Tema 5. Extracción de reglas de asociación

5.1.Extracción de reglas asociación y dependencia. Cobertura, confianza, interés y correlación.
5.2. Reglas de asociación multinivel
5.3. Reglas secuenciales

5.1.Extracción de reglas asociación y dependencia. Cobertura, confianza, interés y correlación.

5.1.1.Reglas de asociación

Las reglas de asociación son una manera muy popular de expresar patrones de una base de datos. Básicamente son reglas de tipo "si__ entonces_" donde se especifica que si se da el ítem del antecedente se da (se tiende a dar con gran frecuencia) el ítem del consecuente.
Las técnicas de inducción de estas rectas tienen importantes aplicaciones prácticas en el análisis de la cesta de la compra,, estudios de textos, búsquedas de patrones en páginas web etc.
Para ilustrar la situación consideremos el conjunto de datos siguiente en el que se dispone de información de distintas situaciones meteorológicas en las que se jugó o no un partido de tenis.

pronóstico	temperatura	humedad	viento	juego
soleado	cálida	normal	FALSO	no
soleado	cálida	alta	VERDADERO	no
cubierto	cálida	normal	FALSO	si
lluvioso	normal	alta	FALSO	si
lluvioso	normal	normal	FALSO	si
lluvioso	fría	baja	VERDADERO	no
cubierto	fría	baja	VERDADERO	si
soleado	normal	alta	FALSO	no
soleado	fría	baja	FALSO	si
lluvioso	cálida	normal	FALSO	si
soleado	cálida	baja	VERDADERO	si
cubierto	normal	alta	VERDADERO	si
cubierto	cálida	normal	FALSO	si
lluvioso	normal	fría	VERDADERO	no

Un reglas de asociación es una regla del tipo:

SI α entonces β o $\alpha \Rightarrow \beta$

Donde α y β son dos conjuntos de ítems disjuntos donde entendemos por ítem una asignación de un valor a una atributo del tipo: "pronóstico=soleado"

Ejemplos de reglas de asociación podrían ser

1. pronóstico = cubierto==> juego= si
2. temperatura = fría ==> humedad = baja
3. humedad = normal & viento = falso ==> juego= si
4. pronóstico = soleado & juego= no ==> humedad = alta
5. pronóstico = soleado & humedad = alta ==> juego= no

6. pronóstico = lluvioso & juego= si==> viento = falso
7. pronóstico = lluvioso & viento = falso ==> juego= si
8. temperatura = fría & juego= si==> humedad = normal
9. pronóstico = soleado & temperatura = cálida ==> humedad = alta
10. temperatura = cálida & juego= no ==> pronóstico = soleado

Vemos que las hay con uno o varios ítems en el antecedente y que las distintas reglas expresan situaciones que se dan con mayor o frecuencia en la base de datos. Así la regla 1 se da en 4 ocasiones, la regla 2 en 3 ocasiones, etc

Llamamos **cobertura** (support) o soporte al número de instancias que la regla predice correctamente.

Llamamos **confianza** (confidence) o precisión a la proporción (o porcentaje) de veces que la regla se cumple de entre las que se podría cumplir (es decir, de entre las que se da el antecedente de la regla)

Las reglas anteriores tendrían por cobertura y confianza los siguientes valores.

regla	cobertura	confianza
1. pronóstico = cubierto==> juego= si	4	100 %
2. temperatura = fría ==> humedad = baja	3	100%
3. humedad = normal & viento = falso ==> juego= si	4	4/5 = 80 %
4. pronóstico = soleado & juego= no ==> humedad = alta	2	2/3 = 66%
5. pronóstico = soleado & humedad = alta ==> juego= no	2	100%
6. pronóstico = lluvioso & juego= si==> viento = falso	3	100%
7. pronóstico = lluvioso & viento = falso ==> juego= si	3	100 %
8. temperatura = fría & juego= si==> humedad = normal	0	0
9. pronóstico = soleado & temperatura = cálida ==> humedad = alta	1	1/3=33%
10. temperatura = cálida & juego= no ==> pronóstico = soleado	2	2/3= 66%

Un algoritmo apropiado para la extracción de reglas de asociación interesantes debe ser capaz de darnos reglas con una alta confianza y con una alta cobertura. Ya que si las reglas seleccionadas no tiene alta confianza no nos garantizan su éxito y si, aun teniendo alta confianza, no tienen una suficiente cobertura no será razonable creer que en situaciones futuras (extramuestrales) siga siendo una regla “confiable”.

Es cierto, sin embargo que esta doble exigencia puede llevarnos a descartar reglas de alta fiabilidad en el caso en el que estén poco representadas en la base de datos. Pero esto es muy poco probable que ocurra si el el tamaño de la base de datos es lo suficientemente grande.

La tarea de encontrar reglas de gran confianza y cobertura puede parecer ardua ya que el número de conjuntos de ítems crece exponencialmente con el número de variables consideradas. Sin embargo, en los casos reales los conjuntos de ítems de

frecuencia elevada (que formarán parte de reglas de alta cobertura) no suelen ser tantos.

El algoritmo básico para la extracción de reglas de asociación es el llamado **algoritmo a priori.** El algoritmo se plante trabajar con un mínimo de confianza exigido (sólo seleccionará reglas con esa confianza o más) y de entre ellas considerará las de mayor cobertura.

El algoritmo comienza por buscar conjuntos de ítems con una determinada (alta) cobertura.

1.- Busca los conjuntos formados por 1 solo ítem con máxima cobertura. (Probablemente no habrá ninguno con una cobertura igual a N=número de instancias de la base de datos; iremos reduciendo hasta encontrar algun conjunto con un solo ítem que tenga una frecuencia maximal)

En el ejemplo, conjuntos de un ítem con 14 casos no hay ninguno. Buscaríamos uno con 13, 12, 11, etc. El conjunto con un solo con mayor frecuencia es **juego=si** con una cobertura de 9

2.- Usamos los conjuntos de un ítem encontrados para formar conjuntos de dos ítems de cobertura maximal .

En el ejemplo que seguimos , en una primera ejecución del algoritmos buscamos parejas que incluyan juego=si de cobertura 9, no hay; 8, no hay; 7, no hay;… Hay conjuntos de dos de cobertura 4:

{viento= falso, juego=si } , {humedad=normal, juego=si}, {pronostico=cubierto, juego=si}

Con cada una de estas parejas podemos construir dos reglas (a⇒b y b⇒a)

Comprobaríamos la confianza de estas 6 reglas y conservaríamos aquellas que tienen una confianza mayor que la requerida.

regla	Conf.	regla	Conf.
viento= falso ⇒ juego=si	4/5	juego=si ⇒ viento= falso	4/9
humedad=normal ⇒ juego=si	4/5	juego=si ⇒ humedad=normal	4/9
pronostico=cubierto ⇒ juego=si	4/4	juego=si ⇒ pronostico=cubierto	4/9

Si la confianza requerida fuera el 90% sólo seleccionaríamos la reglas pronostico=cubierto ⇒ juego=si , si fuera hasta el 80% consideraríamos también

humedad=normal ⇒ juego=si y viento= falso ⇒ juego=si

3.-Usaríamos, ahora los conjuntos de 2 items y cobertura 4 encontrados en dos para buscar conjuntos de 3 items y cobertura 4

Encontraríamos el conjunto { humedad = normal , viento = falso , juego= si}

Con el que podríamos formar las siguientes 6 reglas cuyas confianzas serían:

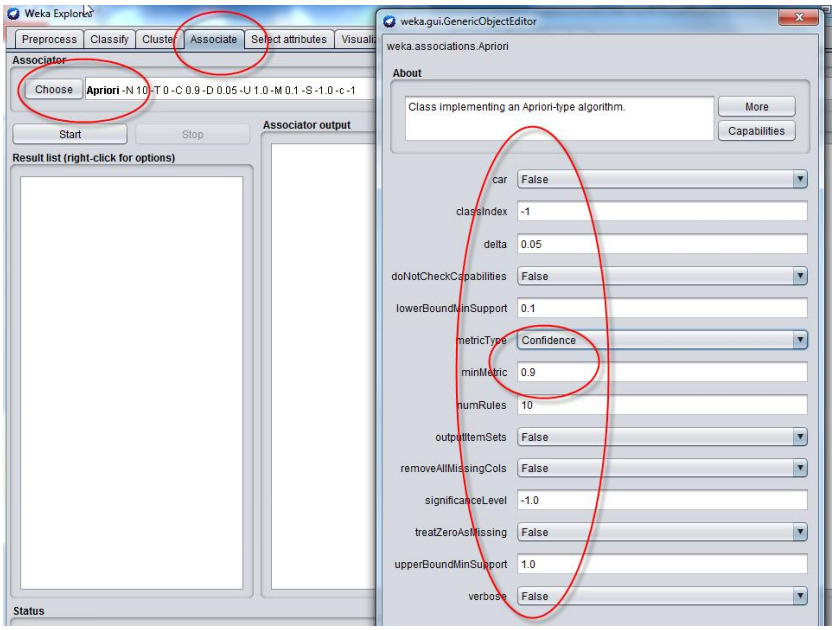
regla	confianza
humedad = normal & viento = falso ==> juego= si	4/5= 80%
humedad = normal & juego= si ==> viento = falso	4/4= 100%
viento = falso & juego= si ==> humedad = normal	4/4= 100%
humedad = normal ==> juego= si & viento = falso	4/5= 80%
viento = falso ==> juego= si & humedad = normal	4/8=50%
juego= si ==> viento = falso & humedad = normal	4/9= 44%

Selecionaríamos las reglas de confianza mayor a la exigida y

En un paso 4 buscaríamos conjuntos de 4 ítems de cobertura 4 y no encontraríamos ninguno por lo que volveríamos al paso 2 con cobertura exigida de 3 (supuesto que 3 aún superara la cobertura exigida)

Repetiríamos el proceso para coberturas inferiores mientras se cumplan las condiciones exigidas.

Para poder lidiar con el problema de la explosión de rescursos de cómputo , especialmente en el caso de grandes bases de datos se han introducido diferentes mejoreas sobre este algoritmos básico que van desde el muestreo de registros, a la estructuración en árboles de registros frecuentes, la acotación de la búsqueda sólo para algunos items o la búsqueda en paralelo.



En weka , el algoritmo a priori puede implementarse con las opciones de:

Límites superior e inferior de cobertura exigida, confianza mínima u otras valores mínimos exigidos para otras métricas (como *lift* = confianza / número de instancias en

las que aparece el consecuente de la reglas ; *leverage* es la proporción de casos adicionales cubiertos tanto por antecendente(s) como por consecuente en relación a los que se darían si antecende y consecuente fueran independientes; y Conviction= (1-cobertura(consecuente)/(1-confinaza(regla)), limitarse a reglas que tengan como consecuente el atributo clase , número de reglas a encontrar, y algunas más.

5.1.2. Reglas de dependencia

Nos podemos encontrar en nuestros datos con una situación en la que la dependencia entre ítems no sea capturada por la selección de reglas basada en cobertura-confianza.
Recordemos que podemos hablar de dependencia entre ítems cuando la probabilidad (o la frecuencia relativa) de que se den conjuntamente (intersección) es diferente al producto de las probabilidades (frecuencia relativa). Si es mayor hablamos de dependencia positiva y si es menor, de dependencia negativa.
Para subsanar este problema se han propuesto métodos de inducción de reglas de dependencia basados en el contraste de independencia de la χ^2, que seleccionar las reglas de dependencia en función de la significación asociada a este contraste.

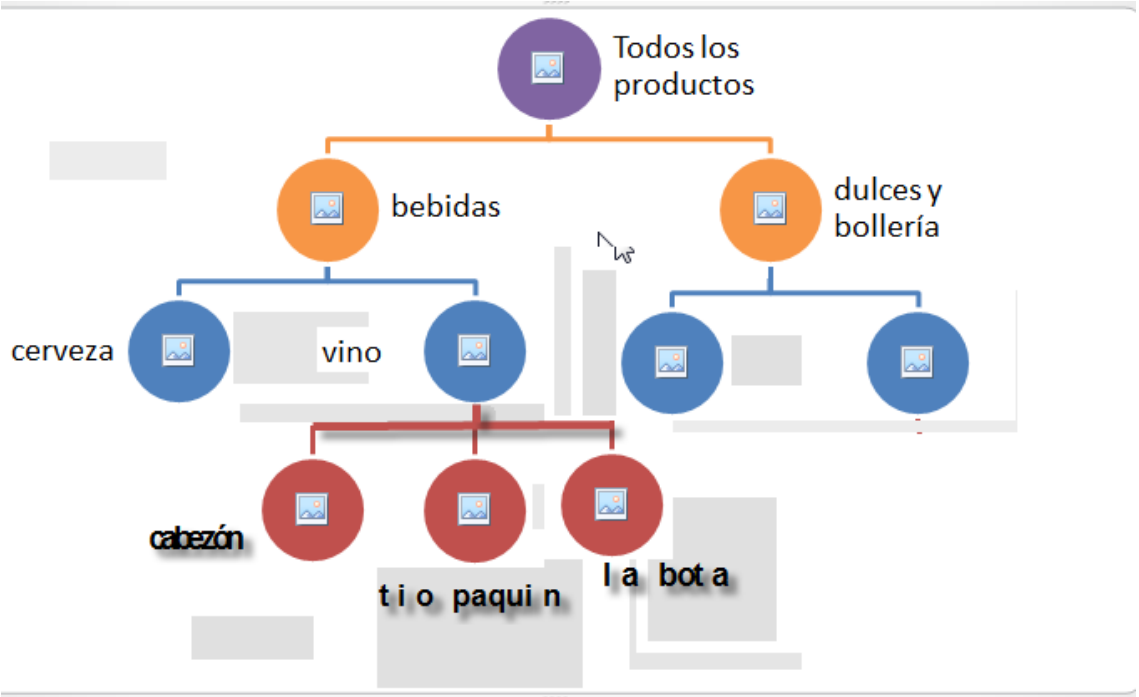
Para poder comparar la "intensidad" de la dependencia detectada entre distintas reglas de dependencia selecionadas no nos podría servir el estadístico χ^2 ya que este depende, entre otras cosas del tamaño de la muestra (número de instancias) y por ello suelen utilizarse los valores de otros indicadores el **interés** y la **correlación** (cuando sea de aplicación) . Por interés (o interés de dependencia) entenderemos el valor del indicador:

$$I(x,y) = \frac{P(x \cap y)}{P(x).P(y)}$$

5.2.Reglas de asociación multinivel

En ocasiones es difícil encontrar relaciones (asociaciones) interesantes debido a la "dispersión " de la información en la base de datos, esto es la existencia de una gran cantidad de atributos con respecto al pequeño número de ítems que hay en cada registro. Pensemos en el problema de la cesta del supermercado existe una enorme cantidad de productos sin embargo cada cliente (cada registro) adquiere unos pocos productos diferentes. Como consecuencia la mayor parte de la matriz de datos está vacía. Una forma de subsanar este problema es agrupar los atributos en categorías (en el supermercado podríamos considerar productos de limpieza, higiene, envasados, fruta y verdura, cárnicos, etc.). La idea es que al utilizar estas categorías en el proceso de aprendizaje de reglas sea más fácil encontrar reglas con la cobertura/confianza requerida.

Las reglas de asociación que utilizan varios niveles de conceptos para expresar las relaciones se denominan reglas multinivel. Para utilizar reglas multinivel es necesario proporcionar, además de los datos una jerarquía de conceptos que contiene un árbol de relaciones entre los atributos. En un nivel 0 o superior de la jerarquía tendríamos la conjunción de todos los coneceptos que irían subdividiéndose hasta llegar al nivel inferior dónde tendríamos los ítems particulares.



El algoritmo de aprendizaje de reglas multinivel sería similar al algoritmo general pero limitando las búsquedas por niveles y secuencialmente se va profundizando en los niveles siguientes con la posibilidad de mantener uniformes o no los criterios de cobertura

5.3.Reglas de asociación secuenciales

Este tipo de reglas expresarían patrosnes de comportamiento que se dan en distintos instantes de tiempo cercanos o en un distinto orden temporal. Una aplicación donde la consideración de la secuencialidad es muy importante es, por ejemplo, el análisis de las visitas a páginas web (Reglas del tipo “tras visitar página de cartelera a los pocos días se entra en una página de venta de entradas “)

El aprendizaje de este tipo de reglas se basará en encontrar las secuencias más frecuentes aunque conceptualmente es semejante al caso general es preciso adecuar los datos para mantener por un lado una identificación de cada instancia (cliente, visitante-web, o el individuo del tipo que sea) y una identificación del la secuencia temporal .