Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Reglas de Asociación
(Supervisado)

1

Profesora: *Dra. Fabiola Ocampo Botello*

2

Larose & Larose (2015) establecen que El análisis de afinidad se refiere al estudio o características que "van juntas". Los métodos para el análisis de afinidad son conocidos como "análisis de la canasta de mercado", la cual busca descubrir asociaciones entre los atributos, es decir, busca descubrir reglas para cuantificar la relación entre dos o más atributos. Las reglas de asociación son de la forma:

Si el antecedente entonces el consecuente.

Considerando una medida de soporte (*support*) y una medida de confianza (*confidence*).

3

Existen dos términos en las reglas de asociación (Han, Kamber & Pei, 2012):

- Soporte s, donde s es el porcentaje de transacciones en D que contienen A∩B, la unión de los conjuntos de A y B, o ambos A y B.
- Confianza c, en el conjunto de transacciones D, donde c es el porcentaje de transacciones en D que contienen A y también a B.

$$\text{Soporte } (A \Rightarrow B) = P(A \cap B)$$

$$\text{Confianza } (A \Rightarrow B) = P(B \mid A)$$

4

Figura tomada de Tan, Steinbach & Kumar (2014).

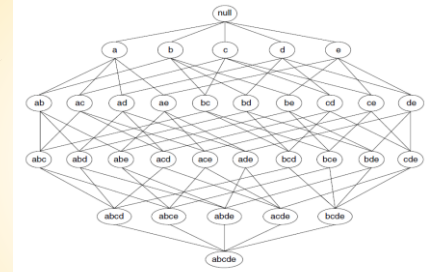


Figure 6.1. An itemset lattice.

5

Las reglas estrictas no son necesariamente interesantes

Si una regla es interesante o no, puede evaluarse subjetiva u objetivamente. En última instancia, solo el usuario puede juzgar si una regla determinada es interesante, y este juicio, al ser subjetivo, puede diferir de un usuario a otro (Han, Kamber & Pei, 2012).

Sin embargo, las medidas objetivas de interés, basadas en las estadísticas "detrás" de los datos, pueden utilizarse como un paso hacia el objetivo de eliminar las reglas poco interesantes que de otro modo se presentarían al usuario. (Han, Kamber & Pei, 2012).

6

García, Romero, Ventura, de Castro & Calders (2011) establecen las siguientes premisas:

- Aunque el apoyo y la confianza permiten eliminar muchas asociaciones, a menudo es deseable aplicar también otras restricciones; por ejemplo, sobre los atributos que deben o no pueden estar presentes en el antecedente o consecuente de las reglas descubiertas.
- Las medidas subjetivas cobran cada vez más importancia, es decir, medidas que se basan en factores subjetivos controlados por el usuario.
- Algunas medidas subjetivas sugeridas son:
  - lo inesperado (las reglas son interesantes si son desconocidas para el usuario o contradicen el conocimiento del usuario) y
  - la capacidad de acción (las reglas son interesantes si los usuarios pueden hacer algo con ellas en su beneficio).

7

García et al (2011) establecen las siguientes premisas:

- Un factor de gran importancia para determinar la calidad de las reglas extraídas es su **comprensibilidad**.
- La experiencia previa y el conocimiento del dominio de esta persona juegan un papel importante en la evaluación de la comprensibilidad. Esto contrasta con la precisión que puede considerarse como una propiedad de las reglas y que puede evaluarse independientemente de los usuarios.
- Existen algunas técnicas tradicionales que se han utilizado para mejorar la comprensibilidad de las reglas descubiertas, como restringir el número de elementos en el antecedente o consecuente de la regla, o realizar una discretización de valores numéricos.

8

Comprensibilidad

Rokach, L. & Maimon, O. (2015) establece que El criterio de **comprensibilidad** (también conocido como **interpretabilidad**) se refiere a qué tan bien los humanos captan el clasificador inducido. Mientras que el error de generalización mide cómo el clasificador se ajusta a los datos, la comprensibilidad mide el "ajuste mental" de ese clasificador.

9

Discretización de los datos

(García et al.,2011)
La discretización divide los datos en clases categóricas que son más fáciles de entender para el analista (los valores categóricos son más familiares para el analista que las magnitudes y rangos precisos) (García et al.,2011).

Hay varios métodos globales no supervisados para transformar atributos continuos en atributos discretos, como el método de igual ancho, el método de igual frecuencia o el método manual (en el que debe especificar los puntos de corte). Las etiquetas que se pueden utilizar son FALLO, PASADO, BUENO y EXCELENTE, y en todos los demás atributos: BAJO, MEDIO y ALTO (García et al.,2011).

10

Discretización de los datos

Para los datos categóricos:

Larose & Larose (2015) establece que algunos métodos de análisis, como los métodos de regresión requieren que los predictores sean valores numéricos, por lo que si el analista desea usar predictores categóricos en la regresión es necesario que recategorice las variables en una o más variables bandera, también conocida como variable dummy o variable indicador debido a que sólo puede tomar uno de dos posibles valores: 1 ó 0.

Ejemplo:

Suponga la variable *región* tiene k = 4 posibles valores {norte, sur, este, oeste}, por lo que al final deberán quedar 4 variables

norte_flag: if región = norte then norte_flag = 1 else norte_flag = 0
sur_flag: if región = sur then sur_flag = 1 else sur_flag = 0
este_flag: if región = este then este_flag = 1 else este_flag = 0
oeste_flag: if región = oeste then oeste_flag = 1 else oeste_flag = 0

11

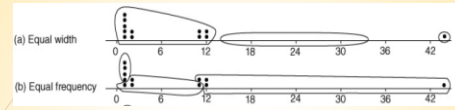
Binning (contenedores) de variables numéricas

Se utiliza para categorizar valores numéricos. Por ejemplo cuando se tiene una serie de datos numéricos y se desea crear categorías de los mismos. Por ejemplo: precio de una casa, sueldo de las personas, edad, etc. (Larose & Larose, 2015)

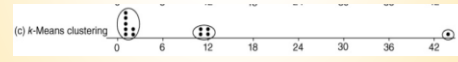
1. El binning de igual amplitud (*Equal width binning*), divide el predictor numérico en k categorías de igual amplitud, donde k es elegido por el cliente o el analista.

2. El binning con la misma frecuencia de ocurrencia de los elementos (*Equal frequency binning*), categorías que contengan la misma cantidad de elementos, en este caso la frecuencia divide el predictor numérico en k categorías, cada una con k/n registros, donde n es el número total de registros.

12



3. El binning por agrupamiento (*Binning by clustering*) utiliza un algoritmo de agrupamiento (cluster), como el k-means para calcular automáticamente la partición "óptima".



13

4. Binning basado en el valor predictivo (*Binning based on predictive value*). Los métodos (1)-(3) ignoran la variable objetivo; el binning basado en el valor predictivo particiona el valor numérico considerando el efecto de cada partición tiene sobre el valor de la variable objetivo.

14

Referencias bibliográficas

García, Romero, Ventura, de Castro & Calders Toon. (2011). Association Rule Mining in Learning Management Systems. In: *Handbook of educational data mining*. Edited by: Romero Cristóbal, Ventura Sebastian, Pechenizkiy Mykola, and Baker Ryan. CRC Press. Taylor & Francis Group. Pp. 93-106.

Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). Data Mining: concepts and techniques. Third edition. Morgan Kaufman Series.

Larose, T. Daniel & Larose, D. Chantal. (2015). *Data Mining and Predictive Analytics*. Second Edition. Wiley.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.

Tan Pang-Ning, Steinbach Michael, Karpatne Anuj, Kumar Vipin. (2005). Introduction to data mining. Second Edition. Pearson