Regles d'associació

Luis Carlos Molina Félix
Ramon Sangüesa i Solé

P03/05054/01038

www.uoc.edu

Índex

Introducció 5

Objectius 6

1. Què són les regles d'associació? 7
1.1. Construcció de regles d'associació simples 9
1.1.1. Terminologia 9
1.2. Generació de regles 12
1.3. La clau de tot el mètode: com es poden trobar conjunts freqüents 14

2. Ponderació de les regles d'associació 16

Resum 18

Activitats 19

Exercicis d'autoavaluació 19

Bibliografia 21

Introducció

A les grans superfícies els interessa saber, per exemple, quins són els productes que formen el cistell de la compra dels seus clients. El fet de saber, per exemple, que dos productes tan dispars com la cervesa i els bolquers apareixen sovint en els grups de productes enregistrats en les transaccions dels punts de venda, pot permetre de redistribuir els objectes dins els prestatges de manera que s'augmenti la probabilitat que es comprin. Amb les informacions recollides a les caixes enregistradores, es construeix una base de dades en què els atributs són els diversos productes. Cada atribut representa una transacció que correspon a un únic client; per exemple, patates, llet, cafè, pa torrat, cervesa, bolquers, sucre, vi, pa integral, salmó, cassets verges, piles d'1,5 V, Coca-Cola, Fanta, formatge, pernil, embotits, cacauets, paté, galetes d'aperitiu, gots de plàstic, plats de plàstic i tovallons de paper.

Els **episodis freqüents** són coocurrències del mateix esdeveniment.

Exemple d'episodi freqüent

En el cas del cistell de la compra, l'esdeveniment és la compra d'un producte.

Hi ha altres informacions que serien útils però que aquesta mena de models no considera com són les quantitats i els preus de cada producte. No és rellevant. L'objectiu no és altre que conèixer la composició del cistell de la compra per tipus de producte. Per tant, la representació que tenen les transaccions no és altra que una tupla amb valors binaris (0 o 1, cert o fals, present o absent) que indica si un client ha comprat o no un tipus de producte determinat. No cal remarcar que el nombre d'atributs d'aquesta base de dades pot estar entorn dels milers. Vegeu també que moltes de les transaccions poden tenir una gran quantitat de valors a zero que indiquen que el client no ha comprat el producte corresponent.

Ús de l'anàlisi d'episodis freqüents

L'anàlisi d'episodis freqüents no es limita a l'àmbit de l'anàlisi del cistell de la compra.

Una última observació: per a aquest tipus d'estudi la identitat del client és completament irrellevant. Vosaltres mateixos us podeu imaginar altres aplicacions en àmbits com les transaccions bancàries, les assegurances, el suport a clients, accessos a un lloc web, el diagnòstic d'avaries freqüents, etc.

Objectius

Amb els materials didàctics associats a aquest mòdul, l'estudiant assolirà els objectius següents:

1. Conèixer les característiques principals de les regles d'associació.

2. Saber fer-ne ús per a diverses tècniques de mineria de dades.

1. Què són les regles d'associació?

Una **regla d'associació**, tal com nosaltres l'entenem, és una expressió de la forma

$$X \Rightarrow Y$$

on tant X com Y són conjunts d'elements. S'entén per **elements** els atributs que poden prendre valors binaris. Aquests elements permeten de formar una expressió lògica composta de conjuncions, disjuncions i negacions.

La interpretació que cal fer d'una expressió com $X \Rightarrow Y$ és bàsicament freqüentativa: s'observa que, en un determinat conjunt de dades, quan hi apareix un objecte que conté l'element X, tendeix a aparèixer-hi també l'element Y. Un exemple típic és que els clients que compren cassets (X) també compren piles (Y).

Regles d'associació

Atenció, no us confongueu. Una regla d'associació no és una regla de classificació.

Fixarem la nomenclatura i la notació per tal de poder explicar més formalment com actuen els diversos mètodes de construcció de regles d'associació.

Fixem primer la forma de què estan formats els antecedents d'aquesta mena de regles.

Sigui $L = \{e_1, e_2, ..., e_n\}$ un conjunt de literals (en el sentit que rep aquest terme en lògica). Cada e_i és un *element*. Sigui B un conjunt de dades, on cada component C d'aquest conjunt està definit sobre els mateixos literals: $C \subseteq L$.

Per a caracteritzar aquests elements segons la nomenclatura que hem estat fent servir fins ara hauríem de dir que C és un conjunt d'atributs binaris, $C = \{e_1, e_2, ..., e_n\}$, o que cada e_i pot prendre només els valors {0, 1} que indiquen que aquell atribut és absent (0) o present (1).

Suposarem que cada component C (un registre de la base de dades, una transacció de la base de dades de vendes, etc.) del conjunt de dades B té un identificador únic C_{id}.

Un conjunt d'elements X, $X \subset L$ es denomina **grup d'elements**. Direm que un component C del conjunt de dades (un registre o una transacció) conté un grup X si $X \subseteq C$.

Una regla d'associació és una implicació de la forma $X \Rightarrow Y$ que compleix les propietats següents:

Propietats de les regles d'associació

Les propietats de les regles d'associació es poden resumir amb la frase següent: intersecció buida d'antecedent i consegüent.

a) $X \subset L$, tots els elements del grup X pertanyen al conjunt d'elements sobre els quals està definit el conjunt de dades (registres, tuples, transaccions).

b) $Y \subset L$, tots els elements del grup Y pertanyen al conjunt d'elements sobre els quals està definit el conjunt de dades (registres, tuples, transaccions).

c) $X \cap Y = \varnothing$, no hi ha cap element repetit a banda i banda de la implicació.

La implicació de la regla d'associació $X \Rightarrow Y$ es compleix en el conjunt de dades B amb una **confiança *c*** si la c per cent (c %) dels components (registres, tuples, transaccions) de B que contenen el grup X també contenen el grup Y.

La regla d'associació $X \Rightarrow Y$ té un **suport *s*** dins el conjunt de dades B si el s per cent (s %) dels components de B contenen la unió del seu antecedent i el seu consegüent, $X \cup Y$.

Exemple de regla d'associació

Aquí tenim part de la base de dades, *B*, del nostre gimnàs Hyper-Gym:

Horari	Act1	Act2	Entrenador	Ús de la piscina
Tarda	Aeròbic	Stretch	No	Sí
Tarda	Aeròbic	Stretch	No	Sí
Matí	Aeròbic	Ioga	No	Sí
Tarda	TBC	Steps	No	No
Tarda	TBC	Stretch	No	Sí
Matí	Ioga	TBC	Sí	Sí
Matí	Stretch	TBC	No	Sí
Tarda	TBC	TBC	No	Sí
Tarda	TBC	TBC	No	Sí
Tarda	TBC	Steps	No	Sí

De fet, per a poder aplicar correctament el mètode de construcció de regles d'associació, caldria distingir entre els atributs que corresponen a la primera i segona activitat. Per tant, hauríem de desdoblar aquests dos atributs en Act1-TBC-Sí, Act1-Ioga, Act1-*Stretch*-Sí, Act1-Aeròbic, Act1-TBC, Act2-TBC, Act2-TBC-Ioga, Act2-*Stretch*, Act2-Aeròbic. Podem veure que la regla següent:

{'Tarda', 'Act1-TBC'} ⇒ {'No entrenador personal', 'Piscina'},

que ens assenyala que els usuaris de l'horari de tarda que practiquen com a primera activitat el TBC no tenen entrenador personal, però fan ús de la piscina té una confiança del 80% i un suport del 40%.

En canvi, la regla següent:

{'Tarda', 'Act1-TBC'} ⇒ {'No entrenador personal'},

que ens diu que els qui van al gimnàs a la tarda i fan com a primera activitat TBC no tenen entrenador personal té una confiança del 100% i un suport del 50% també.

La regla següent:

{'Tarda'} ⇒ {'No entrenador personal'},

que ens indica que els qui van a l'horari de tarda no tenen entrenador personal té una confiança del 100% i un suport del 70%.

El **problema de l'obtenció de regles d'associació** és aconseguir, a partir d'un conjunt de dades *B*, totes les regles que tenen un suport mínim determinat per l'usuari (que denotarem sup_{min}) i una confiança mínima també determinada per l'usuari (que denotarem $conf_{min}$).

Activitat

1.1. Amb les dades de l'exemple que acabem de presentar, si haguéssim demanat que s'obtingués un conjunt de regles amb un suport del 85% i una confiança del 95%, no s'haurien descobert les regles que hem descrit, però què hauríem obtingut?

Les regles d'associació descobreixen, no comproven.

És important adonar-se que en tractar de descobrir patrons o regles amb determinades característiques de qualitat (suport i confiança) ens situem en un terreny diferent del de la comprovació, que efectivament verifica una hipòtesi determinada. Som en un objectiu lleugerament diferent del de les proves d'hipòtesi en estadística, que intenten verificar si es compleix una determinada hipòtesi, per exemple que les persones que van a la tarda no tenen entrenador personal. Aquí no es tracta de determinar si una hipòtesi es compleix, sinó de saber quina hipòtesi es compleix. No es tracta de comprovar sinó de descobrir.

1.1. Construcció de regles d'associació simples

Vegem quin mètode una mica astut es pot construir per a obtenir regles que compleixin els requeriments de suport i confiança que els demanem. No se'ns escaparà que som davant un problema computacional gens trivial. En efecte, hauríem d'anar comprovant si un grup d'un únic atribut compleix els requeriments i, si els compleix, anar intentant afegir-hi altres atributs fins a trobar el conjunt màxim que permeti de construir la regla. El nombre de possibles combinacions que cal provar pot ser força gran.

1.1.1. Terminologia

En primer lloc, fixarem una mica de nomenclatura.

Donat un conjunt d'atributs *R*, una **base de dades binària *r* sobre *R*** és una col·lecció (o conjunt múltiple) de subconjunts de ***R***. Els elements de ***R*** els anomenem **ítems**, i els de ***r***, **línies**. El **nombre de línies de *r*** es denota $|r|$ i la **mesura de *r*** es denota de la manera següent:

$$\|r\| = \sum_{t \in r} |t| .$$

Utilitzarem les lletres majúscules de l'inici de l'alfabet *A*, *B*... per a denotar els ítems (o elements). Denotarem el conjunt de tots els ítems per *R*. Denotarem altres conjunts per les lletres finals de l'alfabet com *X* i *Y*. Els símbols en negreta denotaran la col·lecció de subconjunts, per exemple, *S*. Les bases de dades es denoten per lletres minúscules com *r*, i les línies per lletres com *t* i *u*.

La primera propietat que ens interessa per a tot conjunt d'elements de *R*, $X \subseteq R$, és la que ens permet de saber en quantes línies apareix el que hem anomenat *suport*. A continuació definim el suport d'una manera més formal.

Conceptes bàsics

- Concordança
- Suport
- Confiança

Sigui *R* un conjunt i *r* una base binària definida sobre *R*, sigui X, $X \subseteq R$ un conjunt d'ítems (elements). El conjunt d'ítems *X* concorda amb una línia $t \in r$, si $X \subseteq t$. El conjunt de línies de *r* amb què coincideix *X* es denota com a $M(X,r)$, per exemple, $M(X,r) = \{t \in r \mid X \subseteq t\}$. El suport de *X* en *r*, denotat per $sup(X,r)$, és el següent:

$$\frac{|M(X,R)|}{|r|}.$$

Escriurem simplement $M(X)$ i $sup(X)$ si no introduïm cap ambigüitat en parlar d'aquesta manera de la base de dades en un context determinat. Donat un llindar de suport $min\_sup \in [0, 1]$, el conjunt *X* està *suportat* si $sup(X,r) \geq min\_sup$.

En la literatura es fa servir també el terme *freqüència* per a definir el suport.

Donada una col·lecció de conjunts d'ítems, les regles d'associació fan una descripció de la manera en què apareixen diferents combinacions d'ítems en els mateixos conjunts.

Exemple de binarització de les bases de dades

Descobrir regles d'associació obliga a vegades a "binaritzar" la base de dades original, derivant nous atributs dels ja existents. Per exemple, donada una base de dades binària *r* sobre el conjunt $R = \{A, ..., K\}$:

Base de dades	
ID de la línia	**Línia**
t_1	{*A, B, C, D, G*}
t_2	{*A, B, E, F*}
t_3	{*B, I, K*}
t_4	{*A, B, H*}
t_5	{*E, G, J*}

podem prendre el subconjunt $\{A, B\}$. Llavors podem definir els subconjunts o regles d'associació que es poden organitzar sobre aquesta base de dades com a $M(\{A, B\},r) = \{t_1, t_2, t_4\}$. Prenem $sup(\{A, B\}, r) = 3/5 = 0{,}6$. Com ja hem remarcat abans en l'exemple de l'Hyper-Gym, podem expandir la base de dades en forma de relació en què tots els atributs $\{A, ..., K\}$ són binaris. Llavors la taula anterior quedaria així:

Expansió a valors binaris											
ID de la línia	A	B	C	D	E	F	G	H	I	J	K
t_1	1	1	1	1	0	0	1	0	0	0	0
t_2	1	1	0	0	1	1	0	0	0	0	0
t_3	0	1	0	0	0	0	0	0	1	0	1
t_4	1	1	0	0	0	0	0	1	0	0	0
t_5	0	0	0	0	1	0	1	0	0	1	0

Com ja hem dit abans, un conjunt d'atributs X serà prou freqüent (tindrà prou suport) si coincideix, com a mínim, amb una proporció *min_sup* de les línies en la base de dades *r*. El **llindar de suport** *min_sup* és un paràmetre que el dóna l'usuari i depèn de cada aplicació. El definim més formalment a continuació.

Sigui R un conjunt, r una base de dades binària sobre R i *min_sup* el **llindar de suport**. La col·lecció de conjunts que tenen prou suport a r segons a *min_sup* es denota per $F(r,min\_sup)$, i es defineix amb l'expressió següent:

$$F(r,min\_sup) = \{X \subseteq R \mid sup(X,r) \geq min\_sup\}.$$

Simplificarem la notació a $F(r)$ quan no hi hagi ambigüitat. La col·lecció de conjunts amb suport de mesura suficient l es denota per l'expressió següent:

$$F_1(r) = \{X \in F(r) \mid |X| = l\}.$$

Exemple de conjunt amb suport mínim

Suposem que el llindar de suport és 0,3. La col·lecció $F(r,0,3)$ de conjunts amb un suport m en la base de dades r de la taula d'expansió a valors binaris de l'exemple anterior és $\{\{A\}, \{B\}, \{E\}, \{G\}, \{A, B\}\}$, atès que cap altre conjunt diferent del buit no apareix en més d'una línia. El conjunt buit $\emptyset$ té trivialment el suport mínim en qualsevol base de dades binària; per tant, mai no el considerarem com un cas interessant.

Vegeu la base de dades de l'"Exemple de binarització de bases de dades" més amunt en aquest mateix subapartat.

Per a cada regla podem tenir el seu suport i la seva confiança.

Sigui R un conjunt, r una base de dades binària sobre R i $X,Y \subseteq R$ conjunts d'ítems. Llavors l'expressió $X \Rightarrow Y$ és una **regla d'associació sobre *r***. La confiança de $X \Rightarrow Y$ a r, denotada per $conf(X \Rightarrow Y,r)$, es defineix com segueix:

$$\frac{|M(X \cup Y,r)|}{|M(X,r)|}.$$

El **suport** $sup(X \Rightarrow Y,r)$ **de** $\boldsymbol{X \Rightarrow Y}$ **a** $\boldsymbol{r}$ és $sup(X \cup Y,r)$.

Escriurem $sup(X \Rightarrow Y)$ per simplificar, quan no hi hagi ambigüitat.

Donats un llindar de suport *min_sup* i un llindar de confiança *min_conf*, la regla d'associació $X \Rightarrow Y$ és vàlida en *r* si, i només si, $sup(X \Rightarrow Y,r) \geq min\_sup$ i $conf(X \Rightarrow Y,r) \geq min\_conf$.

Validesa d'una regla

Una regla és vàlida només si té el suport i la confiança mínims.

En altres paraules, la confiança $conf(X \Rightarrow Y,r)$ és la probabilitat condicional que una línia escollida aleatòriament dins *r* que coincideixi amb *X* també coincideixi amb *Y*. El suport de la regla és la quantitat d'evidència que es pot obtenir a partir de la base de dades a favor de la regla. Perquè una regla sigui considerada interessant ha de ser prou freqüent i forta (tenir prou suport i confiança).

La tasca de descobriment de regles d'associació...

... implica descobrir conjunts que compleixin certs requeriments.

En aquest cas la *qualitat* del model que cal construir s'avalua segons el suport i la confiança.

Ja podem establir amb més propietat quina és la tasca de descobriment de regles d'associació. En efecte, donats *R*, *r*, *min_sup* i *min_conf*, cal trobar totes les regles d'associació $X \Rightarrow Y$ que siguin vàlides a *r* respecte a *min_sup* i *min_conf*, perquè *X* i *Y* siguin conjunts disjunts i no buits.

Exemple de descobriment de regles d'associació

Tornem a la base de dades de la taula d'expansió a valors binaris de l'exemple anterior. Suposem que tenim un llindar de suport $min\_sup = 0{,}3$ i un llindar de confiança $min\_conf = 0{,}9$.

L'única regla d'associació amb parts esquerra i dreta no buida i vàlida en la base de dades és $\{A\} \Rightarrow \{B\}$. El suport de la regla és $0{,}6 \geq min\_sup$ i la seva confiança és $1 \geq min\_conf$. En canvi, la regla $\{B\} \Rightarrow \{A\}$ no és vàlida dins aquesta base de dades perquè la seva confiança és 0,75, més petita que *min_conf*.

Vegeu la base de dades de l'"Exemple de binarització de bases de dades" més amunt en aquest mateix subapartat.

Observeu que les regles d'associació no tenen propietats de monotonia respecte a l'expansió o contracció de la part esquerra. Si $X \Rightarrow Y$ és vàlida, llavors $X \cup \{A\} \Rightarrow Y$ no cal que sigui necessàriament vàlida, atès que $X \cup \{A\} \Rightarrow Y$ no tindrà necessàriament prou suport o confiança. D'altra banda, les regles d'associació tampoc no mantenen propietats de monotonia respecte a l'expansió de la seva part dreta. En efecte, si $X \Rightarrow Y$ és vàlida, llavors $X \Rightarrow Y \cup \{A\}$ no ha de ser necessàriament vàlida amb prou suport o confiança.

Monotonia

La part esquerra de les regles no conserva la monotonia respecte al suport.

La part dreta sí que ho fa.

Les regles d'associació només mantenen la propietat de monotonia respecte a la contracció de la part dreta. En efecte, si $X \Rightarrow Y \cup \{A\}$ és vàlida, llavors segur que $X \Rightarrow Y$ també ho és.

Activitat

1.2. Comproveu amb els exemples que heu vist en aquest subapartat les propietats de monotonia en les regles d'associació i les regles obtingudes.

Lectura recomanada

Trobareu la divisió de fases que proposa Agrawal en l'obra següent:

R. Agrawal; T. Imielinski; A. Swami (1993). "Mining Association Rules between Sets of Items in Large Databases". A: P. Buneman; S. Jajodia (ed.). *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data: SIGMOD'93* (pàg. 207-216). Washington: ACM.

1.2. Generació de regles

Agrawal proposa la divisió de fases següent dins tot mètode de construcció de regles d'associació:

a) Trobar totes les combinacions d'elements que tenen un valor de suport al qual s'ha fixat com a mínim, sup_{min}. Aquestes combinacions d'elements es denominen *grups grans* (*large itemsets*).

b) Utilitzar els grans grups per a construir finalment les regles. La idea és que si tenim, per exemple, dos grups grans (ABCD i AB) cal determinar si es compleix la regla AB ⇒ CD. Per a fer això cal veure si té la confiança necessària o superior a la confiança exigida per l'usuari, $conf_{min}$.

Normalment els algorismes de construcció de regles d'associació fan, almenys, dues passades sobre el conjunt de dades: l'una per a extreure els grans grups i l'altra per a construir les regles. La manera en què es construeixen els grans grups d'elements és a còpia d'anar considerant un nou element cada vegada i anar trobant les regles que tenen aquest nou element al consegüent i que tenen la confiança màxima (100%).

El procés habitual consisteix a seguir els passos següents:

1) Escollir una *llavor* de grups d'elements grans.

2) Utilitzar-la per a generar nous possibles grups grans, els grups candidats.

3) Avaluar-los comparant-ne el suport respecte al suport mínim; si superen el suport mínim, sup_{min}, es consideren grups grans.

4) Aquests grups es converteixen en les llavors de la fase següent.

5) Repetir el procés fins que no es trobin més conjunts grans.

Donarem a continuació l'algorisme de construcció de regles d'associació proposat per Agrawal i altres, i millorat pel mateix Agrawal el 1995.

La idea, com ja hem comentat, és obtenir primer tots els conjunts d'ítems $X \subseteq R$ i calcular-ne els suports respectius. Llavors cal comprovar separadament per a tot $Y \subset X$, $Y \neq \emptyset$ si la regla $X \setminus Y \Rightarrow Y$ és vàlida amb prou confiança. L'algorisme següent utilitza aquesta aproximació per a generar totes les regles d'associació vàlides amb la base de dades d'entrada. En el subapartat següent discutim la part clau de tot algorisme, que és com podem trobar els conjunts que tenen prou suport en la base de dades.

L'algorisme d'Agrawal presenta l'entrada i la sortida següents:

- Entrada: un conjunt R, una base de dades binària r sobre R, un llindar de suport *min_sup* i un llindar de confiança *min_conf*.
- Sortida: les regles d'associació que són vàlides en r respecte a *min_sup* i *min_conf* i els seus suports i confiances respectius.

Els passos que segueix l'algorisme d'Agrawal són els que enumerem a continuació:

1) Trobar els conjunts freqüents.
2) Calcular $F(r,min\_sup) := \{X \subseteq R \mid fr(X,r) \geq min\_sup\}$.

Lectures recomanades

Trobareu l'algorisme de construcció de regles d'associació proposat per Agrawal i l'algorisme millorat, respectivament, en les obres següents:

R. Agrawal; T. Imielinski; A. Swami (1993). "Mining Association Rules between Sets of Items in Large Databases". A: P. Buneman; S. Jajodia (ed.). *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data: SIGMOD'93* (pàg. 207-216). Washington: ACM.

R. Agrawal; H. Mannila; R. Srikant; H. Toivonen; I. Verkamo (1995). "Fast discovery of Association Rules". A: U.M. Fayyad; G. Piatetsky-Sahpiro; P. Smyth; R. Uthurusamy (ed.). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

3) Executar el bucle següent per a generar regles:

per a tot X ∈ F(r,*min_sup*) **fer**
  per a tot Y ⊂ X com a Y ≠ ∅ **fer**
    si *sup*(*X*) / *sup*(*X* \ *Y*) ≥ min_conf **llavors**
      crear la regla *X* \ *Y* ⇒ *Y* , *sup*(*X*) y *sup*(*X*) / *sup*(*X* \ *Y*);
    fsi
  fper
fper

Ara veurem per què funciona aquest algorisme. En primer lloc, cal notar que es verifica la igualtat següent:

$$conf(X \Rightarrow Y,r) = \frac{|M(X \cup Y,r)|}{|M(X,r)|} = \frac{sup(X \cup Y,r)}{sup(X,r)}$$

Clarament totes les regles d'associació $X \Rightarrow Y$ que produeix l'algorisme són validades respecte a la base de dades *r*:

$sup(X \Rightarrow Y) \geq min\_sup$, atès que (tercera línia del bucle):

$sup(X \cup Y) \geq min\_sup$ (línia 2) i $conf(X \Rightarrow Y) \geq min\_conf$.

L'algorisme genera totes les regles d'associació $X \Rightarrow Y$ que són vàlides amb la base de dades d'entrada.

Com que $sup(X \Rightarrow Y) \geq min\_sup$, llavors també $sup(X \cup Y) \geq min\_sup$ i $X \cup Y$ ha de ser dins de $F(r,min\_sup)$ (punt **2**). Si això passa, la possible regla $X \Rightarrow Y$ quedarà verificada (punt **3**). Com que $conf(X \Rightarrow Y) \geq min\_conf$, la regla corresponent, efectivament, es generarà (tercera línia del bucle).

1.3. La clau de tot el mètode: com es poden trobar conjunts freqüents

Cercar exhaustivament els conjunts freqüents –és a dir, conjunts amb prou suport, el que Agrawal anomena *large itemsets*– és una tasca clarament feixuga des del punt de vista computacional, per no dir impossible. Només resulta factible per als conjunts *R*: "petits". En efecte, l'espai de cerca de conjunts freqüents està format per $2^{|R|}$ subconjunts de *R*.

Exemple d'espai de cerca de conjunts freqüents

Amb $R = 100$, que representa un nombre minúscul de productes per a una cadena de supermercats, el nombre de configuracions per a explorar és aquesta xifra terrorífica: $1{,}26765060022823 \cdot 10^{30}$.

S'imposa una altra aproximació. El concepte clau és el de **conjunt candidat**.

La idea és anar construint progressivament conjunts més grans. Cal anar provant diverses cardinalitats o mesures dels conjunts. Per a cada cardinalitat, cada $l = 1, 2, \ldots$, es determina primer una col·lecció de conjunts C_l d'elements de mesura *l* tals que $F_l(r) \subseteq C_l$. Amb això s'obté la col·lecció de **conjunts candidats freqüents** (*large itemsets*) $F_l(r)$ que en calculen les freqüències.

Per a una base de dades gran amb un nombre gran d'atributs i amb conjunts candidats grans, el càlcul de la freqüència és prou costós. Per aquesta raó resulta útil minimitzar el nombre de candidats igualant-lo al de la fase de generació.

El problema de trobar conjunts candidats

Trobar conjunts candidats representa un problema que requereix clarament una aproximació diferent de l'exploració exhaustiva de combinacions possibles.

Per a construir una col·lecció de conjunts candidats petita però suficient cal fixar-se en les propietats dels conjunts de dades que ara descrivim.

a) Un subconjunt de línies és, com a mínim, tan freqüent (té el mateix suport) com el superconjunt dins el qual està inclòs. En altres paraules, el suport és monòton respecte a la contracció del conjunt per al qual es calcula. Això vol dir que per a qualsevol subconjunt X, Y d'ítems tals que $Y \subseteq X$, es compleix $M(Y) \supseteq M(X)$ i $sup(Y) \geq sup(X)$, i a més, si X té prou suport (és prou freqüent) llavors Y també.

b) A partir de la propietat anterior, podem aconseguir informació útil per a la generació de candidats. En efecte, donat un conjunt X, si qualsevol dels subconjunts de X no té prou suport, llavors podem descartar X com a conjunt freqüent possible –és a dir, com a candidat possible–, i l'eliminem del conjunt de candidats $C_{|X|}$.

La proposició següent, a més, estableix que això és suficient per a poder conèixer si tots els subconjunts de X tenen prou suport.

Proposició: sigui $X \subseteq R$ un conjunt. Si qualsevol dels subconjunts propis Y de X, $Y \subset X$, no té prou suport, llavors es verifiquen les propietats següents:

1) X no té prou suport.
2) Hi ha un subconjunt Z, $Z \subset X$, sense prou suport, de mesura: $|X| - 1$.

Demostració: el primer punt es deriva directament de l'observació del fet que si X té prou suport, llavors tots els seus subconjunts de X, $Y \subset X$ tenen prou suport. El mateix argument es pot aplicar per al segon punt: per a qualsevol Y, $Y \subset X$ hi ha Z tal que $Y \subseteq Z \subset X$ i $|Z| = |X| - 1$. Si Y no té prou suport, llavors Z tampoc no en té.

Exemple de conjunt candidat

Si sabem que un conjunt:

$$F_2(r) = \{\{A, B\}, \{A, C\}, \{A, E\}, \{A, F\}, \{B, C\}, \{B, E\}, \{C, G\}\},$$

podem concloure que $\{A, B, C\}$ i $\{A, B, E\}$ són els únics components de $F_3(r)$ possibles, atès que són els únics conjunts de mida 3 per als quals tots els subconjunts de mida 2 apareixen a $F_2(r)$. Ara ja sabem que $F_4(r)$ ha de ser buit.

Posteriorment a aquest algorisme, s'han introduït una sèrie de millores.

Lectures recomanades

Amb relació a l'algorisme d'Agrawal millorat, consulteu la primera de les obres esmentades a continuació. Vegeu també altres algorismes alternatius en les obres següents, la referència completa de les quals trobareu en l'apartat de bibliografia del mòdul:

R. Agrawal; H. Mannila; R. Srikant; H. Toivonen; I. Verkamo (1995).

R. Agrawal; R. Srikant (1994).

M. Houtsma; A. Swami (1993).

H. Mannila; H. Toivonen; I. Verkamo (1994).

2. Ponderació de les regles d'associació

Les regles d'associació resulten especialment temptadores en moltes aplicacions de marquèting, però tenen uns requeriments tant d'emmagatzemament com computacionals prou alts. A continuació n'esmentem alguns:

1) D'una banda, treballen sobre **grans conjunts d'atributs**: per exemple, els productes d'una cadena de supermercats.

Quant a transformació de dades, requereixen l'eliminació d'algunes de les dades que normalment es recullen en les transaccions (quantitat, preu, etc.). Com a pas següent obliguen a la binarització de les dades: és a dir, a transformar els valors de tots els atributs en els equivalents a {'Present', 'Absent'}. Això comporta l'extensió de la dimensió d'una taula de la base de dades que ja té d'entrada una dimensió prou alta.

A més, cal tenir en compte que en aplicacions reals el nombre d'atributs de la base de dades pot ser molt alt, entorn de milers en el cas d'aplicacions d'anàlisi del cistell de la compra. En cada transacció, hi haurà un nombre petit d'atributs presents. Això fa que, en construir els conjunts candidats de dimensió *i*, ens vegem obligats a guardar els resultats intermedis entre l'escombrada *i* sobre la base de dades i l'escombrada *i* + 1, no en memòria principal, sinó en disc.

Conjunts candidats de mesura *i*

Per a generar conjunts de mesura 5, per exemple, hem de partir dels de mesura 4 i efec-tuar comptatges de freqüència un altre cop. El cost de refer els comptatges no és trivial.

La manera en què es van construint els conjunts candidats és ideal per a intentar aportar solucions des del procés paral·lel.

Lectura recomanada

Vegeu una discussió general sobre el paral·lelisme en mineria de dades, amb una secció sobre regles d'associació, en l'obra següent:

M. Holsheimer; M.L. Kersten (1994). "Architectural Support for Data Mining. Knowledge Discovery in Dtabases". *Papers from the 1994 AAAI Workshop* (pàg. 217-228). Menlo Park (Califòrnia).

2) D'altra banda, es pot veure que el **cost** també depèn del suport exigit per l'usuari. Com més suport, més passades cal fer sobre la base de dades.

Igual que passa amb altres sistemes de construcció de regles (per exemple, de classificació) les col·leccions de regles d'associació resultants poden ser molt grans i complicar-ne la interpretació.

3) La **capacitat predictiva** que tenen les regles obtingudes és un altre aspecte que cal considerar. El **suport** i la **confiança** tenen relació amb la capacitat de predicció, evidentment, però no són determinants. Que una proporció molt gran de transaccions mostri una associació determinada entre l'objecte *X* i l'objecte *Y* amb una confiança prou alta, no vol dir que s'hi pugui generalitzar amb molta facilitat.

Tot i els aparents inconvenients esmentats, les regles d'associació resulten força comprensibles i s'utilitzen en moltes aplicacions pràctiques.

L'extensió de les regles d'associació a dades no binàries és un problema força interessant que té potencialment moltes aplicacions. Per a una discussió de mètodes de descoberta de regles d'associació sobre atributs que poden prendre diversos valors (vegeu Miller, 1997).

Les regles d'associació exploren, en general, dependències, la qual cosa és un problema que des de les bases de dades ja s'havia atacat amb anterioritat. En efecte, és important detectar en les dades relacions de dependència funcional que permetin de modificar després el disseny inicial de les bases de dades (vegeu Mannila, 1994).

Finalment, s'explora la relació entre les regles d'associació i les de classificació i s'han trobat mètodes que combinant totes dues perspectives permeten de millorar la precisió de les classificacions obtingudes (vegeu Liu, 1998). També és interessant veure com es poden treure regles de classificació amb poder predictiu a partir de grans conjunts de regles d'associació (consulteu Klemettinen, 1994).

Lectura recomanada

Hi ha una discussió interessant respecte a la capacitat predictiva de les regles d'associació en l'obra següent:

A. Siebes (1994). "Homogenenous Discoveries Contain no Surprises: Inferring Risk-Profiles from Large Databases". *Knowledge Discovery in Databases*. Papers from the 1994 AAAI Workshop (pàg. 97-108). Menlo Park (Califòrnia).

Resum

Les regles d'associació són un model que descriu un domini segons les dependències entre conjunts de valors.

Els atributs que apareixen a les parts esquerra i dreta d'una regla corresponen a atributs, els valors dels quals coocorren en el conjunt de dades original amb un suport i una confiança determinats per l'usuari.

Les dades han d'estar en format binari: els valors dels atributs només poden ser presents o absents. Això comporta normalment un primer procés de transformació del conjunt original de dades.

La complexitat de la descoberta de conjunts de regles amb les característiques requerides és un procés costós computacionalment. Cada grup d'atributs d'una certa cardinalitat C requereix almenys C escombratges previs per a trobar subconjunts de cardinalitat més petita.

La propietat de no-monotonia del suport per a subconjunts d'un conjunt donat és la que permet de millorar els algorismes de construcció de regles d'associació.

És un model prou comprensible, si bé les seves propietats estadístiques (en concret, característiques de predicció) no són prou clares.

Així i tot, tenen gran quantitat d'aplicacions.

Activitats

1. Accediu a l'adreça d'Internet que es dóna al marge i compareu les especificacions dels diversos programaris adreçats a construir-hi regles d'associació.

Per a fer l'activitat 1, accediu a l'adreça http://www.kdnuggets.com.

2. Per al problema que us havíeu proposat a l'activitat 1 del mòdul "Extracció de coneixement a partir de dades" d'aquest curs, us serveixen les regles d'associació? Quin mètode creieu que us resultaria més convenient?

3. Seguiu les activitats suggerides en el nucli de coneixement que correspon a aquest mòdul.

Exercicis d'autoavaluació

1. Donada la base de dades següent, que correspon a l'exemple de les lents de contacte:

Vegeu l'exemple de les lents de contacte en el subapartat 2.3 del mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

Edat	Diagnòstic	Astigmatisme	Llàgrima	Recomanació
Jove	Miop	No	Reduïda	Cap
Jove	Miop	No	Normal	Toves
Jove	Miop	Sí	Reduïda	Cap
Jove	Miop	Sí	Normal	Dures
Jove	Hipermetrop	No	Reduïda	Cap
Jove	Hipermetrop	No	Normal	Toves
Jove	Hipermetrop	Sí	Reduïda	Cap
Jove	Hipermetrop	Sí	Normal	Dures
Prepresbícia	Miop	No	Reduïda	Cap
Prepresbícia	Miop	No	Normal	Toves
Prepresbícia	Miop	Sí	Reduïda	Cap
Prepresbícia	Miop	Sí	Normal	Dures
Prepresbícia	Hipermetrop	No	Reduïda	Cap
Prepresbícia	Hipermetrop	No	Normal	Toves
Prepresbícia	Hipermetrop	Sí	Reduïda	Cap
Prepresbícia	Hipermetrop	Sí	Normal	Cap
Presbícia	Miop	No	Reduïda	Cap
Presbícia	Miop	No	Normal	Cap
Presbícia	Miop	Sí	Reduïda	Cap
Presbícia	Miop	Sí	Normal	Dures
Presbícia	Hipermetrop	No	Reduïda	Cap
Presbícia	Hipermetrop	No	Normal	Toves
Presbícia	Hipermetrop	Sí	Reduïda	Cap
Presbícia	Hipermetrop	Sí	Normal	Cap

a) Transformeu les dades per poder extreure'n regles d'associació.
b) Trobeu un conjunt de regles amb suport mínim 0,2 i confiança 0,4.
c) Podeu trobar cap conjunt amb suport i confiança superiors?

2. Donada la base de dades següent:

Horari	Act1	Act2	Entrenador personal	Ús de la piscina
Tarda	Aeròbic	Stretch	No	Sí
Tarda	Aeròbic	Stretch	No	Sí
Matí	Aeròbic	Ioga	No	Sí
Tarda	TBC	Steps	No	No
Tarda	TBC	Stretch	No	Sí
Matí	Ioga	TBC	Sí	Sí
Matí	Stretch	TBC	No	Sí
Tarda	TBC	TBC	No	Sí
Tarda	TBC	TBC	No	Sí
Tarda	TBC	Steps	No	Sí

a) Binaritzeu-la.
b) Quin és el suport de la regla {'Tarda', 'TBC', '*Stretch*'} ⇒ {'Entrenador personal', 'Piscina'}?
c) I la seva confiança?
d) Trobeu, si n'hi ha, la regla mínima de suport 0,9 i confiança 0,9.
e) Quin és el conjunt de suport i confiança mínims superiors a 0,1?
f) I quin és el conjunt màxim?
g) Repetiu els apartats **e)** i **f)** amb el valor 0,4.

Bibliografia

Agrawal, R.; Imielinski, T.; Swami, A. (1993). "Mining Association Rules between Sets of Items in Large Databases". A: P. Buneman; S. Jajodia (ed.). *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data: SIGMOD'93* (pàg. 207-216). Washington: ACM.

Agrawal, R.; Srikant, R. (1994). "Fast Algoritmics for Mining Tools". *Proceedings of the International Conference on Very Large DataBases,VLDB-94.*

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, I. (1995). "Fast discovery of Association Rules". A: U.M. Fayyad; G. Piatetsky-Sahpiro; P. Smyth; R. Uthurusamy (ed.). *Advances in Knowledge Discovery and Data Mining.* AAAI / MIT Press.

Holsheimer, M.; Kersten, M.L. (1994). "Architectural Support for Data Mining. Knowledge Discovery in Dtabases". *Papers from the 1994 AAAI Workshop* (pàg. 217-228). Menlo Park (Califòrnia).

Houtsma, M.; Swami, A. (1993, octubre). A: *Set-Oriented Mining of Association Rules*. Research Report RJ 9567. San José (Califòrnia): IBM Almaden Research Center.

Klemettinen, M.; Mannila, H.; Ronkainen, H.; Toivonen, H.; Verkamo, A. (1994). "Finding Interesting Rules from Large Sets of Discovered Association Rules". *Proceedings of CIKM'94, Conference on Information and Knowledge Management* (pàg. 40-407).

Liu, B.; Hsu, W.; Ma, Y. (1998, agost). "Integrating Classification and Association Rule Mining". *4th International Conference on Knowledge Discovery and Data Mining: KDD 98* (pàg. 23-27). Nova York.

Liu, B.; Hsu, W.; Ma, Y. (1998). *Buildi an Accurate Classifier using Association Rules. Technical report.*

Mannila, H.; Toivonen, H.; Verkamo, I. (1994). "Efficient Algorithms for Discovering Association Rules". *A Knowledge Discovery in Databases. Technical report WS-94-03.* American Association for Artificial Intelligence.

Mannila, H.; Räihä, K.-J. (1994). "Algorithms for Inferring Functional Dependencies from Relations". *Data & Knowledge Engineering* (vol. 12, núm. 1, pàg. 83-99).

Miller, R.J.; Yang, Y. (1997). "Association Rules over Interval Data". *ACM SIGMOD. International Conference on the Management of Data* (vol. 27, núm. 2, pàg. 452-461).

Siebes, A. (1994). "Homogenenous Discoveries Contain no Surprises: Inferring Risk-Profiles from Large Databases". *Knowledge Discovery in Databases.* Papers from the 1994 AAAI Workshop (pàg. 97-108). Menlo Park (Califòrnia, EUA).