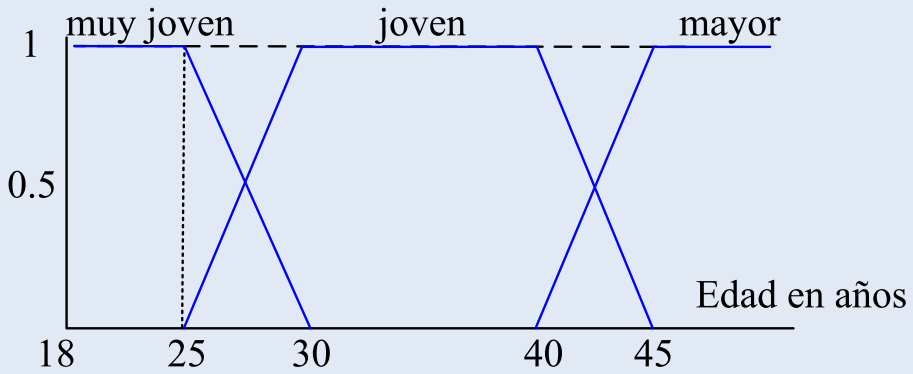
- `Auto-mpg` (cilindrada, peso, aceleración,...)
- 398 tuplas
- 8 atributos que han sido fuzzificados con las etiquetas lingüísticas: bajo, medio y bajo.

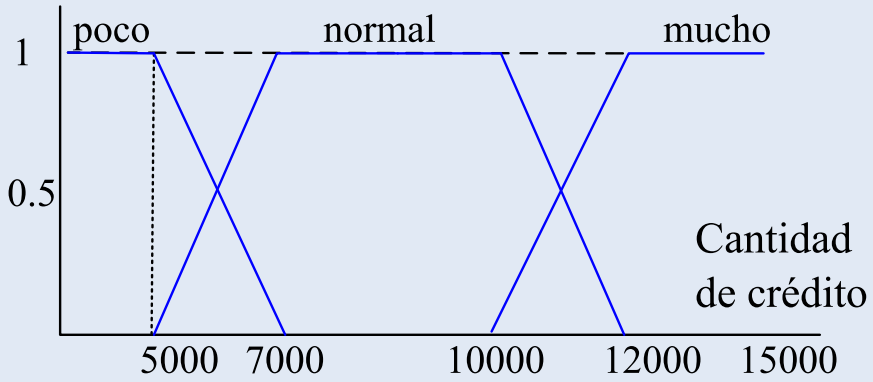
Experimentos y Resultados

`German-statLog` (sobre créditos y clientes de un banco alemán)

- 1000 transacciones
- 21 atributos: 18 categóricos o numéricos, 3 continuos (fuzzificados)







Experimentos y Resultados

- Ejemplo de regla obtenida en `Auto-mpg`

"SI `consumo=alto`
$ENTONCES$ `cilindrada = 4` ($\mathrm{Fsop} = 0.206$ & $\mathrm{FFC} = 0.912$)
O `n. de cilindros = en la media` (inusual $\mathrm{FFC}_1 = 1$, $\mathrm{FFC}_2 = 1$)

Interpretación:

SI el consumo de gasolina es alto ENTONCES la cilindrada es 4 con soporte 0.206 y factor de certeza 0.912, *O BIEN inusualmente* el número de cilindros está en la media con factor de certeza 1.

Conclusiones y Trabajos Futuros

- Hemos propuesto el uso de reglas anómalas difusas pare representar el comportamiento anómalo o inusual que se desvía del normal.
- Usamos el Factor de Certeza, obteniendo un conjunto más pequeño y fiable de reglas.
- Proponemos adaptar un algoritmo ya conocido para reglas difusas.
- Probado en distintas bases de datos donde se han fuzzificado varios atributos continuos.

Futuro: Desarrollar algoritmos más eficientes.

- Extender otras propuestas parecidas, e.g. reglas de excepción, que pueden ser de utilidad para aplicarlas en contextos de detección de fraude y seguridad.

Referencias

[Berzal et al., 2004] F. Berzal, J.C. Cubero, N. Marín, and M. Gámez. **Anomalous association rules.** In *IEEE ICDM Workshop Alternative Techniques for Data Mining and Knowledge Discovery*, 2004.

[Delgado et al., 2011] M. Delgado, M.D. Ruiz, and D. Sánchez. **New Approaches for Discovering Exception and Anomalous Rules.** *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 19, No. 2 pp. 361-399, 2011.

ESTYLF 2014

Muchas gracias. ¿Alguna pregunta?