El procés de descobriment de coneixement a partir de dades

Objectius, fases i problemàtica

Ramon Sangüesa i Solé

P03/05054/01033

UOC
www.uoc.edu

Índex

Introducció 5

Objectius 6

1. Descobriment de coneixement en grans volums de dades 7
1.1. Què entenem per coneixement? 9

2. Les fases del procés d'extracció de coneixement 11
2.1. Definició de la tasca de mineria de dades 11
2.2. Origen de les dades 17
2.3. Preparació de dades 18
2.3.1. Neteja de dades 19
2.3.2. Transformació de dades 22
2.3.3. Reducció de la dimensionalitat 25
2.4. Mineria de dades: el procés de construcció de models 28
2.4.1. Mecànica general del procés de cerca 29
2.4.2. Diversitat dels models de cerca 30
2.5. Avaluació i interpretació del model 33
2.6. Integració dels resultats en el procés 35
2.7. Observacions finals 36

3. Les eines de mineria de dades i les àrees relacionades 37
3.1. Eines de visualització 37
3.2. *Data warehouse* 44
3.3. Mètodes OLAP 47
3.4. Sistemes OLTP 48
3.5. Estadística 49
3.6. Aprenentatge automàtic 49

4. Cas d'estudi: la cadena Hyper-Gym 51

Resum 57

Activitats 59

Glossari 59

Bibliografia 60

Introducció

La mineria de dades s'ha d'inscriure dins un procés d'abast més ampli: el descobriment de coneixement dins de grans bases de dades, o KDD.

KDD és la sigla del terme anglès *knowledge discovery in databases.*

En aquest mòdul definim els objectius del procés de descobriment de coneixement a partir de bases de dades, en delimitem les diverses fases, ens centrem en la fase de mineria de dades i comentem els problemes més freqüents i importants inherents a tot el procés. Finalment, presentem en un cert detall l'exemple que serà el fil conductor durant tota l'assignatura, que ens permetrà de comparar les fortaleses i febleses de cadascun dels mètodes, i que en serà el veritable nucli temàtic.

Mineria de dades

La traducció literal del terme anglès *data mining* seria 'mineria de dades', encara que en català és preferible el concepte de *prospecció de dades*.

Objectius

Els materials associats a aquest mòdul permetran a l'estudiant d'assolir els objectius següents:

1. Tenir una idea clara de totes les fases que comporta un projecte de mineria de dades.

2. Conèixer la raó de ser de cadascuna de les fases del projecte.

3. Saber els problemes concrets que es presenten en cadascuna de les fases del projecte.

1. Descobriment de coneixement en grans volums de dades

La mineria de dades s'ha de considerar com una fase del descobriment de coneixement a partir de dades. Alguns autors utilitzen les dues denominacions (*descobriment de coneixement* i *mineria de dades*) indistintament. Nosaltres, però, ens adherim a la divisió establerta en el primer taller o *workshop* sobre KDD l'any 1989.

Així, quant al concepte de descobriment de coneixement en grans bases de dades (KDD) Piatetsky-Shapiro proposa la definició següent:

"El procés de KDD (*knowledge discovery in databases*) és el procés no trivial de descobrir patrons vàlids, nous, potencialment útils i comprensibles dins un conjunt de dades."

G. Piatetsky-Shapiro; C. Mateus; P. Smyth; R. Uthurusamy (1993). "KDD-93: Progress and Challenges in Knowledge Discovery in Databases". *AI Magazine* (núm. 15, vol. 3, pàg. 77-87).

Terminologia

KDD és la sigla del terme anglès *knowledge discovery in databases,* que correspon a 'descobriment de coneixement en bases de dades'.

Remarquem aquí uns quants aspectes que poden passar per alt en una primera lectura de la definició que acabem de donar:

Cal obtenir patrons vàlids, útils i comprensibles.

a) "patrons vàlids": s'ha d'entendre com a 'coneixement correcte contrastable amb la realitat'.

b) "potencialment útils": la utilitat està en relació amb l'objectiu que ens proposem en portar endavant el procés de mineria de dades.

Exemple d'utilitat potencial

El fet de saber que els clients amb rendes altes compren aparells electrònics pot ser útil si volem distingir els patrons de compra per a cada nivell de renda, però és completament inútil per a decidir si un client serà fidel a l'empresa els propers sis mesos.

c) "comprensibles": la comprensibilitat està en relació amb l'usuari que manega el coneixement, els patrons, etc., extrets de les dades. Per tant, no és una propietat absoluta dels patrons obtinguts.

Exemple de comprensibilitat

Posem per cas que volem predir la distribució geogràfica de vendes. Per a un estadístic, un model de regressió és prou comprensible. Per a un usuari menys preparat, potser és més comprensible veure un gràfic en pantalla.

Una altra definició que remarca la part que correspon a la mineria de dades és la que presentaren Holsheimer i Siebes.

Holsheimer i Siebes defineixen la mineria de dades de la manera següent:

"La mineria de dades és el procés de trobar models comprensibles a partir de grans volums de dades."

M. Holsheimner; A. Siebes (1994, gener). "*Data Mining*: the Search for Knowledge in Databases". *Report Tecnic CS-R9406*. Amsterdam: Centrum voor Winskunde en Informatica, CWI.

En aquesta definició, el que és important és la paraula *model*.

Un **model** és una descripció articulada i abstracta d'una realitat.

Per a descriure un edifici podem tenir un model que es basi únicament en conceptes estructurals o bé que només tingui en compte la distribució de conductes elèctrics. El que és important tenir en compte és que hi ha uns models que s'adiuen millor que d'altres al tipus d'utilitat que volem treure de les dades. Ja tornarem sobre això més endavant.

Vegeu el subapartat 2.5 d'aquest mòdul.

L'extracció del coneixement a partir de bases de dades és un procés complex que s'adreça a l'obtenció de models de coneixement a partir de dades recollides en una o més bases de dades quant a uns objectius determinats. Aquest procés de passar de dades a coneixement comporta un canvi de llenguatge d'expressió, i implica diverses fases.

Passar de dades a coneixement comporta un canvi de llenguatge d'expressió.

Els **objectius del projecte de mineria de dades** determinen el tipus de model de coneixement que hem d'extreure. També determinen que una dada o una relació entre dades sigui significativa o no en relació amb els objectius que ens hem marcat.

Relacions entre dades significatives

El coneixement que els clients que compren margarina viuen a la perifèria de la ciutat i els que compren mantega al centre pot ser una relació significativa o no segons l'objectiu que ens hàgim proposat per portar endavant el procés de descobriment.

El **canvi de llenguatge d'expressió** és important, i aquí és on rau la potència del descobriment de bases de dades. El que fa un procés d'aquesta mena és extreure de les dades un resum en un llenguatge diferent, però:

a) comprensible per a qui porta endavant el projecte quan n'analitza el coneixement extret, i

b) directament integrable en les operacions de l'empresa, encara que no sempre ho és.

Exemple de canvi de llenguatge d'expressió

Si partim d'un conjunt de registres de bases de dades extrets a partir dels codis de barres del terminal de punt de venda tindrem una repetició de valors de productes (per exemple, margarina) i codis postals (per exemple, 08012). Un resum que només tregui aquesta coocurrència representa un canvi de llenguatge poc potent. Un que ens tregui una regla del tipus:

Si codi postal = 08012 aleshores compra (margarina)

ha efectuat un canvi vers un llenguatge més comprensible.

1.1. Què entenem per coneixement?

El coneixement és creença justificada

El coneixement sempre està sotmès a revisió, és provisional. Per tant, l'extracció de coneixement a partir de dades és un procés continu.

No entrarem aquí en discussions gaire profundes respecte de què és el coneixement. A efectes pràctics podem assimilar el concepte *coneixement* a una informació que ha estat interpretada, classificada, aplicada i revisada, de manera que té un cert valor per a l'usuari de la informació inicial quant als seus objectius. Si es vol, podem adherir-nos al concepte de *coneixement* com a 'creença justificada', en el sentit que interpretem en relació amb el que sabem i amb el fet de relacionar les dades amb el que sabem. Així és com obtenim nou coneixement. Noteu que en aquesta accepció, el coneixement sempre està sotmès a revisió tenint en compte noves informacions.

El coneixement com a creença justificada

Podem tenir la creença intuïtiva que la majoria de clients d'una empresa són de Barcelona, però si les dades ens diuen que quinze mil clients són de Barcelona i trenta mil de la resta de Catalunya, haurem d'interpretar el fet com que la majoria de clients d'una empresa són de fora de Barcelona. La nostra creença inicial no queda, doncs, justificada, i no és un coneixement vàlid.

Només podem arribar a una justificació de la nostra creença quan relacionem les dades en brut (el nombre de clients de cada zona) amb una altra dada procedent del nostre coneixement previ (per exemple, que tenim quaranta-cinc mil clients). Faríem una altra interpretació si el nombre de clients fos dos-cents cinquanta mil: hauríem de relativitzar la creença, ja contrastada, que la majoria de clients (trenta mil) són de fora de Barcelona.

El que es demana a la mineria de dades és que aporti un primer nivell d'interpretació a força d'extreure relacions a partir de dades en brut. D'aquest primer nivell, encara se'n pot extreure coneixement més elaborat.

Hi ha un cert consens a considerar que en el procés d'extracció de coneixement, les **dades** són la matèria primera; quan algú els atribueix un significat, tenim **informació**; quan es fa una abstracció d'aquesta informació en relació amb els conceptes necessaris i relacionats amb un objectiu, tenim un **coneixement**.

La figura següent reflecteix molt bé la jerarquia d'importància que es desenvolupa en el procés d'extracció de coneixement mateix:

Font: Molina, 1998.

Exemple de procés d'extracció de coneixement

Aquí tenim part d'un conjunt de dades que correspon als clients d'un gimnàs. Els conceptes que utilitzem per a expressar el coneixement que volem extreure són la renda i el tipus d'activitat esportiva:

Grup	Centre	Horari	Act1	Act2	Renda	Edat	Sexe
1	1	Matí	Ioga	*Stretch*	Alta	68	D
2	3	Tarda	Ioga	*Steps*	Mitjana	32	H
4	3	Tarda	*Stretch*	Ioga	Baixa	44	D
2	3	Tarda	*Steps*	Peses	Mitjana	23	H
1	3	Tarda	Peses	*Stretch*	Mitjana	35	D
2	1	Matí	Peses	Peses	Mitjana	45	D
2	1	Matí	Ioga	*Steps*	Baixa	19	D
1	2	Matí	*Stretch*	*Stretch*	Alta	21	D
3	3	Matí	*Steps*	Aeròbic	Alta	56	H
3	1	Matí	Aeròbic	*Steps*	Baixa	30	D

Un primer resum de les dades ens diu que hi ha tres clients de renda alta, quatre de renda mitjana i tres de renda baixa. També ens diu que hi ha tres clients que fan com a activitat principal (Act1) ioga, dos que fan peses, dos que fan estiraments (*stretch*), un que fa aeròbic i dos que fan *steps*. Ara, aplicant tècniques tradicionals d'estadística sobre tot el conjunt original, que conté més registres que els que hem presentat en aquesta taula, podem extreure que hi ha una correlació de 0,8 entre la renda i l'activitat principal, la qual cosa ja és un primer canvi de llenguatge.

Aplicant alguna altra tècnica de mineria de dades podríem extreure una llista de regles d'aquest mena:

```
If ACT1 is steps Then
   Renda is Mitjana
   Rule's probability: 0, 981.
   The rule exists in 52 records.
   Significance Level: Error probability < 0,2.
```

que ens diu que si l'activitat principal d'un client és *steps*, es pot assegurar amb una probabilitat del 98% que la seva renda és mitjana, que això ha estat observat en cinquantados casos originals, i que la significació d'aquesta regla és menor del 2%, la qual cosa és un nou canvi de llenguatge i, fins i tot, més entenedor.

El canvi de llenguatge des de la matèria primera fins a una expressió del coneixement entenedora i/o operativa requereix diverses fases. Les comentem a continuació.

2. Les fases del procés d'extracció de coneixement

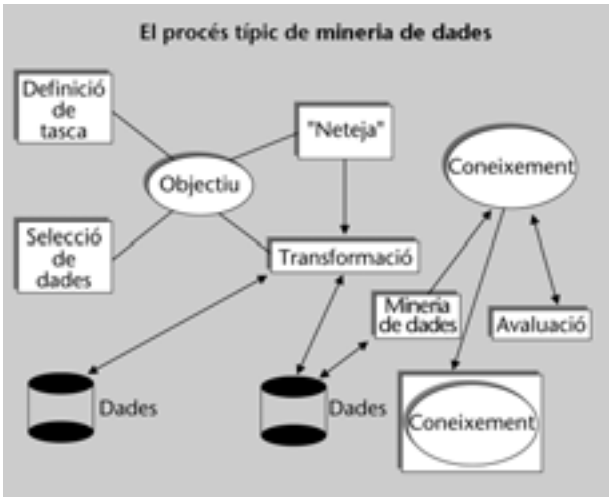
Com hem dit, la mineria de dades no és un procés que es dugui a terme en una sola fase. No és que decidim expressar un tipus d'objectiu determinat (predir la continuïtat dels clients, per exemple) i automàticament es generi un model que ens resolgui l'objectiu (que ens digui si un nou client serà fidel durant un cert temps). És més complicat i menys automàtic. S'ha de passar per diverses fases.

Lectura recomanada

Consulteu les fases del procés d'extracció de coneixement en l'obra següent:

U. Fayyad; R. Uthurusamy (ed.) (1995). *Proceedings of the First International Conference on Knowledge Discovey and Data Mining (KDD-95)*. Cambridge: MIT Press.

Hi ha diferents formulacions de les fases en què s'ha de dividir el procés. Seguirem Fayyad i esquematitzarem el **procés de descobriment** segons les fases que expressem gràficament en la figura següent:



Font: adaptat de Fayyad i altres (1996).

Aquestes fases són: definició de la tasca de mineria de dades, selecció de dades, preparació de dades, mineria de dades pròpiament dita, avaluació i interpretació del model, i integració. Com ja hem remarcat, aquest procés no és lineal, sinó que es realimenta i continua: nous canvis en la situació poden fer que el nostre coneixement deixi de ser correcte i calgui extreure'n de nou.

Vegeu el subapartat 1.1 d'aquest mòdul.

Ara comentarem cada fase detallant-ne els objectius, les activitats que li corresponen i la problemàtica que presenta.

2.1. Definició de la tasca de mineria de dades

Aquest és el punt en què precisem quin és l'objectiu del projecte de mineria de dades. És el moment de decidir si es tracta, per exemple, de trobar dependències entre variables (la renda i l'activitat, per exemple), si volem saber què distingeix un tipus d'usuari d'un altre, si volem conèixer tendències, etc.

Important

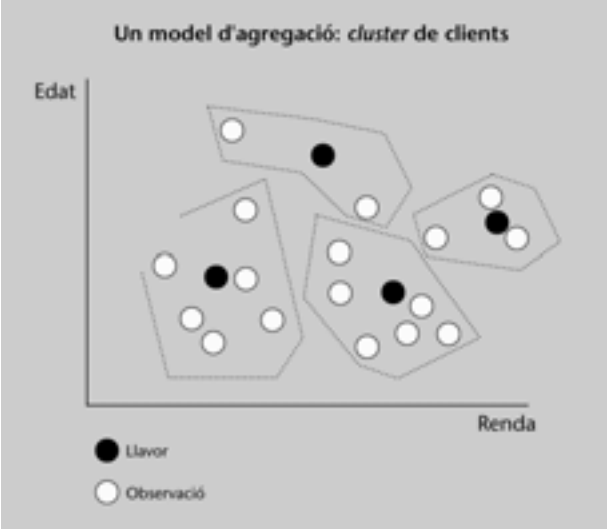
És molt important aclarir de bon principi quin és l'objectiu del projecte de mineria de dades.

La tasca principal de cadascun d'aquests projectes normalment es pot assimilar a alguna de les següents:

a) Trobar similituds i **agrupar objectes semblants**: correspon a un projecte en què tenim *poca* informació del domini i volem començar a tenir-ne una idea més clara. Els models típics per a assolir aquests objectius són els **models d'agregació** (***clustering***) procedents de l'anàlisi de dades o de l'aprenentatge automàtic i els models associatius.

Exemple de model d'agregació

Un exemple típic de projecte de mineria de dades per a trobar similituds consisteix a trobar grups de clients semblants.



b) Classificar objectes: la tasca d'aquests projectes de mineria de dades no és exactament igual que la del punt anterior. Aquí el més habitual és partir d'una situació més informada, sabent que hi ha grups ja definits. El que es vol és estudiar millor les diferències entre un grup i un altre, les seves característiques peculiars. Normalment la classificació d'objectes és el pas previ per a efectuar prediccions; és a dir, per a saber alguna conducta d'interès a partir d'unes quantes dades. Això es pot refinar obtenint coneixement predictiu.

Exemple de classificació d'objectes

Si ja hem pogut separar els nostres clients en diversos grups, volem saber quin és l'atribut que distingeix millor un grup de clients o l'altre (la renda, la seva àrea de residència, etc.).

Alguns **models classificatoris típics** són els arbres de decisió com CART, ID3, C4.5 i C5.0, les xarxes neuronals per classificació, i les regles de classificació:

- Els **arbres de decisió** donen una estructura en què a cada node se li fa una pregunta sobre un atribut determinat; el valor que prengui indica que cal seguir la branca corresponent a l'atribut. Els nodes finals corresponen a conjunts d'exemples que pertanyen a la mateixa classe. Resseguint les branques des de l'arrel fins a les fulles s'obté un seguit de condicions que permeten de classificar les noves observacions. Per exemple, en la figura següent es pot veure l'estructura d'un arbre de decisió per a predir si un client d'un gimnàs demanarà tenir un entrenador personal o no:

Lectures complementàries

Trobareu més informació sobre el model de decisió CART, els models de classificació ID3, C4.5 i C5.0 i els models de xarxes neuronals, en les obres següents respectivament:

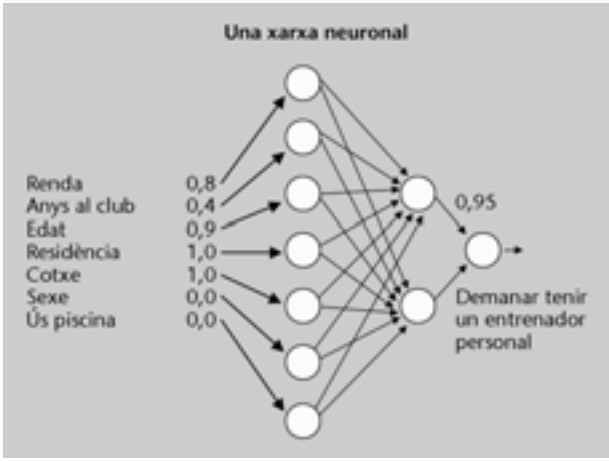
L. Breiman; H. Friedman; R. Olshen; C. Stone (1984). *Classification and Regression Tress*. Wadsworth: Chapman & Hall.

J.R. Quinlan (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

B.D. Ripley (1994). "Neural Networks and Related Methods for Classification". *Journal of the Royal Statistical Society* (vol. 3, núm. 56, pàg. 409-437).

Es pot comprovar que hi ha una relació clara entre els arbres de decisió i certs mètodes de construcció de regles de classificació. Normalment, els mètodes de construcció d'arbres de decisió s'acompanyen amb més informació, que permet de saber per a cada node i els seus possibles valors com queda dividit el conjunt d'observacions en diverses classes.

- Les **xarxes neuronals** també són bons models classificatoris i predictius. Tenen certes analogies amb la manera en què estan connectades les neurones cerebrals i s'organitzen en forma de molts nodes de procés connectats que donen una o més sortides. Les diverses capes de nodes estan connectades entre si amb més o menys força a través d'uns factors o pesos que indiquen la importància de les sortides produïdes per cada node. En conjunt, el que fan és aprendre a ajustar els valors d'aquests pesos per a ser tan predictives com sigui possible. En la figura següent podeu veure una xarxa neuronal utilitzada també per a predir si un client demanarà entrenador personal o no:



Les entrades de la xarxa recullen els valors d'altres atributs que cal tenir en consideració (sexe, edat, renda, residència, etc.) i la sortida té el valor 1 si efectivament, a la combinació de valors (descripció d'un client), li correspon normalment un entrenador personal.

- Les **regles de classificació** tenen una expressió d'aquesta forma:

$$\text{Antecedent} \rightarrow \text{Conseqüent}.$$

Aquestes regles imposen una sèrie de condicions sobre els valors que prenen els atributs d'entrada per tal d'indicar a quina classe poden pertànyer. En el nostre exemple, la classe estava determinada per l'atribut *Entrenador personal*. Hi ha dues classes, la dels clients que demanen entrenador i la dels que no; per tant, la forma que té la regla ens indica sota quines condicions un client demanaria entrenador:

```
If ACT1 is Steps Then
   Entrenador Personal is No
   Rule's probability: 0,981.
   The rule exists in 52 records.
   Significance Level: Error probability < 0,2.
```

Aquesta no és l'única forma que admeten les regles de classificació.

Altres formes adoptades per les regles de classificació

Per exemple, les regles de classificació que s'obtenen amb mètodes com el CN2 (Clark, 1991) o les que s'obtenen a partir de mètodes de lògica inductiva (Muggleton, 1992; Muggleton i Raedt, 1994) són una conjunció de condicions lògiques sobre els valors que poden prendre els atributs (igualtat, comparació, etc.) i no acostumen a tenir cap indicació respecte al grau de validesa, significació o error. Per exemple, la regla anàloga a l'anterior, que es pot obtenir utilitzant CN2, té la forma següent:

If (ACT1 = *Steps* $\Rightarrow$ (Entrenador = no).

Lectures complementàries

Trobareu algunes de les diferents formes que adopten les regles de predicció en l'obra següent:

J. Cendrowska (1987). "PRISM: An algorithm for inducing modular rules". *International Journal of Man-Machine Studies* (vol. 27, núm. 4, pàg. 349-370).

P. Clarck; R. Boswell (1991). "Rule Introduction with CN2: Some Recent Improvements". A: Y. Kodratoff (ed.). *Proceedings of the European Working Session on Learning: Machine Learning* (pàg. 151-163). Berlín: Springer Verlag.

c) Predir: es tracta d'obtenir coneixement que ens permeti de predir allò que ens interessi. Té moltes similituds amb la classificació. Efectivament, en una classificació binària, de dues classes possibles (clients que demanen entrenador i els que no en demanen) es tracta de predir el valor de la classe dins un conjunt limitat de valors (en aquest cas [0, 1], on el 0 representa els qui no demanen entrenador i 1 els qui sí que en demanen). En altres casos amb més classes, la classificació es pot entendre com un procés de predicció que ha d'indicar el valor d'aquesta etiqueta.

Exemples de models predictius

Alguns exemples de models que ens permeten d'arribar a aquest objectiu són els arbres de decisió i els models predictius clàssics de l'estadística (per exemple, els models de regressió). Aquí podríem incloure també els models que no detecten valors concrets (per exemple, sí o no per a la fidelitat d'un client), sinó que detecten tendències. Típicament, inclouríem en aquest tipus els estudis de sèries temporals i totes les variacions que s'han aportat des de l'aprenentatge automàtic, com les xarxes neuronals per a la predicció de sèries temporals.

Ara bé, a vegades el que ens interessa predir no és un atribut que prengui valors en un conjunt finit de valors (numèrics o no), com acabem de veure. Per exemple, ens pot interessar predir la durada de les trucades que efectua un departament a partir d'altres dades. En aquest cas intentem obtenir un valor que varia en una escala contínua de valors (fins i tot podríem tenir decimals); per tant, el nombre d'*etiquetes de classe* possibles és infinit: n'hi ha tantes com nombres reals hi hagi entre el valor mínim i el màxim detectats.

d) Descriure: tot i que obtenir agrupacions d'objectes és un primer nivell de descripció, per *coneixement descriptiu* ens referim normalment a trobar i expressar associacions significatives o causals entre diverses variables.

Un exemple típic d'aquesta mena de models són les **xarxes bayesianes** i, encara que amb menys potència, les **regles d'associació**. Les presentem a continuació:

Exemple de model descriptiu

En el cas de les assegurances, un model descriptiu posarà en relleu que els factors determinants de la perillositat d'un client són la seva edat, el color del cotxe que compri i el valor dels atestats presentats el darrer any. A més, ens indicarà el tipus de força d'associació entre totes aquestes variables.

- Considerem una **xarxa bayesiana** d'exemple, com la que veiem en la figura següent:

Aquesta xarxa bayesiana indica coses interessants. Està extreta d'un conjunt de dades simplificades que relaciona diverses variables utilitzades en assegurances per a predir si un client és de risc o no. Només inspeccionant visualment el model, ens diu que les variables que influeixen més directament en la classe de risc d'un client són: el color del cotxe, el quilometratge i l'any de matriculació. També podem veure que altres relacions no influeixen directament en la classe: per exemple, l'edat de la persona que condueix, que hi influeix indirectament a través del color i en combinació amb altres variables, com la renda del client. A més, per a cada enllaç podem saber la probabilitat condicional existent entre les diverses variables connectades. En aquest cas, podem veure els valors següents:

Taula de probabilitats condicionals			
Sexe	**Edat**	**Quilometratge alt**	**Quilometratge baix**
Dona	Jove	0,708	0,292

Taula de probabilitats condicionals			
Sexe	**Edat**	**Quilometratge alt**	**Quilometratge baix**
Dona	Vella	0,582	0,418
Home	Jove	0,714	0,296
Home	Vell	0,255	0,645

Per tant, tenim una idea de la influència mútua entre les variables, quines són realment rellevants per a conèixer el valor d'una variable donada i, a més, quan observem una determinada combinació de valors, podem predir els valors més probables per a la resta de variables relacionades.

- Les **regles d'associació** cerquen trobar coocurrències prou significatives entre grups de variables. L'únic requeriment que imposen és que s'indiqui el "nivell de suport" que es vol que tinguin a partir de les dades, la proporció de les dades que es vol cobrir amb aquesta regla. Llavors cal trobar grups de variables i combinacions de valors que arribin a tenir aquest grau de suport. Per exemple, es dóna el cas següent:

Una regla d'associació

(Renda > 35000000) & (Edat > 40) & (Residència = C) ⇒
⇒ (Entrenador = Sí) & (Piscina = Sí) (0,95)

En altres termes, en un 95% dels casos, quan el client té una renda que supera els 3,5 milions, una edat superior als quaranta anys i viu a la zona que hem designat com a *C*, llavors també resulta que ha demanat entrenador personal i utilitza la piscina.

e) Explicar: aquí es tracta d'obtenir models que ens puguin donar les raons de per què s'ha donat un comportament determinat (Anand, 1992). Per exemple, si havíem predit que un client determinat adquiriria una sèrie de productes determinada i no ho ha fet, a què es pot deure? Exemples de models d'aquesta mena són les xarxes bayesianes, que permeten de trobar quin és el conjunt de variables i amb quins valors per a un valor observat per una variable determinada, que més probablement poden donar raó del valor observat.

Lectura complementària

Trobareu més informació sobre els models explicatius en l'obra següent:

T. Anand; G. Kahn (1992). "SPOTLIGHT: A Data Explanation System". *Proceedings of the Eighth IEEE Conference on Applied Artificial Intelligence* (pàg. 2-8). Washington D.C.: IEE Computer Society.

Per exemple,...

... les xarxes bayesianes poden ser utilitzades com a models descriptius, predictius i explicatius.

És important tenir clar que per a cadascuna d'aquestes tasques es pot utilitzar més d'un model.

A més, per a construir cada tipus de model tenim a la nostra disposició diversos mètodes que estaran inclosos, o no, en les eines comercials existents en el mercat. D'altra banda, cadascuna d'aquestes es pot adir, o no, amb els condicionants tècnics del sistema de base de dades de què disposem i del sistema d'informació en què s'ubica.

Per tant, a l'hora de definir la tasca de mineria de dades hem de ser capaços de fer el següent:

- Aproximar l'objectiu a alguna de les tasques genèriques que hem esmentat. Aquest punt és el més crític i per al qual resulta més difícil donar regles d'aplicabilitat general. Decidir que ens cal conèixer millor els nostres clients, per exemple, és un objectiu massa general que s'ha de precisar. Volem conèixer quins grups de clients tenim? Llavors hem d'obtenir un model d'agregació. Volem determinar les característiques distintives de grups separats geogràficament o per renda? Llavors hem de classificar. Però sempre queden objectius que requereixen un esforç combinat. Potser haurem d'agrupar primer, classificar després i extreure un model predictiu finalment.

- Decidir el model que ens cal. Aquest punt ens obliga a conèixer bé els diversos models que hi ha, per a què serveixen i per a què no. Els podem comparar, per exemple, en termes de la seva capacitat expressiva, de la seva comprensibilitat, de la seva facilitat d'implementació i integració al sistema d'informació de l'empresa.

Factors de comparació

Comprensibilitat
Facilitat d'implementació
Facilitat d'integració

- Seleccionar el mètode necessari per a construir-lo. Aquest tercer punt va molt lligat al segon i ens permetrà d'avaluar les possibilitats i les característiques de les diverses eines existents: complexitat de mètode, cost computacional.

Els dos últims punts estan, a més, condicionats per l'entorn en què s'ha de desenvolupar el projecte per la implementació dels resultats: sistema d'informació i entorn d'explotació.

Un cop definit quin és l'objectiu i un cop l'hem posat en relació amb la tasca principal del projecte, quins models ens interessen més i quins mètodes i eines ens calen, cal passar a trobar la matèria primera: les dades. No és pas tan senzill com sembla. 

A continuació entrem a descriure detalladament cada fase, la qual cosa ens permetrà d'abordar els mòduls restants amb una perspectiva global.

2.2. Origen de les dades

Trobar les dades que ens calen és més fàcil de dir que de fer. És l'equivalent en la mineria real a saber on hem de començar a perforar per trobar petroli.

Idealment, la **tecnologia de *data warehousing*** (literalment, 'magatzems de dades') (Inmon, 1996) està especialment orientada a facilitar la localització de les dades dins una empresa en relació amb diversos tipus d'utilitats. Un *data warehouse* integra dades procedents de les diverses dades de cada departament d'una empresa.

Data warehousing

Una tecnologia especialment dirigida a facilitar els processos de la mineria de dades.

Amb aquesta tecnologia s'assegura, per exemple, que per als diversos registres, el camp *nivell de risc del client* té el mateix significat, tot i que ha resultat de fusionar les diverses interpretacions que li dóna el departament de cobraments o el de pòlisses, si parlem, per exemple, d'una companyia d'assegurances. A més, els *data warehousing* guarden dades històriques de l'empresa, de manera que permeten de predir tendències. En principi, són una tecnologia dirigida a la presa de decisions. Quin problema presenten, doncs? Que la majoria d'empreses no tenen aquesta tecnologia instal·lada! I això és més cert com més petita és l'empresa. En conseqüència, el més normal és que, com a primera fase, s'hagin de localitzar les fonts de dades a partir del següent:

a) Les bases de dades disperses pels diversos departaments que considerem que poden ser rellevants per al projecte que portem endavant.

b) Les bases de dades transaccionals, que recullen les operacions dia a dia i que acumulen informació històrica que pot ser rellevant per a obtenir el model que cerquem.

Cap de les dues operacions no és senzilla. De fet, cal que l'empresa –si no té un *data warehouse*– porti endavant una bona política de gestió de dades amb vista a la presa de decisions. Probablement, caldrà crear i introduir nous processos per a obtenir les dades que ens calen.

Necessitat d'introduir nous processos

Quan una agència de viatges va decidir que volia recomanar millor als seus clients les possibles destinacions de viatge en relació amb quins viatges havia fet cada client anteriorment, es va adonar que no guardava informació històrica útil de cada client més enllà de dos mesos i va haver d'introduir un nou procés i modificar la base de dades corresponent per disposar de dades anteriors de més temps.

Normalment, el primer projecte de mineria de dades que s'emprèn en una organització fa paleses les mancances que hem esmentat i es pot utilitzar com una oportunitat per a redissenyar la política de gestió de dades amb vista a millorar la presa de decisions.

2.3. Preparació de dades

Un cop localitzades les fonts de dades, aquestes s'han de preparar perquè se'ls pugui aplicar els mètodes o eines que construiran el model volgut. Aquesta fase, encara que sembli senzilla, conjuntament amb la de selecció de dades, s'emporta el 80% de l'esforç en els projectes de mineria de dades de nova implantació.

Lectura complementària

Trobareu més informació sobre la fase de preparació de dades en l'obra següent:

E. Simoudis; U.M. Fayyad (1997, març). "Data Mining Tutorial". *First International Conference on the Practical Applications for Knowledge Discovery in Data Bases*. Londres.

En aquest punt cal assegurar-se d'unes quantes coses:

a) Que les dades tinguin la qualitat suficient: és a dir, que no continguin errors, redundàncies o que presentin altres menes de problemes. També s'en-

tén la qualitat de les dades com la propietat que asseguri la qualitat del model resultant (per exemple, que asseguri que serà prou predictiu, si es tracta de crear un model d'aquest tipus). No cal dir que aquesta darrera accepció de la paraula *qualitat* és encara més problemàtica de garantir.

b) Que les dades siguin les necessàries: potser n'hi haurà que no ens caldran i potser n'hi haurem d'afegir. Normalment és molt estrany que les dades que realment necessitem ja hagin estat recollides pel sistema amb el propòsit de portar endavant justament el tipus d'estudi de mineria de dades que volem emprendre. Això normalment comportarà afegir camps nous a les diverses relacions d'una base de dades procedents d'altres relacions o d'altres bases de dades.

c) Que estiguin en la forma adequada: molts mètodes de construcció de models requereixen que les dades estiguin en un format determinat que no ha de coincidir necessàriament amb el format en què que estan emmagatzemades. Comprovarem que hi ha diverses interpretacions d'aquesta diferència de format que obliguen a efectuar diferents transformacions que estudiarem en el seu moment. La més típica és que les dades siguin valors numèrics continus i els mètodes només admetin valors discrets.

A continuació comentem breument les tècniques que es fan servir per a assegurar els tres aspectes que hem comentat. Són la neteja de dades, la transformació de les dades i la reducció de la dimensionalitat.

2.3.1. Neteja de dades

La **neteja de dades** consisteix a processar les dades per tal d'eliminar les que siguin errònies o redundants.

També hi ha qui inclou en aquesta fase el pas d'eliminar alguns dels atributs de les dades, el que s'anomena **selecció d'atributs**, que cerca assegurar que amb menys dades es puguin obtenir models de la mateixa qualitat. Comentarem aquest aspecte més endavant.

Vegeu la reducció de la dimensionalitat en el subapartat 2.3.3 d'aquest mòdul.

Tot i estar en la forma adequada, acostuma a passar que les dades no són perfectes en un 100%. Les dades introduïdes a mà o bé procedents de la fusió de diverses bases de dades acostumen a mostrar factors de distorsió importants. Revisem quins són els més freqüents:

Un exemple de dades incompletes...

... es té quan s'emplena el camp corresponent al carrer però s'oblida, o no es representa correctament, el corresponent al número o al pis. Una manera de completar les dades que manquen és substituint-les per un "valor raonable".

a) Dades incompletes: pot passar que tinguin un valor "indefinit"; és a dir, que falti algun valor dels que són comuns per als registres de la base de dades que considerem. I això es dóna especialment quan a l'hora de dissenyar

el procés corresponent d'entrada de dades es decideix que determinats atributs no són obligatoris o tenien format lliure.

Normalment, el que es fa és "completar" els valors que no apareixen. Per als valors numèrics es complementa calculant el valor mitjà observat per a un atribut.

Completar valors numèrics

Si, per exemple, ens manca informació sobre el salari d'un empleat, podem adscriure-li la mitjana de valors dels salaris de la resta d'empleats de la companyia o de la seva secció. No cal dir que aquest mètode presenta els seus problemes i pot induir a errors o rebaixar la qualitat del model resultant. Però a vegades és tot el que es pot fer.

b) Dades redundants: a vegades es repeteixen tuples que corresponen al mateix objecte.

Exemple de repetició de tuples

Sovint es donen situacions en què un mateix client és donat d'alta diverses vegades fins i tot amb el mateix número d'identificació. Aquest és un resultat típic de la fusió de dades procedents de bases diferents. Una variació són els casos en què un conjunt de valors corresponents al mateix objecte rep identificadors diferents. Posem per cas un client que ha estat donat d'alta diverses vegades però amb identificadors diferents. Aquí tenim un exemple senzill que relaciona els clients i les seves dades amb els centres de compra on adquireixen els seus productes dins una cadena de botigues.

Identificador	Nom	Adreça	Centre
24.567	**Poch**	**Roca, 33-1**	**1**
32.456	Martínez	Travessera, 222	2
24.567	**Poc**	**Roca, 33-1**	**1**
33.400	Sala	Diagonal, 556	1
33.441	Arregui	Diagonal, 222	2

No sabem si el client 24.567 es diu "Poc" o "Poch", però hi ha moltes probabilitats que es tracti de la mateixa persona. És clar que ens calen eines per a poder decidir quina de les dues interpretacions és la correcta o bé si es tracta d'un problema d'assignació d'identificadors. Aquest és un procés difícil i moltes vegades cal fer servir eines estadístiques tan sols per a poder-lo detectar. És molt normal, però, en dades que procedeixen d'informació voluntàriament (mal) donada pels clients. Els bancs i altres entitats coneixen el perquè d'aquesta conducta d'alguns clients, que volen ocultar els diversos canvis de domicili o "despistar".

c) Dades incorrectes o inconsistents: molt comú quan el tipus de valors que pot rebre un atribut no està controlat perquè està declarat com a "text lliure" o bé està definit com un tipus determinat (cadena alfanumèrica, per exemple), però no s'han mantingut els processos de control d'errors necessaris. Per exemple, un client amb una edat superior a cinquanta anys que rep descomptes per Carnet Jove; o per exemple, un client que té un carrer que no correspon al codi postal que té assignat; o bé una població que no correspon al codi postal.

Exemple de dades inconsistents

En el cas següent, si suposem que només hi ha deu botigues en una determinada cadena comercial, és evident que passa alguna cosa estranya amb el client "Martí":

Identificador	Nom	Adreça	Centre
24.567	Poch	Roca, 33-1	1
32.456	Martínez	Travessera, 222	2

Identificador	Nom	Adreça	Centre
33.345	**Martí**	**Roca, 33-1**	**144**
33.400	Sala	Diagonal, 556	1
33.441	Arregui	Diagonal, 222	2

És evident que, donat el número de centres de la cadena, el valor de "centre" del client "Martí" no és correcte.

d) Errors de transcripció: molt típics i que poden donar lloc a algun dels problemes anteriors. Per exemple, majúscules/minúscules.

Exemple d'error de transcripció

El mètode que explori les dades pot decidir que "Barcelona" i "BARCELONA" són dues poblacions diferents.

e) Dades envellides: certes dades es converteixen en incorrectes perquè no han estat actualitzades adequadament.

Exemples de dades envellides

Un exemple típic és el domicili o domiciliació bancària quan no es notifiquen els canvis corresponents. Un altre cas es podria donar quan es tracta de treballar amb rangs d'edats i en les dades cada persona apareix amb el rang d'edat corresponent. Posem, per exemple, que en comptes de guardar la data de naixement es guarda l'edat del client quan es dóna d'alta. Si no hi ha un procediment de posada al dia de les edats, el que passa és que l'assignació d'un client a una edat no queda modificada. Per exemple, tots els clients que l'any 1994 tenien cinquanta-nou anys, l'any 2000 en tenen seixanta-cinc. Han passat de la categoria de clients "veterans" a "jubilats", però en la base de dades es continuen considerant clients "veterans".

f) Variacions en les referències als mateixos conceptes: per exemple, un advocat pot estar considerat com a "professional liberal", mentre que un altre client que també ho sigui pot estar categoritzat com a "autònom". És més probable si la mateixa informació es guarda en bases de dades diferents.

Exemple de variacions en les referències als mateixos conceptes

Aquí tenim un exemple senzill de la situació que considerem. És una simplificació d'un cas real de l'àmbit bancari en el qual el banc també ofereix als seus clients assegurances per a automòbils. En aquest cas, la divisió d'assegurances procedia de l'adquisició per part del banc d'una companyia d'assegurances. La fusió a corre-cuita dels sistemes d'informació de les dues empreses va portar, entre altres conseqüències, que durant un llarg temps les dades dels clients comuns es guardessin duplicadament en dues bases de dades diferents en què, a més, alguns atributs (com el de *Professió*) prenien valors de conjunts diferents. Aquí ho teniu. Fixeu-vos en els clients "Martínez" i "Sala": sapigueu que tots dos són professors universitaris. Però per al banc, el Sr. Martínez és professor i per a l'asseguradora el Sr. Sala és mestre. Com que el banc també tenia la categoria professional *Mestre* entre les que podien ser assignades als seus clients, doncs va passar el que va passar...

Identificador	Nom	Professió	Risc
24.567	Poch	Advocat	Tot Risc
32.456	**Martínez**	**Professor**	**Tercers**
33.345	Martí	Construcció	Tot Risc
33.400	**Sala**	**Mestre**	**Tot Risc**
33.441	Arregui	Cuiner	Tercers

Aquí tenim les dades corresponents als tipus de préstec demanat per cada client i la quantitat corresponent. Què li passa al Sr. Martínez? I al Sr. Sala? Quina és la seva veritable professió? Com ho detectem?

Ident.	Nom	Professió	Préstec	Import	Saldo actual	Saldo mitjà
24.567	Poch	Advocat	Personal	10.000.000	4.500.000	3.200.000
32.456	**Martínez**	**Professor**	**Hipoteca**	**25.000.000**	**1.000.000**	**1.567.000**
33.345	Martí	Construcció	Personal	3.000.000	4.000.000	6.563.316
33.400	**Sala**	**Mestre**	**Hipoteca**	**6.000.000**	**2.000.000**	**5.012.233**
33.441	Arregui	Cuiner	Personal	5.500.000	40.000.000	3.245.678

Els *data warehousing* intenten aportar solucions a aquesta mena de problemes.

g) Dades esbiaixades: aquesta mena de problema es pot donar amb dades que compleixen tots els altres requisits de qualitat esmentats fins al moment. Es tracta de la mena de dades que, en conjunt, reflecteixen preferentment un valor determinat o conjunt de valors o que procedeixen d'un conjunt d'objectes molt determinat. A vegades els estudis de mineria de dades van precisament en la direcció de trobar aquesta mena de subconjunts. Altres vegades no interessa disposar d'aquest tipus de conjunts de dades.

Exemple de dades esbiaixades

Per exemple, pot ser que hàgim escollit sense adonar-nos-en un conjunt de clients que són majoritàriament joves o d'un determinat tipus de professió. Segons quin sigui l'objectiu del nostre estudi, pot no interessar aquesta mena de biaix.

Suposant que després de fer la "neteja" hàgim aconseguit deixar les dades en un estat de qualitat acceptable, encara hi ha més coses a fer.

2.3.2. Transformació de dades

No sempre les dades estan en la forma més adequada per a poder aplicar els mètodes que calen per a la tasca que s'ha de portar a terme i el model que es vol obtenir. En general ens trobarem que haurem d'efectuar alguna d'aquestes transformacions:

1) Dades numèriques a categòriques: les dades categòriques són atributs que prenen valor en un conjunt finit d'etiquetes simbòliques. Per exemple, l'atribut *Edat*, per a una determinada tasca pot ser descrit prou bé com a *Vell* o *Jove* perquè són els grups d'edat que interessa distingir i estudiar. Ara bé, pot succeir que en la base de dades o en diverses bases de dades utilitzades aquest camp tingui un valor numèric (edat entre els valors numèrics 18 a 100, per exemple). La solució consisteix a assignar una categoria a cada rang de valors que ens calgui, bé fixant una correspondència entre els valors numèrics i la categoria (per exemple, *Vell* pot correspondre als valors més grans que seixanta) o bé automàticament mitjançant procediments de discretització.

2) Dades categòriques a numèriques: disposem de dades que apareixen descrites mitjançant valors categòrics i el que ens cal realment és disposar dels valors numèrics corresponents. Hem d'efectuar el procés invers, adscrivint una traducció a cada categoria al conjunt corresponent de valors numèrics. Per

exemple, fent que la categoria *Jove* de l'atribut *Edat* equivalgui al rang de valors divuit-vint-i-cinc anys. El problema és que per cada aparició en la base de dades, potser no podem posar un interval sinó un valor únic. En aquest cas cal efectuar noves transformacions.

3) **Altres transformacions**: moltes vegades, per a simplificar la representació cal efectuar altra mena de transformacions.

- **Simplificació de valors**: per exemple, dividir els sous per 1.000 o 1.000.000.

- **Agrupació de valors continus**: per exemple, totes les compres entre les 8 i les 10 corresponen al valor 1, les de 10 a 12 al valor 2, etc.

- **Normalització de dades**: posar els valors numèrics en un interval determinat. Per exemple, moltes xarxes neuronals obliguen que els valors numèrics estiguin entre 0,0 i 1,0.

- **Afegir-hi una etiqueta** que digui a quina classe pertany un registre. Per exemple, en el cas de les assegurances, si un client pertany a la classe de risc o no. Això es pot haver derivat de l'experiència o bé a partir d'un altre mètode de mineria de dades.

- **Expansió d'un atribut**: pel fet que el valor d'un atribut pot prendre valors en un conjunt limitat de categories. Per exemple, l'atribut *Risc d'incendi* pot prendre valors en les categories *Alt*, *Baix* i *Mitjà*. Pel fet que calgui expressar les dades en forma numèrica podem disgregar l'atribut de *Risc d'incendi* en els atributs *Risc-alt*, *Risc-mitjà* i *Risc-baix,* cadascun dels quals pot prendre el valor 0 o 1, que indiquen l'existència o no de cada tipus de risc.

No totes aquestes transformacions –en absència d'altres eines– es poden fer utilitzant el llenguatge de consulta i manipulació de bases de dades que es tingui a disposició. Com més evolucionades són les eines de mineria de dades, més facilitats donen en aquest sentit i més transparent resulta a l'usuari la interacció de l'eina amb el sistema de bases de dades subjacent. Els *data warehouses* es caracteritzen per donar encara més facilitats en aquest sentit.

Exemples de transformacions de dades

Aquí tenim un exemple en què es poden veure diverses transformacions de dades. De la taula original:

Vegeu l'"Exemple de variacions en les referències als mateixos conceptes" en aquest mateix subapartat.

Ident.	Nom	Professió	Préstec	Import	Saldo actual	Saldo mitjà
24.567	Poch	Advocat	Personal	10.000.000	4.500.000	3.200.000
32.456	Martínez	Professor	Hipoteca	25.000.000	1.000.000	1.567.000
33.345	Martí	Construcció	Personal	3.000.000	4.000.000	6.563.316
33.400	Sala	Mestre	Hipoteca	6.000.000	2.000.000	5.012.233
33.441	Arregui	Cuiner	Personal	5.500.000	40.000.000	3.245.678

Expandint atributs i aplicant transformacions numèriques obtenim la taula que veiem a continuació:

Ident.	Nom	Professió	Personal	Hipoteca	Import	Saldo actual	Saldo mitjà
24.567	Poch	Advocat	1	0	10	4,5	3,2
32.456	Martínez	Professor	0	1	25	1	1,6
33.345	Martí	Construcció	1	0	3	4	6,6
33.400	Sala	Mestre	0	1	6	2	5,0
33.441	Arregui	Cuiner	1	0	5,5	40	3,2

Altres conjunts de canvis i transformacions s'originen per motius diferents. En efecte, la nostra font o fonts de dades poden reunir informació sobre un cert conjunt d'atributs i pot ser que el que ens calgui sigui un conjunt diferent. Comentem els problemes i solucions més habituals.

a) Derivació de dades: podem utilitzar els atributs de les dades existents per a derivar atributs nous (i generar, de fet, un conjunt nou de dades) que ens siguin més útils per a la mena d'estudi de *mineria de dades* que es porti a terme.

Exemple de derivació de dades

Típicament es pot derivar l'atribut *Edat* de la diferència entre la data actual i la data de naixement declarada.

Per a un estudi mèdic sobre l'obesitat pot passar que disposem de les dades següents: edat, alçada (en m), pes (en kg), sexe i professió:

Identificador	Alçada	Pes	Sexe	Professió
24.567	1,90	88	Dona	Advocat
32.456	1,85	92	Home	Mestre
33.345	1,78	73	Home	Construcció
33.400	1,70	65	Dona	Representant
33.441	1,78	110	Home	Cuiner

Però ens interessa obtenir l'índex de massa corporal (IMC), que correspon a aquesta senzilla fórmula:

$$\text{Pes (en kg)} / \text{Alçada (en cm)} \times \text{Alçada (en cm)}$$

Identificador	Alçada	Pes	Sexe	Professió	IMC
24.567	190	88	Dona	Advocat	24,4
32.456	185	92	Home	Mestre	26,9
33.345	178	73	Home	Construcció	23,0
33.400	170	65	Dona	Representant	22,5
33.441	178	110	Home	Cuiner	34,7

De fet, en fer aquest càlcul per a tot el conjunt de dades el que aconseguim és reduir el conjunt d'atributs original, ja que l'IMC és equivalent a la informació combinada del pes i l'alçada. Normalment, però, les derivacions acostumen a ser més complexes, a involucrar-hi més d'un atribut i a generar també més d'un de resultat. Per cert, hi ha hagut alguna altra transformació pel mig que no hem citat.

Com es pot entendre, des del punt de vista de bases de dades, la derivació de dades comporta crear una relació nova, que és la resultant d'incloure un atribut nou a la relació original.

b) Fusió de dades o enriquiment: pot interessar afegir dades procedents d'altres relacions o, fins i tot, altres bases de dades aportades des d'altres fonts.

Exemple d'enriquiment

Podem afegir a la informació sobre els clients el resultat d'una enquesta en què els haguéssim preguntat quin cotxe voldrien comprar i si pensaven canviar de cotxe l'any vinent.

Vegeu "Exemple de transformacions de dades" en aquest mateix subapartat.

Aquí tenim la taula de clients original:

Ident.	Nom	Professió	Préstec	Import	Saldo actual	Saldo mitjà
24.567	Poch	Advocat	Personal	10.000.000	4.500.000	3.200.000
32.456	Martínez	Mestre	Hipoteca	25.000.000	1.000.000	1.567.000
33.345	Martí	Construcció	Personal	3.000.000	4.000.000	6.563.316
33.400	Sala	Representant	Hipoteca	6.000.000	2.000.000	5.012.233
33.441	Arregui	Cuiner	Personal	5.500.000	40.000.000	3.245.678

Aquí, el que van contestar a una enquesta telefònica:

Identificador	Tipus de cotxe	Any vinent
24.567	BMW	Sí
32.456	Skoda	No
33.345	Nissan	Sí
33.400	Mercedes	No
33.441	Smart	Sí

Fusionant les dues taules en una taula nova, cosa que en una base de dades relacional es pot fer amb una operació de "Join", obtenim:

Ident.	Nom	Professió	Préstec	Import	Saldo actual	Saldo mitjà	Tipus de cotxe	Any vinent
24.567	Poch	Advocat	Personal	10.000.000	4.500.000	3.200.000	BMW	Sí
32.456	Martínez	Mestre	Hipoteca	25.000.000	1.000.000	1.567.000	Skoda	No
33.345	Martí	Construcció	Personal	3.000.000	4.000.000	6.563.316	Nissan	Sí
33.400	Sala	Representant	Hipoteca	6.000.000	2.000.000	5.012.233	Mercedes	No
33.441	Arregui	Cuiner	Personal	5.500.000	40.000.000	3.245.678	Smart	Sí

Qui ens demanarà un crèdit l'any vinent?

2.3.3. Reducció de la dimensionalitat

Què vol dir gran?

Una de les justificacions més freqüents per a l'ús d'eines de mineria de dades és la seva capacitat de treballar amb grans conjunts de dades. Ara bé, la grandària d'un conjunt de dades, o d'un problema de mineria de dades, la dóna

tant la quantitat de registres que té com el nombre d'atributs que es maneguen. El que passa és que a partir de certs nivells de registres i atributs, l'eficiència dels algorismes de mineria de dades es comença a reduir. Per tant, si és possible treballar amb menys dades i obtenir els mateixos resultats, doncs millor.

Els **mètodes de reducció de dimensionalitat** cerquen justament treballar amb menys dades i obtenir els mateixos resultats.

A continuació presentem les tècniques habituals de reducció de la dimensionalitat.

Reducció del nombre de registres que cal tractar

La reducció del nombre de registres que cal tractar consisteix a trobar un conjunt de dades més petit per a construir el tipus de model que necessitem amb el nivell de qualitat necessari.

Com a exemple...

... de reducció del nombre de registres que cal tractar, suposem que per a predir algun comportament determinat dels nostres clients potser no cal tractar tota la base de dades de clients, sinó una mostra significativa més reduïda.

L'estadística ha desenvolupat eines per escollir conjunts suficients de dades per a construir models. Per tant, és bo recórrer a les seves tècniques per a obtenir un conjunt més reduït, però igualment potent, de dades inicials. Els problemes que hem d'evitar aquí són principalment que el conjunt escollit no sigui massa esbiaixat –un conjunt d'objectes amb característiques molt concretes i poc representatives–, com ja hem comentat abans. També succeeix que si el conjunt és massa reduït, les conclusions que s'extreguin del model resultant final no seran prou significatives. Per tant, a vegades cal conservar un conjunt alt de registres per a mantenir la qualitat suficient.

Vegeu les dades esbiaixades en el subapartat 2.3.1 d'aquest mòdul.

Una diferència que cal destacar de la manera en què s'han de seleccionar les mostres de dades en estadística i en mineria de dades és que, mentre que en la primera els casos extrems (*outliers*) es descarten sistemàticament, en la segona són els que més interessen. Per tant, els mètodes per a reduir el nombre de registres que s'emprarien en l'anàlisi de dades tradicional i en mineria de dades són lleugerament diferents.

Reducció del nombre d'atributs que cal tractar

La reducció del nombre d'atributs per tractar també s'anomena *selecció d'atributs*.

La **selecció d'atributs** consisteix a eliminar els atributs que no aportin informació per a obtenir el tipus de model que volem.

Exemple de selecció d'atributs

Posarem un exemple de reducció d'atributs en el qual hem treballat els autors directament: en certes depuradores d'aigües residuals urbanes es recollien dades de cent quaranta variables i es va poder demostrar que només triant-ne tretze es podien obtenir models igualment útils per a la tasca assignada.

Això representa haver de detectar atributs irrellevants que no tenen efecte sobre la qualitat del model final (és a dir, que el model ens permet de contestar les mateixes preguntes amb aquests atributs o sense), i atributs o combinacions d'atributs equivalents (atributs que permeten de fer el paper de grups d'altres atributs sense afectar la qualitat final del model).

Els mètodes de selecció d'atributs tenen una certa complexitat i els estudiarem adequadament amb un parell d'exemples.

Nota

És convenient que mireu els dos exemples presentats aquí.

Exemples de mètodes de selecció d'atributs

Exemple 1

Vegem un exemple senzill. Suposem que tenim les dades d'obesitat d'una sèrie de clients del banc:

Identificador	Alçada	Pes	Sexe	Professió	IMC
24.567	190	88	Dona	Advocat	24,4
32.456	185	92	Home	Mestre	26,9
33.345	178	73	Home	Construcció	23,0
33.400	170	65	Dona	Representant	22,5
33.441	178	110	Home	Cuiner	34,7

Sembla raonable pensar que la combinació d'atributs (*Alçada*, *Pes*) és equivalent a l'IMC, ja que aquest darrer atribut es calcula a partir dels altres dos. Podem dir que la combinació d'atributs (*Alçada*, *Pes*) aporta la mateixa informació que l'atribut IMC. Per tant, podem substituir els dos atributs *Alçada* i *Pes* per l'atribut IMC amb una senzilla operació SELECT de les bases de dades relacionals:

Identificador	Sexe	Professió	IMC
24.567	Dona	Advocat	24,4
32.456	Home	Mestre	26,9
33.345	Home	Construcció	23,0
33.400	Dona	Representant	22,5
33.441	Home	Cuiner	34,7

Amb un exemple tan senzill com aquest és evident que no estalviem gaire, però en bases de dades de milions de clients és important fer una anàlisi prèvia per a trobar aquesta mena d'equivalències.

Exemple 2

Posem un altre exemple no tan directe i que té a veure amb els atributs irrellevants. Tornem a la nostra base de dades del banc imaginari. Hi hem afegit un atribut nou que correspon a la *Classe Client*. Cada client no pot pertànyer sinó a una sola classe. La classe 0 és la dels clients de poca morositat. La classe 1 és la d'alt risc de morositat. No cal que ens preocupem ara de la manera en què s'ha obtingut aquesta classificació (precisament aprendrem a classificar en el mòdul "Arbres de decisió" d'aquesta assignatura). A més, hem introduït alguns canvis en els valors per deixar més clar el que volem dir:

Ident.	Nom	Professió	Préstec	Import	Saldo actual	Saldo mitjà	Tipus de cotxe	Any vinent	Classe
24.567	Poch	Advocat	Personal	10	4,5	3,2	BMW	Sí	0
32.456	Martínez	Mestre	Hipoteca	25	1	1,6	Skoda	No	1

Ident.	Nom	Professió	Préstec	Import	Saldo actual	Saldo mitjà	Tipus de cotxe	Any vinent	Classe
33.345	Martí	Construcció	Personal	3	4	6,6	Nissan	Sí	0
33.400	Sala	Representant	Hipoteca	6	2	5,0	Mercedes	No	1
33.441	Arregui	Cuiner	Personal	5,5	40	3,2	Smart	Sí	0

Aquí hi ha dos atributs que, tot i semblar interessants, ens presenten un problema. Tothom que té un préstec personal ha contestat *sí* a la pregunta de si es comprarà un cotxe l'any vinent. Tothom que té una hipoteca ha contestat *no*. Per tant, aquests dos atributs són molt poc informatius. Evidentment si hi hagués més tipus de préstecs, això no seria així. Però llavors el que passaria és que tindríem unes dades insuficients. És clar que això ho hauríem de matisar. Si volem establir associacions o dependències sembla que hi ha una forta dependència entre el tipus de préstec i la intenció de compra, la qual cosa ja és prou significativa. Si volem agrupar els clients per la seva similitud, potser aquests dos atributs no ens calen per a res. Ja parlarem amb més profunditat d'aquesta mena de problemes en el mòdul "Preparació de dades" d'aquesta assignatura.

Activitat

1.1. Considereu el segon exemple plantejat a "Exemples de mètodes de selecció d'atributs" i plantegeu-vos la pregunta següent: podem pensar seriosament que els camps *Ident.* i *Nom* són fonamentals per a agrupar clients semblants en un camp d'aplicació com la banca?

2.4. Mineria de dades: el procés de construcció de models

En la fase de mineria de dades tenim les dades amb la qualitat i en el format adequats i hem seleccionat els atributs i els registres aparentment necessaris i rellevants.

Tenim decidit quina mena de model es vol obtenir. Per tant, ara cal escollir un mètode de construcció de models entre la munió de mètodes que permeten d'obtenir el tipus de model que ens interessa.

De la mateixa manera que la diversitat aparent de models ens pot ocultar les similituds envers la tasca per a la qual s'ha d'aplicar, tampoc no podem deixar de remarcar la gran similitud existent en el procés de construcció dels diversos tipus de models.

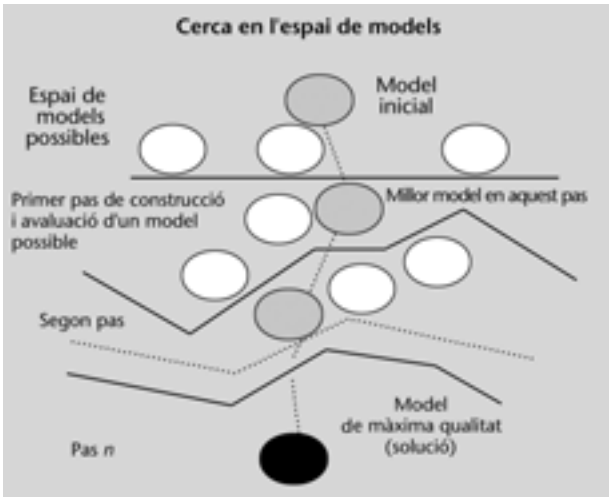
Diversitat de models

Per a una mateixa tasca hi ha diversos models que ens poden ser útils.

En efecte, el procés de construcció de models consisteix a trobar el model (el coneixement) que respon millor a les característiques implícites dins les dades. Aquest tipus de problema s'acostuma a conceptualitzar com un *procés de cerca*.

Un **procés de cerca** consisteix a explorar un espai de models possibles (per exemple, de trobar la millor xarxa neuronal entre totes les possibles, donades les dades d'entrada i de sortida) per a trobar el que tingui la millor qualitat.

Podem representar el procés de cerca amb l'esquema següent:



2.4.1. Mecànica general del procés de cerca

En què consisteix un procés de cerca?

Tot **procés de cerca** parteix d'un model inicial (que pot ser el model buit, inexistent) i a cada pas el modifica mitjançant un conjunt d'operadors de modificació de models. A cada pas és possible aplicar més d'un operador i es poden generar diversos models nous alternatius:

- Si algun d'aquests models ja té prou qualitat, es pot considerar el model final i hem trobat la solució.

- Si el model no és encara la solució, n'hem d'escollir algun dels altres i continuar aplicant operadors de transformació fins que trobem una solució o fins que no hi hagi més combinacions possibles per obtenir.

Quins models parcials escollirem?

Com que la majoria de les vegades el conjunt de combinacions possibles és tan gran que resulta impossible generar-les totes en un temps raonable, cal escollir encertadament a cada pas un subconjunt de models parcials.

Escollirem models que semblin assegurar un model final.

Com coneixem el millor model?

Per a saber quin és el millor model utilitzarem alguna mena de funció d'avaluació, que dóna un valor a un model parcial, més alt com més probable sigui que es trobi en el camí cap a un model final bo, de qualitat alta.

No se'ns ha d'ocultar que els principals problemes d'aquesta mena de conceptualització són els següents:

a) Disposar d'una mesura de qualitat que qualifica com el millor possible un model que tan sols és relativament bo respecte als models que té a prop seu,

però que no és el model òptim dins de tot l'espai de solucions. Aquest problema s'anomena **problema de l'obtenció de mínims locals**.

b) Obtenir un model que sigui bo tenint en compte només les dades de què disposem. Per posar un exemple, si modelitzem el comportament dels clients d'una empresa i disposem d'una base de dades de tres mil clients per a fer proves, de la qual obtenim un model classificatori, no volem que en veure el cas tres mil u, el model s'equivoqui i li assigni la classe que no li correspon (per exemple, no volem que un client amb risc de morositat sigui classificat com a no morós). Aquest problema d'obtenir models que generalitzen malament i són massa específics al que han vist se'l coneix com el **problema de la sobreespecialització** (en anglès, *overfitting*).

c) En el decurs del temps és possible que el model obtingut degeneri, perquè comencen a aparèixer observacions noves per a les quals no generalitza bé. Aquest és el **problema de l'envelliment de models** i cal actualitzar-los.

Veurem que tots els **processos de construcció de models** es distingeixen per les característiques següents:

- El **llenguatge de descripció**: és a dir, allò que expressen els models (associacions, dependències, similituds, etc.).

- Els **operadors de modificació de models parcials**: com es van construint els models.

- La **funció d'avaluació**: avaluen que el model que es construeix és millor o pitjor.

2.4.2. Diversitat dels models de cerca

Vegeu la comprensibilitat del model en l'apartat 1 d'aquest mòdul.

Pel que acabem de dir, sembla com si tots els mètodes de construcció de models funcionessin de la mateixa manera: parteixen de les dades i avancen per l'espai de solucions guiant-se per la seva funció d'avaluació. Això és cert fins a un cert punt i és recomanable en part. De fet, si recordem que el model ha de ser comprensible i utilitzable per algú que pren decisions, convé veure si el procés admet també la presència d'un observador humà que en tot moment pugui avaluar com ha anat funcionant el procés i fer que la cerca prengui una certa direcció.

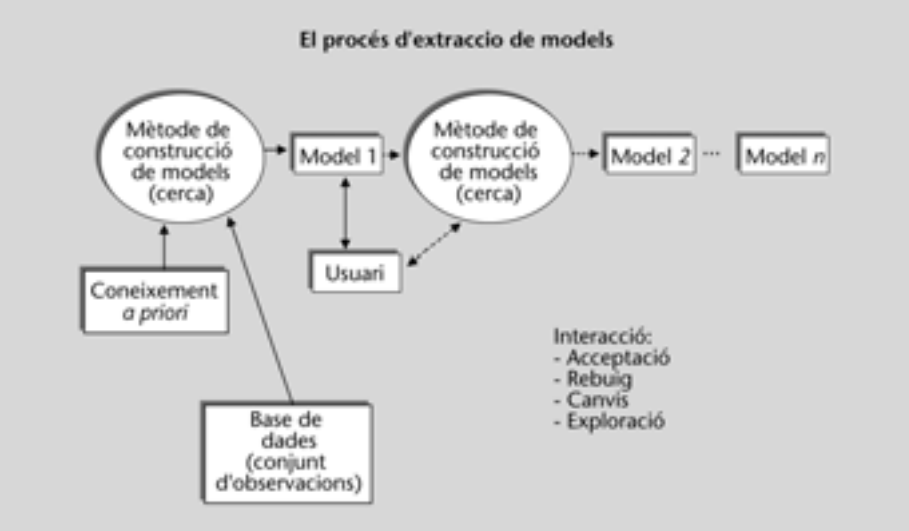
En el fons, es pot considerar que un procés de cerca parteix de més informació que les dades. Això ho esquematitzem en aquesta figura:

Els elements que cal considerar en aquest esquema són els següents:

a) Base de dades: conjunt d'observacions o exemples de què disposem.

b) Coneixement *a priori*: tot tipus de coneixement de què es disposa i es vol utilitzar. Per exemple, saber que determinades pautes de conducta sempre estan associades. En el cas del banc de què hem parlat, podem expressar el coneixement que els clients de renda alta viuen majoritàriament en un barri determinat i, per tant, indiquem que tenim preferència per models que reflecteixin aquest fet. Aquí hi ha un aspecte important: si les dades contradiuen aquest coneixement *a priori*, a qui hem de fer cas? A les dades o a l'usuari humà que ha donat aquesta indicació inicial? Mig i mig? Com alterem el model que es construeix per tal d'equilibrar la influència de les dues fonts d'informació, l'usuari i les dades?

Vegeu l'exemple del banc a "Exemple de variacions en les referències als mateixos conceptes" en el subapartat 2.3.1 d'aquest mòdul.



De quina manera podem expressar aquest coneixement *a priori*? Aquí exposem algunes de les tècniques d'expressió d'aquest coneixement més típiques, utilitzades en alguns dels mètodes, però no sempre aplicables a tots:

- **Distribució *a priori* de la probabilitat dels valors de les dades**: normalment és algun paràmetre que permet de fixar les característiques de la distribució que segueixen les dades. Per exemple, podem suposar que la distribució que segueixen les dades és una normal multivariant i donar els paràmetres que creiem importants.

- **Relacions entre les dades**: per exemple, indicar que considerem que dos atributs s'han de tractar com a dependents. En l'exemple que hem comentat més amunt, els atributs serien el nivell de renda i el barri de residència.

- **Relacions entre les dades i part del model**: per exemple, en el cas d'agregació, forçar que determinades observacions estiguin en el mateix grup d'observacions final.

- **Relacions de coocurrència**: semblants a les de relacions de dependència que hem comentant més amunt. Per exemple, "sempre que algú compra cervesa, també compra cacauets".

c) Control del procés de cerca: en principi, el procés de cerca es guia per les característiques de qualitat dels models que va construint o bé amb una mesura de qualitat que és una combinació del grau amb què el model s'ajusta a les dades i a les preferències de l'usuari. Això no vol dir que el procés de cerca procedeixi automàticament fins a trobar un model de qualitat suficient. L'usuari pot decidir aturar el procés de cerca i recular fins a una situació anterior en què comença a explorar i a introduir una altra informació. Aquest és el sistema seguit, per exemple, per Data Surveyor.

Lectura complementària

Trobareu informació sobre Data Surveyor en l'obra següent:

M. Holsheimer; A. Siebes (1994, gener). "Data mining: the Search for Knowledge in Databases". *MReport Tecnic CS-R9406*. Amsterdam: Centrum voor Winskunde & Informatica, CWI.

Segons aquest esquema, doncs, podem dividir els **mètodes de mineria de dades** pel que fa al coneixement *a priori*, al tipus de dades i al procés de construcció, en els termes següents:

1) Pel que fa al **coneixement *a priori***

- Tipus de coneixement *a priori* que els mètodes permeten d'expressar.
- Llenguatge en què els mètodes permeten d'expressar aquest coneixement.
- Reutilització del coneixement extret per un altre sistema de mineria de dades per a guiar el procés de cerca*.

*** En anglès, *multistrategy learning*.**

2) Pel que fa al **tipus de dades**

- Mètodes que utilitzen només observacions; és a dir, només informació respecte als valors que s'han recollit, sense més informació afegida. Els mètodes que parteixen d'aquesta base se'ls anomena **mètodes no supervisats**. Un exemple típic són els mètodes d'agregació*.

*** En anglès, *clustering*.**

- Mètodes que afegeixen informació discriminant a les observacions. Per exemple, a cada observació se li assigna el nom o l'indicatiu de la classe a la qual pertany. Pròpiament parlem llavors d'**exemples** i **contraexemples**, més que no pas d'observacions. Els mètodes que parteixen d'aquesta situació inicial s'anomenen **mètodes supervisats**. Un exemple típic n'és la classificació.

3) Pel que fa al **procés de construcció**

- **Mètodes batch***:* són mètodes que estan pensats per a utilitzar totes les dades existents com a un únic conjunt de dades per a obtenir un model "en un sol pas".
- **Mètodes incrementals**: són els mètodes que estan pensats per a anar construint un model amb porcions del conjunt total d'observacions (fins i tot d'observació en observació) sense guardar memòria de totes les observacions vistes anteriorment. Són útils per a anar modificant els models a mesura que van canviant les dades en el temps (envelliment de dades o deriva de models).

Lectura complementària

Trobareu informació sobre els mètodes incrementals en l'obra següent:

T.M. Mitchell (1997). *Machine Learning*. Nova York: McGraw-Hill.

- **Mètodes interactius**: mètodes en què l'usuari té un paper a cada pas o sèrie de passos que el procediment efectua, introduint-hi coneixement nou i donant indicacions sobre cap on cal portar la cerca abans de construir el model final.

2.5. Avaluació i interpretació del model

El final de la fase de mineria de dades és un model que representa un tipus determinat de coneixement sobre el domini que estudiàvem. Però com és de bo, aquest model? Podria ser millor? El podem considerar prou bo, o hem de tornar a començar?

No hi ha mesures absolutes i generals de qualitat per a tot tipus de models, ja que hem dit que cada model està dirigit a un objectiu diferent.

Models amb objectius diferents

No es poden comparar models predictius amb models descriptius. Per als primers hem d'usar mesures que ens diguin fins a quin punt són bons fent prediccions. Per als segons, hem de mesurar fins a quin punt s'ajusten al domini descrit.

Típicament, el **procés d'avaluació** consisteix a disposar de dos conjunts de dades procedents del mateix conjunt inicial (i ja preparat): un conjunt de dades que s'utilitza per a construir el model i un altre per a avaluar-lo. A vegades s'introdueix un tercer conjunt entre aquests dos: el **conjunt de validació**. Utilitzem un conjunt de dades per a construir el model; un altre per a donar-lo per bo (validar-lo) i un tercer per a avaluar-lo. Normalment aquests tres conjunts, tot i reflectir informacions sobre els mateixos atributs, procedeixen de conjunts de dades diferents.

Però quines són les mesures utilitzades més sovint per a avaluar la qualitat del model? I com es pot efectuar l'avaluació d'una manera més metòdica que la que tot just hem apuntat?

Exemple de procés d'avaluació

Agafem un cas del món de les assegurances. Si volem extreure un model predictiu que ens digui a partir d'una sèrie de dades dels clients si un client nou pot ser de risc (tenir un nombre excessiu d'accidents que l'asseguradora hagi de pagar), separarem la base de dades de clients, dels quals sabem quins són de risc i quins no, en dues bases de dades: una per a construir el model i una altra per a validar-lo.

Sobre la primera sèrie de dades apliquem un mètode de predicció (per exemple, una classificació que ens digui si un client és de la classe de risc o de la de no-risc) i n'extraiem les combinacions de valors que prediuen bé la pertinença a una classe o a una altra.

Per exemple, suposem que els clients amb edat baixa (*Joves*) que tenen cotxes vermells i un carnet de conduir amb més de dos anys d'antiguitat són els més propensos a pertànyer a la classe de risc. Podem utilitzar aquesta regla de predicció sobre el conjunt de dades d'avaluació per saber si aquest "minimodel" és correcte:

- Si amb aquesta regla es classifica correctament un percentatge de clients significatiu (posem el 95%), tenim un model de bona qualitat. És a dir, si el model classifica correctament clients que el mètode no havia utilitzat per a construir-lo. Això vol dir: si indica com a propensos a ser de risc clients que tenim etiquetats com de risc, i com de no-risc, clients que tenim etiquetats com a tals.

- Si, en canvi, la proporció de "falsos positius" (clients de no-risc que es classifiquen com de risc) o "falsos negatius" (clients de no-risc que en realitat sí que són de risc) és alta, llavors tenim un model predictiu de baixa qualitat.

En aquest darrer cas hem de pensar que alguna cosa ha anat malament en tot el procés de mineria de dades (la qualitat de les dades, una mostra insuficient o esbiaixada, etc.) i reconsiderar els passos corresponents.

Aquest petit exemple ens ha servit per a mencionar una possible mesura de qualitat: l'"error en la predicció".

Interpretació final

Un cop tenim un model que té el nivell de qualitat requerit i que ha estat validat mitjançant el procés d'avaluació, cal interpretar-lo i extreure'n el significat del coneixement que ens mostra. Aquí es corre el risc de les falses interpretacions.

Exemple de falses interpretacions

Un exemple molt senzill de falses interpretacions és el del sistema de predicció que va determinar que el factor més important per a calcular si una persona resident a Londres podia quedar-se embarassada era el sexe, ja que en el 99,99 per cent dels casos les persones que quedaven embarassades eren dones (seria interessant saber què passava amb el 0,01 per cent restant!).

Exemples de models que donen conclusions evidents n'hi ha molts i cal estar a l'aguait en aquesta fase d'interpretació per a no recollir com una gran troballa allò que ja és sabut. D'aquí ve la importància dels mètodes que permeten d'integrar coneixement *a priori*, especialment coneixement negatiu, del que s'està segur que no pot passar o que no és rellevant. Però hi ha problemes que no són evidents.

El cas de l'apendicectomia beneficiosa

Il·lustrem el problema de les falses interpretacions amb el cas discutit per Wen sobre la interacció entre determinats tipus d'operacions quirúrgiques i la taxa de mortalitat en un hospital públic d'Ontario, Canadà, entre el 1981 i 1990.

Lectura complementària

Trobareu el cas estudiat per Wen de l'apendicetoctomia beneficiosa en l'article següent:

S.W. Wen; R. Hernández; C.D. Naylor (1995). "Pitfalls in Nonrandomized Studies: The case of incidental Appendectomy with Open Cholecystectomy". *Journal of the American Medical Association* (núm. 275, pàg. 1687-1691).

Wen es va concentrar en els casos de pacients sotmesos a una colecistotomia primària oberta. Alguns d'aquests pacients també havien estat sotmesos a una apendicectomia en el procés de colecistotomia, el que s'anomena una *apendicetoctomia incidental* o *discrecional*.

En la taula següent es poden veure els resultats que reflecteixen les morts ocorregudes a l'hospital comparant els pacients que havien patit apendicectomia durant l'operació de colecistotomia primària oberta i els que no:

El cas de l'apendicectomia beneficiosa		
	Amb apendicectomia	Sense apendicectomia
Percentatge de morts ocorregudes a l'hospital	21 (0,27%)	1.394 (0,73%)
Percentatge de pacients supervivents a l'hospital	7.825 (99,73%)	190.205 (99,27%)

Es va efectuar un test de significació de χ^2 per a comparar els resultats dels dos grups i esbrinar si mostraven una diferència significativa. Es va trobar que, segons el test, efectivament la diferència era significativa.

Aquest "descobriment" del coneixement del fet que una apendicectomia incidental durant l'operació de colicistotomia pot "millorar" les probabilitats de sobreviure s'ha de prendre amb una mica de calma. Com és possible que una apendicectomia incidental pugui millorar els resultats?

Wen va considerar separadament un grup de pacients de baix risc. Aquest grup de pacients, en canvi, mostrava que l'efecte de l'apendicectomia discrecional tenia resultats força insatisfactoris. Paradoxalment, podria ser que l'apendicectomia casual afectés negativament tant els pacients de baix risc com els d'alt risc però que, considerats plegats,

donés la impressió d'un efecte positiu. D'això se'n diu *paradoxa de Simpson* i se n'ha d'estar prou a l'aguait. Vegem com acaba de funcionar.

En la taula següent es mostren dades fictícies que permetrien d'interpretar aquesta paradoxa:

Divisió en pacients de baix i alt risc				
	Amb apendicectomia		**Sense apendicectomia**	
	Baix risc	Alt risc	Baix risc	Alt risc
Morts	7	14	100	1.294
Supervivents	7.700	125	164.009	26.196

En la taula següent es mostren les proporcions corresponents a les morts dins l'hospital classificades per apendicectomia incidental i pacients de risc corresponents amb les dades de la taula anterior:

Efecte combinat		
	Amb apendicectomia	**Sense apendicectomia**
Baix risc	0,0009	0,0006
Alt risc	0,1000	0,0500
Combinat	0,0030	0,0070

Es pot veure que les categories de risc i les morts estan altament correlacionades. Era més probable que les apendicectomies fossin aplicades a pacients de baix risc que no pas als d'alt risc. Per tant, si no es coneix la categoria de risc (relacionada amb l'edat) d'un pacient, però se sap que ha passat per una apendicectomia, es pot deduir que és més probable que pertanyi a la categoria de "Baix risc" (*Joves*). Ara bé, això no implica de cap manera que passar per una apendicectomia disminueixi el risc d'alguns pacients. Si la informació sobre el risc no apareix en la taula, se'n pot extreure aquesta conclusió purament il·lusòria.

Wen va fer un estudi de regressió tenint en compte més variables (edat, sexe, situació d'entrada a l'hospital) i va concloure que no hi havia cap manera d'afirmar que la millora a curt termini es pogués considerar deguda a l'apendicectomia.

Per tant, la interpretació requereix molta precaució, i més d'una mirada sobre els resultats.

2.6. Integració dels resultats en el procés

El darrer pas consisteix a integrar els resultats de la mineria de dades en el procés típic del sistema d'informació en què s'aplica.

Un exemple senzill és el de procés de documentació textual utilitzat en alguns grans diaris. Cada dia les notícies es classifiquen per diverses categories, de manera que els usuaris dels serveis d'informació d'aquests diaris poden fer consultes per diverses paraules clau. Clarament, hi ha un esforç previ de classificació. Aplicant-hi algorismes de mineria de dades textuals és possible construir un procediment de classificació que automàticament assigni les pa-

Lectura complementària

Trobareu més informació sobre els processos de documentació textual utilitzats en alguns grans diaris en l'article següent:

J. Schmitz; G.I. Armstrong; J.D.C. Little (1990). "CoverStory-Automated News Finding in Marketing". *DSS Transactions*.

raules clau a les diverses notícies del diari. La integració correspondria aquí a la transformació del model de classificació en un programa més de la cadena: edició, etiquetatge i inclusió a la base de dades documental del diari.

No tots els models es poden integrar amb facilitat. La majoria requereixen una transformació al codi de programació corresponent. Bona part dels sistemes comercials de mineria de dades ofereixen la possibilitat de traduir el model obtingut en procediments en el llenguatge de programació corresponent (normalment, C) i inserir-lo després dins un tractament d'informació més general.

2.7. Observacions finals

Ens afanyem a remarcar que, tot i aquesta presentació lineal, el descobriment avança de manera iterativa. No acaba amb la construcció d'un model i amb la generació d'informació resumida. Un cop es disposa d'un model cal treballar-hi, fer preguntes noves, preguntar-se què passa si en comptes de disposar de les relacions que mostra el model se'n donessin unes altres, etc. Això pot representar a la vegada l'aportació de dades que no ens calien inicialment i, per tant, haver d'emprendre un procés de selecció i neteja de dades noves que donaran, amb les dades actualment existents, un model nou, etc.

Assolir aquesta possibilitat de mantenir un procediment obert de descobriment no és trivial; cal preveure els mecanismes que permetin de redefinir fàcilment les dades d'interès, transformar-les, etc. Cal, doncs, una disposició activa per a anticipar els requeriments de dades de cada àrea d'interès possible, la connexió de les fonts de dades adequades, etc.

3. Les eines de mineria de dades i les àrees relacionades

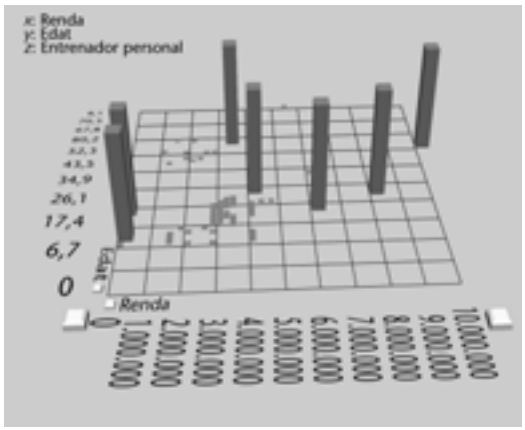
El nucli del programa se centra en les eines de mineria de dades, els models, els mètodes per a construir-los i els algorismes en què es basen. Normalment, però, hi ha tota una sèrie d'eines que també tenen relació amb la mineria de dades i que almenys cal tenir presents.

3.1. Eines de visualització

Una manera molt potent i intuïtiva d'extreure coneixement a partir de dades és per inspecció visual.

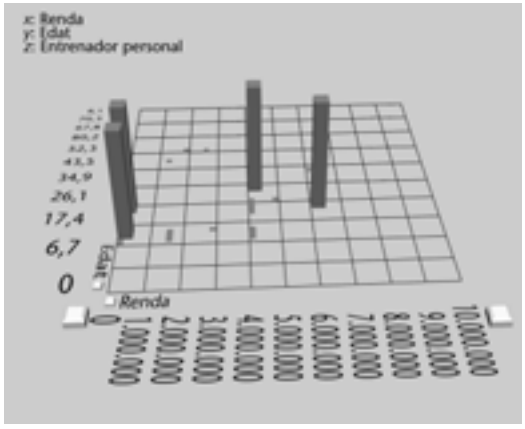
Exemple senzill de la potència de les eines d'inspecció visual

En la figura següent es representa la relació entre el nivell d'ingressos (*Renda*, eix *X*) dels socis del club esportiu que utilitzarem com a exemple al llarg del programa, la seva edat (*Edat*, eix *Y*) i el fet que demanin un entrenador personal (*Entrenador personal*, eix *Z*).



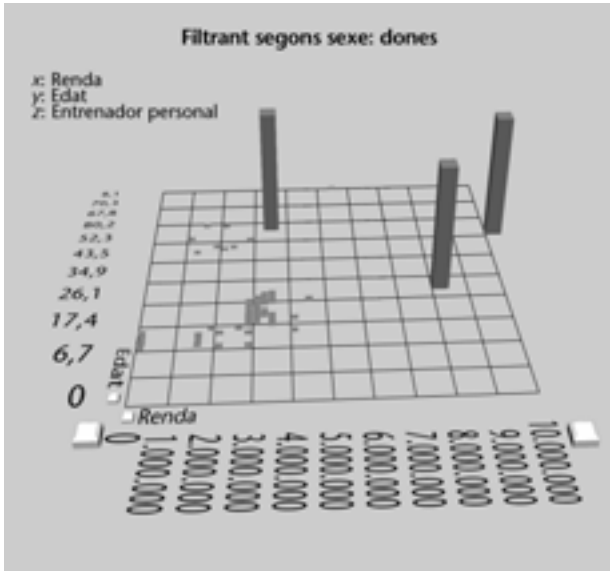
Podem extreure algun coneixement d'aquesta visualització: sembla com si la major part de socis que demanen el servei d'entrenador personal es concentren en les rendes superiors a 4 milions de pessetes. Fem unes quantes comprovacions més:

1) Fem un filtratge de les dades per sexes, de manera que en el gràfic apareguin només els homes:

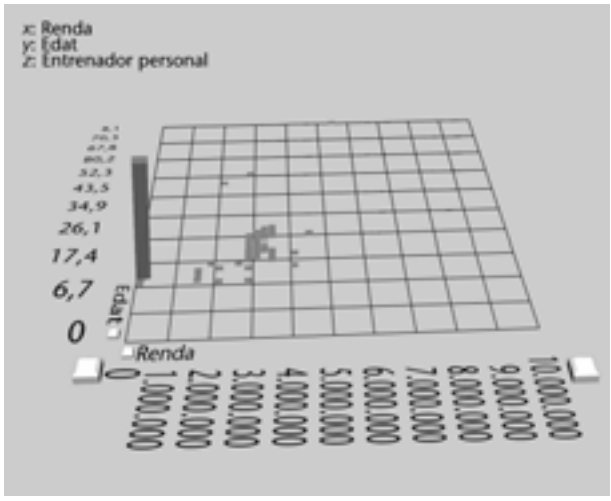


Podem extreure la conclusió que els qui predominantment demanen el servei d'entrenador són homes joves i rendes mitjanes-baixes.

2) I les dones? Fem un filtratge de les dades per veure només les dones:



Sembla, doncs, que es tracta de dones més grans i amb rendes més altes, oi? Si ara filtrem el que tenim segons el districte de residència i només eliminem els socis que procedeixen de la zona A (un barri de classe alta) resulta que gairebé no apareixen socis en la imatge:



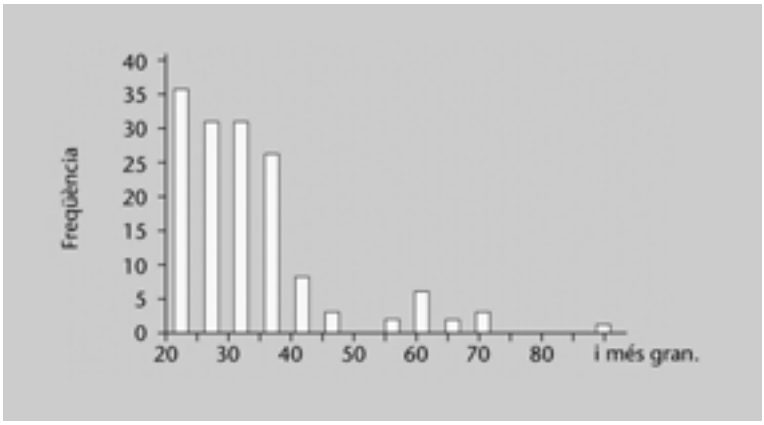
Per tant, sembla que fins ara arribarem a conèixer prou bé els clients que demanen el servei d'entrenador personal en aquest gimnàs: són homes joves de renda baixa-mitjana, però que viuen al districte alt de la ciutat, o bé dones més grans amb renda mitjana-alta que també viuen al mateix barri. Ara, després de tenir una primera idea de les dades, podríem aplicar altres mètodes que ens donessin un resultat numèric més precís i un model de predicció més acurat que el que ens permet la inspecció visual.

Les eines de visualització són força útils en la fase de preparació de dades i la d'interpretació dels models resultants. En la primera fase, aquestes eines permeten d'ajudar-nos a conèixer millor les dades, i es complementen amb les eines típiques d'estadística descriptiva, que ens permeten de trobar els valors més freqüents, la dispersió de valors de cada variable, valors mínims i màxims, valors molt poc freqüents i alguna mena de correlació entre les variables considerades.

Les eines visuals més tradicionals en aquesta part són els **histogrames**.

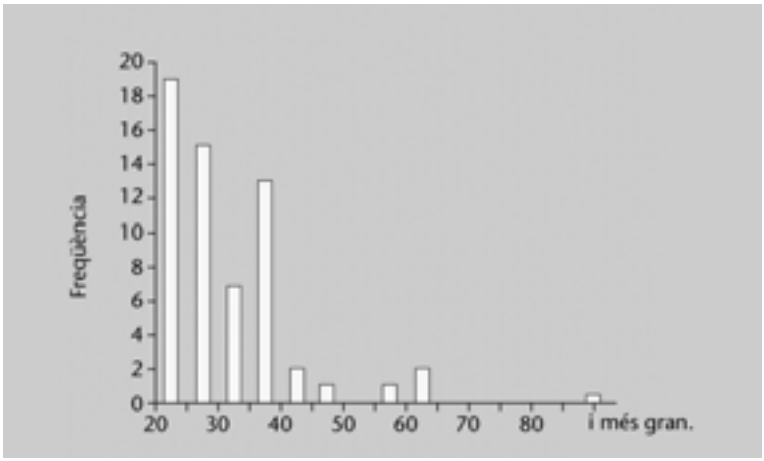
Exemple d'utilitat dels histogrames

Aquí podem veure la distribució dels valors de les edats entre els clients del nostre club:

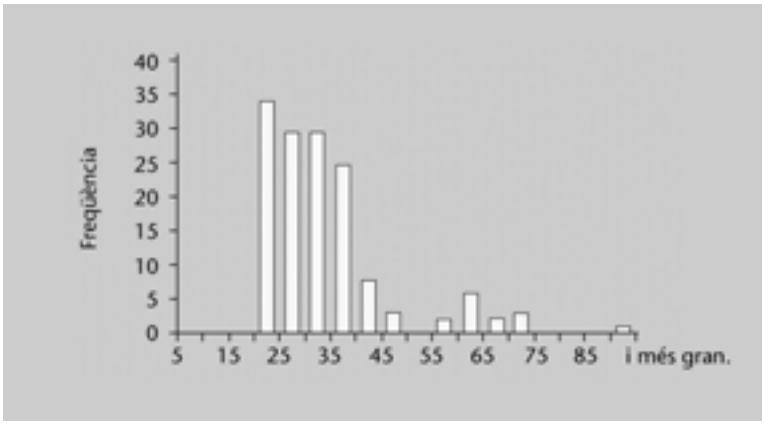


Vegeu l'"Exemple senzill de la potència de les eines d'inspecció visual" en aquest mateix subapartat.

Escollint una variable discriminant (o una variable de classe) es poden fer comparacions entre les distribucions de valors per a diverses classes o combinacions dels valors dels altres atributs. Aquí tenim la distribució de les edats entre les dones que acudeixen al club:



I aquí, entre els homes:



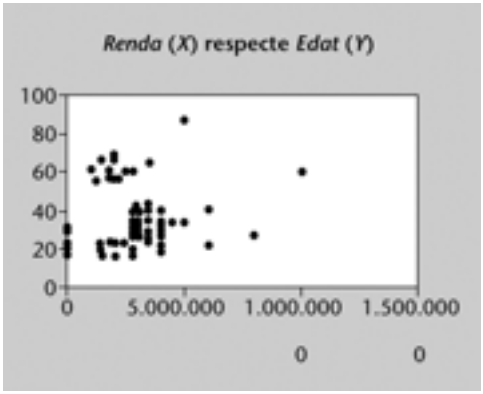
Podem dir que hi han diferències significatives?

Una altra eina útil són els **diagrames de dispersió***, que donen una idea de la relació entre els valors de dues variables.

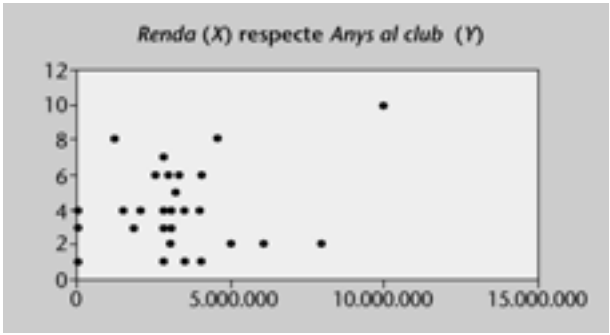
*** En anglès, *scatter plots*.**

Exemple d'utilitat dels diagrames de dispersió

Aquí podem veure la relació entre els valors de les variables *Renda* i *Edat*:



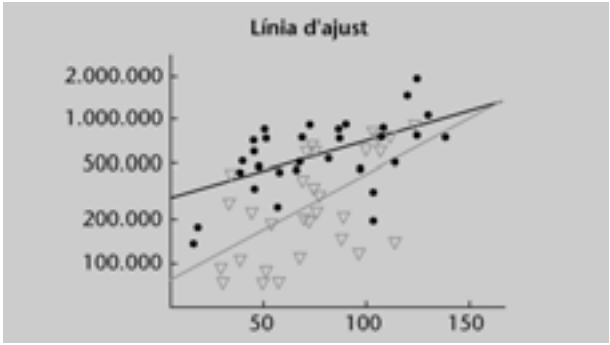
I aquí entre *Renda* i *Anys al club*:



Normalment, l'existència de relació entre els valors de dues variables es pot estudiar construint una funció que dibuixi una línia que expliqui els valors d'una variable en funció dels de l'altra. Igualment es pot derivar el coeficient de correlació entre les dues variables. Ara no entrarem a definir ni explicar aquests conceptes. Només ens centrarem en l'aspecte de visualització.

Exemple d'utilitat d'una funció d'ajust

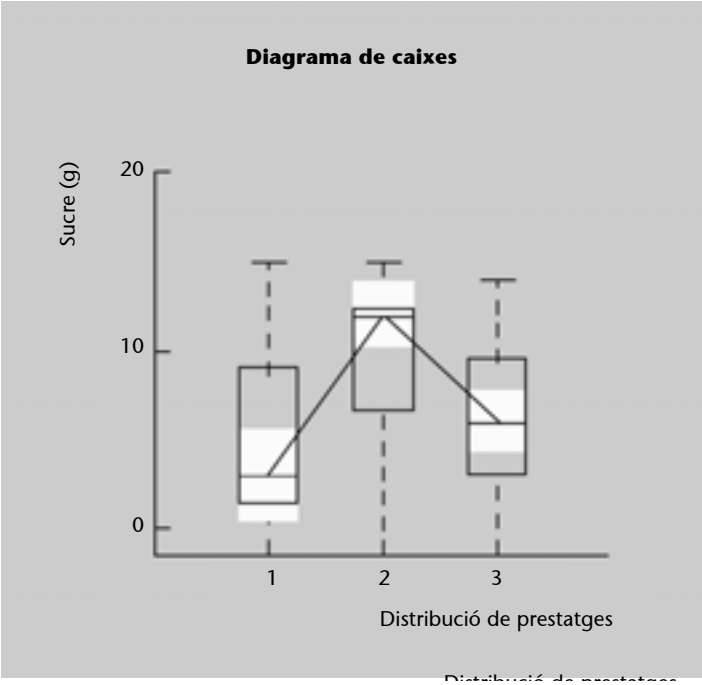
Aquí presentem una gràfica d'exemple en què s'ha trobat una línia d'ajust que indica una relació funcional entre les variables *Ingressos* i *Temps treballat en tres mesos*.



Una altra eina gràfica que informa de la concentració de valors entorn d'un punt són els **diagrames de caixes** o ***boxplot***.

Exemple d'utilitat d'un diagrama de caixes

Aquí tenim una mostra d'un diagrama de caixes que relaciona determinats productes amb la seva disposició als prestatges del supermercat:



El problema de les dades amb dimensionalitat gran (amb un conjunt d'atributs gran) és que no podem visualitzar completament les relacions entre totes les variables de manera simultània. Així, cal projectar conjunts de *n* variables sobre representacions gràfiques de 2 o 3 dimensions i esgotar les diverses combinacions de variables dos a dos per a preguntar-nos sobre els fenòmens d'interès.

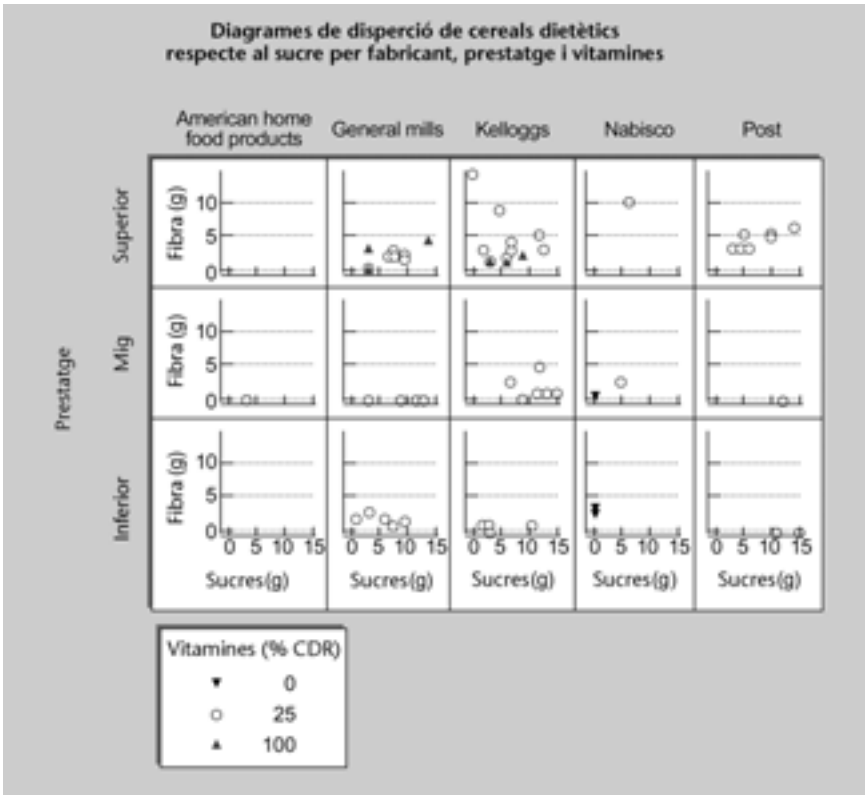
És a dir, si treballem amb observacions que tenen vint atributs, conceptualment treballem en un espai de vint dimensions, que no podem visualitzar de cap manera.

Ara bé, projectant part d'aquests atributs en representacions tridimensionals o bidimensionals podem extreure algun tipus d'intuïció que després podem confirmar o refutar amb una altra mena d'eines, procedents de l'estadística i de l'aprenentatge automàtic, i encetar un autèntic procés de mineria de dades.

El tractament d'aquest problema admet diverses formes. En general, es fan combinacions d'histogrames o gràfics de dispersió per a diverses variables i diverses fonts. Aquestes tècniques són força comunes i útils per a intentar comparar rendiments de centres dispersos geogràficament.

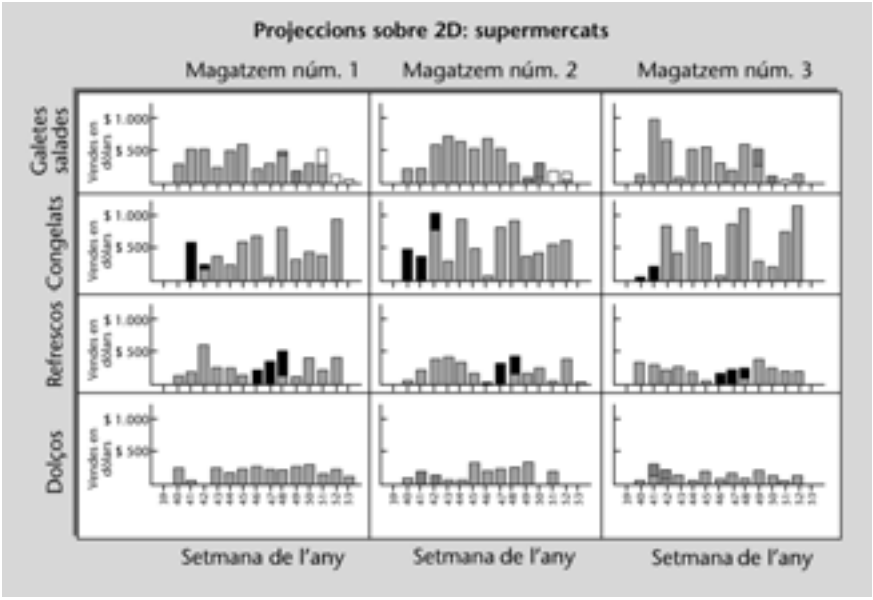
Exemple de tècniques de projecció sobre 2D: sucre i disposició als prestatges

Aquí tenim un diagrama de dispersió que relaciona el contingut de fibra i sucre per a diverses marques de cereals en relació amb la posició que ocupen dins els prestatges d'un supermercat.



Exemple de tècniques de projecció sobre 2D: nivell de vendes en supermercats diferents

El camp d'eines de visualització de dades té una gran activitat. Aquí tenim un altre exemple de projecció sobre 2D:



En aquest exemple tenim una taula que presenta histogrames que relacionen la setmana de l'any en què s'han recollit les dades amb el nivell de vendes assolit per diversos productes (caramels, begudes no alcohòliques, aliments congelats i galetes salades) per a tres supermercats diferents.

Altres eines permeten de combinar diverses formes de visualització.

Exemple d'utilització d'eines de visualització combinades

En la figura següent podem veure com es combina l'estructura d'un arbre de decisió en tres dimensions amb els histogrames que reflecteixen la distribució dels valors de la partició que s'indueix a escala de node:

L'assignació de colors que apareix a la part baixa de la pantalla indica que la variable DDPE5550 té els seus valors distribuïts en rangs segons la partició: valors inferiors a –0,45; valors entre –0,45 i 0,001; valors entre 0,001 i 0,15; valors entre 0,15 i 0,35 i valors superiors a 0,35.

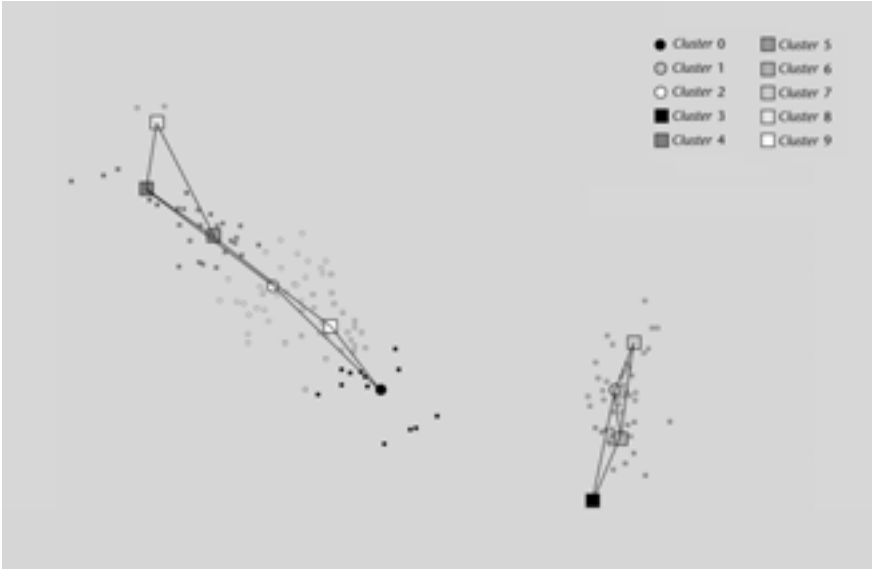
Només per inspecció visual ja es pot veure quins valors predominen en cada partició.

Per al cas de l'**agregació de dades***, en què se cerca trobar grups de dades semblants, la representació de núvols de punts sobre espais bidimensionals permet d'estudiar cada grup d'objectes segons les característiques escollides.

*** En anglès, *clusters*.**

Exemple d'utilitat de projeccions 2D en casos d'agregació

En un cas d'agregació interessa trobar quins grups d'objectes són pròxims entre si. Per tant, una ajuda important és la de representar gràficament el camí que connecta els grups d'objectes (*clusters*) més propers o fortament relacionats.

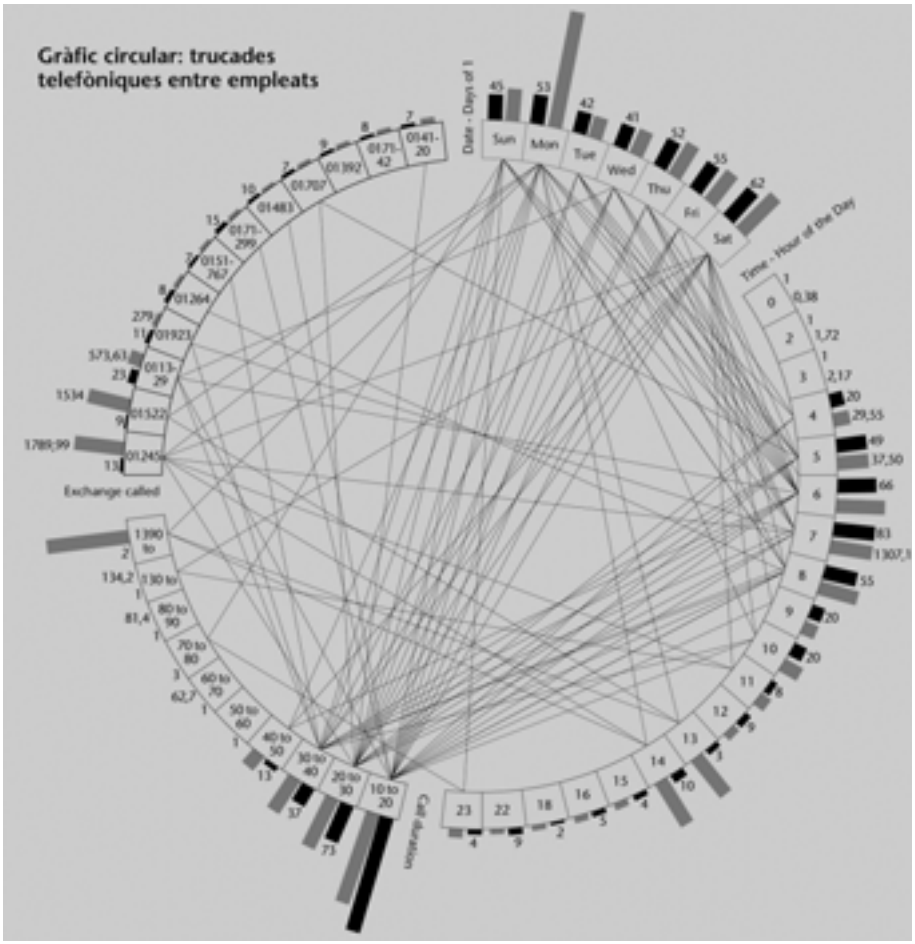


En aquest cas, el gràfic indica que hi ha deu *clusters* i assigna un color diferent als elements de cada un. Llegint els colors es pot veure que hi ha fortes relacions (línies contínues) entre els *clusters* 1, 6 i 4 i entre els grups 5, 8, 0 i 3.

L'àrea de les tècniques de visualització és molt activa i cada cop ofereix més variants de presentació, algunes força curioses, com per exemple els **gràfics circulars**, en què cada node representa alguna mena de variables.

Exemple d'utilitat de gràfic circular

A tall d'exemple de gràfic circular, mostrem aquest gràfic, que relaciona trucades entre diversos números de telèfon dels empleats d'una empresa. Endevineu qui treballa amb qui?



La potència de les diverses eines de visualització es revela per la seva connexió correcta amb la resta d'utilitats per a la selecció i preparació de dades, com també amb els mètodes de mineria de dades posteriors a la primera fase d'intuïció que aporten les claus visuals.

3.2. *Data warehouse*

El concepte original de *data warehouse* va ser presentat per William Inmon i comercialitzat per IBM sota el terme Information Warehousing, que establia l'analogia entre els magatzems físics de les empreses on es podien localitzar de

manera flexible els materials segons la necessitat i l'equivalent pel que fa a les dades d'interès de les diverses àrees de l'empresa.

Lectura complementària

Trobareu informació sobre el concepte original de *data warehouse* en l'obra següent:

W. Inmon (1996). *Building the Data Warehouse* (2a. ed.). Nova York: John Wiley & Sons.

La intenció de la proposta de *data warehouse* és fornir una infraestructura per a prendre decisions amb quatre objectius fonamentals:

a) Regular l'accés als sistemes d'informació i emmagatzemament de dades segons els diversos tipus d'usuaris i grups de treball de manera més flexible i dinàmica que les bases de dades tradicionals.

b) Facilitar la representació de dades i la reconfiguració d'aquesta representació segons les necessitats de prendre decisions de l'empresa, que canvien segons canvia l'entorn competitiu.

c) Construir un model de dades corporatiu que permeti un manteniment i una evolució millors que els models actuals.

d) Mantenir la independència entre els procediments adreçats als usuaris finals i els d'administració de dades, separant un tipus de procediments de l'altre.

El *data warehouse* es pot considerar com una manera d'agrupar dades procedents del sistema de transaccions de l'empresa amb les dades que són necessàries per al treball diari de grups situats en jerarquies intermèdies i els decisors d'alt nivell. No cal dir que cada tipus d'entorn té requeriments diferents i maneres de veure les dades diferents i canviants. El *data warehousing* agrupa diversos tipus d'eines i tecnologies.

Algunes eines i tecnologies...

... que s'agrupen en el *data warehousing* són les següents: plataformes de programari, eines d'extracció i conversió de dades, bases de dades preparades per a consultes complexes i dinàmiques, eines d'anàlisi de dades i eines de gestió de bases de dades.

Algunes de les tecnologies que, sense ser pas noves, són utilitzades o tenen rellevància amb el *data warehousing* són les següents:

1) Sistemes de gestió de bases dades que suportin un procés paral·lel.

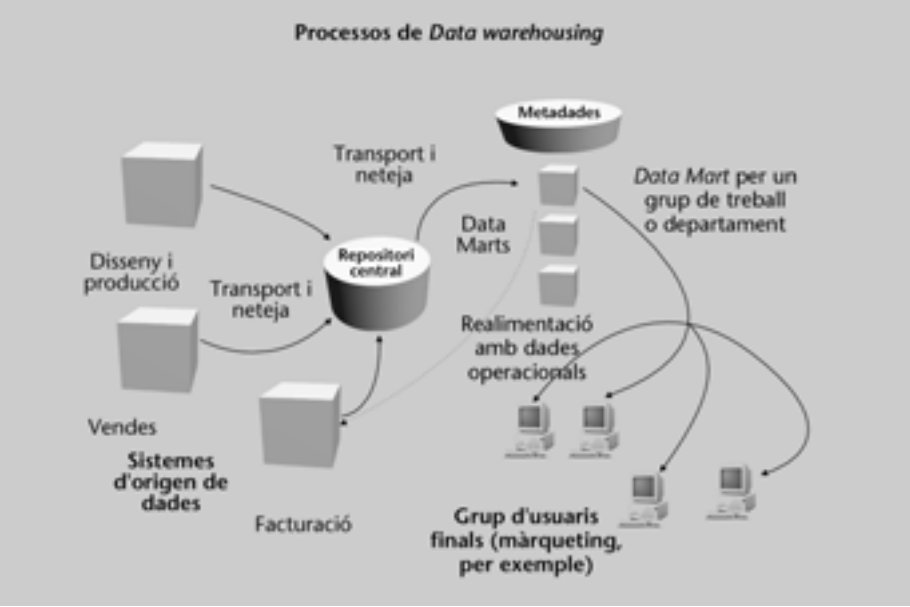
2) Eines de conversió automàtica de dades.

3) Tecnologies client-servidor per a accedir a dades distribuïdes en plataformes diferents.

4) Integració d'eines d'anàlisi i relació amb sistemes de presa de decisió, sistemes de presa de decisió en grup i sistemes d'informació executius.

L'aspecte més crític del *data warehousing* és probablement la modelització dels diversos usos i perspectives que es tenen de les dades, i també la integració transparent dels diversos productes de programari i la seva actualització i millora contínues i tan automàtiques com siguipossible.

Els processos de *data warehousing* fan un ús extensiu de grans volums de dades, en particular dades històriques (de cinc a deu anys) que es manipulen a diversos nivells per a analitzar dades, o sintetitzar-ne de noves, o posar-les en relació amb els factors crítics d'èxit de l'empresa.



Un sistema de *data warehousing* conté normalment els components següents:

1) Sistemes d'origen de dades: sistemes que recullen les dades al nivell més baix (en el sentit que recullen les dades amb un mínim d'abstracció). Per exemple, els que recullen les dades procedents de punts de venda. La tasca del sistema de *data warehousing* és poder integrar les informacions procedents de diverses fonts de manera coherent i aportar una descripció una mica més elevada.

2) Transport i neteja de dades: sistemes de programari que s'encarreguen de "netejar" les dades en el sentit que ja hem explicat, i portar-les cap a altres llocs on es guardaran en la forma o formes adequades. Tradicionalment, aquest tipus de procediment era tasca de programació i resultava difícilment ampliable. Els productes de transport i neteja actualment disponibles adopten una òptica més d'especificació, en la qual s'indica d'on procedeixen les dades i què els ha de passar sense arribar-les a processar. Normalment, aquesta part implica una descripció de les dades en un altre llenguatge de descripció de dades, i crea el que s'anomenen *metadades*.

Vegeu la neteja de dades en el subapartat 2.3.1 d'aquest mòdul.

3) Dipòsit central: lloc principal on es guarden les dades del magatzem. Consta dels elements següents:

a) Maquinari ampliable: l'ampliabilitat del maquinari radica en el fet que permet d'augmentar sense gaire pertorbació tant la rapidesa de càlcul (computació paral·lela) com el volum de dades (entorn dels terabytes).

b) Sistema de bases de dades relacional: les bases de dades relacionals del dipòsit central estan especialment pensades per a millorar la construcció dinà-

mica d'índexs, les operacions de còpia i manteniment i el processament de consultes variades i no estàtiques en el temps.

c) Model lògic de dades: finalment, el model lògic de dades té com a objecte la intercanviabilitat de dades entre els diversos components de l'empresa i la mantenibilitat del dipòsit.

4) Metadades: com hem dit, són dades sobre les dades que introdueixen un grau d'abstracció més elevat respecte als components bàsics, que són les taules i les relacions. Hi ha una gran varietat de components de les metadades que, a més de facilitar la comprensió i l'administració de les dades als administradors del *data warehouse*, cerquen millorar la comprensió i l'accés per part dels usuaris finals.

Els *data warehouse*...

... consten d'eines per a la millora de la comprensió i l'accés per part de l'usuari com ara les següents: anotacions en el model lògic, mapatge del model lògic en els sistemes font de dades, vistes i fórmules més comunes per a accedir a les dades, i informació de seguretat i accés.

5) Data Marts: es tracta de "personalitzar" la visió, els components i els continguts del *data warehouse* segons les necessitats dels diversos grups de treball. Les dades d'una vista combinen les de diverses taules relacionals, probablement distribuïdes.

6) Eines de realimentació* operativa: recullen les dades procedents dels sistemes de presa de decisió i les integren en el dipòsit central. Aquesta és una desviació notable respecte de l'ús tradicional de les eines de presa de decisió operacional. Per exemple, integren criteris per fer comandes a proveïdors en relació amb nivells d'estoc i grau de compliment del proveïdor en qüestió, o integren ajudes per a treballar amb clients. Aquest aspecte del *data warehouse* permet, per exemple, d'oferir suggeriments a un client després que ha respost una sèrie de preguntes directament al personal d'atenció a clients. És un dels aspectes en què les eines de mineria de dades ofereixen més resultats.

*** En anglès, *feedback*.**

La relació entre *data warehousing* i mineria de dades és considerada per alguns com a inclusiva, en el sentit que les eines de mineria de dades formen part de l'entorn de *data warehousing*.

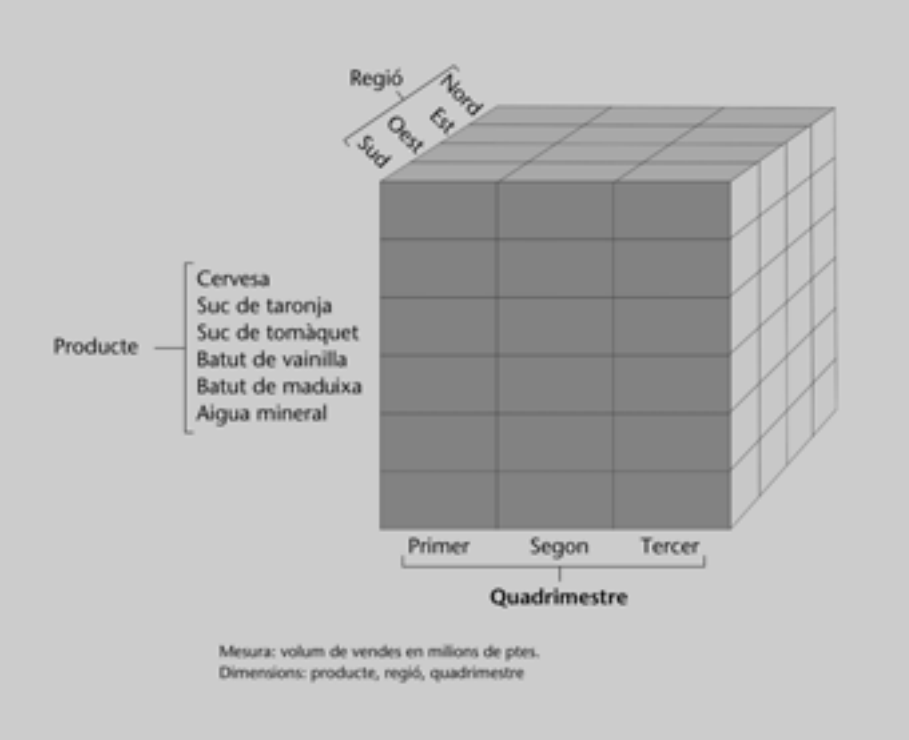
3.3. Mètodes OLAP

Els mètodes OLAP (Codd, 1993) van sorgir per analitzar les dades de vendes i marquèting, com també per processar dades administratives i consolidar dades procedents de diverses fonts amb vista a efectuar una anàlisi de rendibilitat, manteniment de qualitat i altres tipus d'aplicacions que es caracteritzen perquè redefineixen de manera contínua i flexible el tipus d'informació que cal extreure, analitzar i sintetitzar (en comparació amb les bases de dades tradicionals, adreçades a respondre consultes força prefixades i rutinàries).

OLAP és la sigla de l'expressió anglesa *on-line analytical processing*.

Els sistemes OLAP s'alimenten de les dades generades pels sistemes transaccionals (facturació, vendes, producció, etc.). Eines típiques d'OLAP són les que permeten una anàlisi multidimensional de les dades en contra de les típiques

facilitats de creació de resums i informes propis dels sistemes de bases de dades tradicionals.



La **unitat de dades d'OLAP** és el "cub"*, que és una representació de les dades que permet de "tallar-les" i mirar-les des de les perspectives de molts grups diferents d'usuaris. La característica principal dels cubs és que optimitzen les consultes. Normalment es guarden en forma de taula relacional especial que facilita certs tipus de consultes. Per exemple, hi ha columnes de les taules que s'anomen *columnes de dimensió*, que faciliten i preveuen dades per a resums i informes. Les columnes anomenades *columnes agregades* permeten de precalcular quantitats com comptatges, sumes i mitjanes.

*** En anglès, *cube*.**

Construir un cub requereix una anàlisi detallada de les necessitats de dades del grup d'usuaris als quals va adreçat i pot requerir força temps, tant de disseny com d'instal·lació per primera vegada. Compensa pel fet que facilita extraordinàriament les tasques d'anàlisi de dades dels diversos grups d'usuaris i, un cop establert, resulta més senzill de modificar que les taules relacionals tradicionals.

3.4. Sistemes OLTP

Els sistemes de processament de transaccions en línia (OLTP) tenen com a objectiu guardar la integritat de les dades necessàries per a administrar una organització de manera eficient.

OLTP és la sigla de l'expressió anglesa *on-line transactional processing*.

Per tant, els sistemes OLTP cerquen mantenir models de dades que corresponguin a la visió que cada empleat (o tipus d'empleat) té de l'organització. En comptes de veure l'organització com una estructura de dades organitzada entorn de taules i relacions, les eines d'OLTP la presenten en forma de jerarquies i dimensions, de manera que es poden mirar les mateixes dades des de perspectives diferents.

Els sistemes tradicionals són dinàmics en el sentit que sempre són actualitzats amb dades noves. Per a poder-los analitzar cal fer una "fotografia" del seu estat en un moment donat i aplicar-hi les eines d'anàlisi corresponents. Fer aquest tipus de feina només amb les operacions de consulta pròpies de les bases de dades tradicionals no és fàcil i pot induir a una degradació del rendiment general del sistema. Igualment, el sistema de bases de dades pot no estar preparat per a guardar el resultat d'aquestes anàlisis.

En canvi, els sistemes d'OLTP (com els d'OLAP, en certa manera) permeten de descarregar el sistema central (ocupat, potser, en processos transaccionals) i efectuar aquest tipus d'operació al mateix temps que permet de guardar els seus resultats. Les eines de mineria de dades poden donar algun servei en aquesta mena d'anàlisis.

3.5. Estadística

La tasca d'analitzar grans volums de dades ha estat i continua essent el reialme de l'estadística, en concret, l'anàlisi de dades. L'enfocament tradicional de l'estadística s'adreça a la recollida de dades adequada per a la interpretació, en particular a la inferència de característiques d'una població a partir de les mostres recollides.

La idea de mineria de dades ha posat les tècniques estadístiques clàssiques davant una gran oportunitat pràctica i també davant la necessitat de crear eines que, tot i mantenir la sòlida fonamentació teòrica aportada per aquesta disciplina, donin una resposta fàcilment comprensible a usuaris no sempre ben preparats estadísticament dins els límits de temps imposats per la velocitat que requereixen els nous entorns de treball.

Mètodes típics procedents de l'estadística...

... podrien ser bona part dels mètodes de classificació i d'agregació de dades (en anglès, *clustering*), models predictius (anàlisi de regressió, per exemple) i mètodes de construcció de models gràfics (xarxes bayesianes, per exemple) (Pearl, 1981; Whittaker, 1990).

3.6. Aprenentatge automàtic

L'aprenentatge automàtic és la part de la intel·ligència artificial que estudia com els sistemes intel·ligents són capaços de desenvolupar coneixements i habilitats nous a partir de la seva experiència.

En concret, els mètodes d'aprenentatge inductiu (Michell, 1985) cerquen l'extracció de conceptes nous, pautes de conducta noves, plans nous i, en general, coneixements nous a partir de l'observació de les dades de l'entorn o del mateix comportament del sistema intel·ligent. La plètora de mètodes aportats des d'aquest camp i la seva insistència a fer prevaler l'expressió simbòlica i no numèrica del coneixement també han convertit els seus mètodes en rellevants per a la tasca de la mineria de dades.

Mètodes d'aprenentatge inductiu

Bona part (si no tots) dels mètodes basats en lògica, casos, xarxes neuronals, regles de classificació, xarxes neuronals, xarxes d'associació i algorismes genètics procedeixen d'aquesta comunitat de recerca.

Es pot dir, en resum, que la mineria de dades integra resultats de disciplines com són les bases de dades (amb la seva extensió a *data warehousing*, OLAP i OLTP), l'estadística, l'aprenentatge automàtic i la visualització.

Hem de dir que la mateixa proposta de mineria de dades ha generat una gran activitat en tots aquests camps, que s'han vist obligats a modificar alguns dels supòsits per a arribar a donar la qualitat de resultat exigida pels nous objectius.

4. Cas d'estudi: la cadena Hyper-Gym

El cas que utilitzarem per a practicar la metodologia general de mineria de dades, que el tindrem com a "camp de proves" per a les diverses tasques de mineria de dades que ens proposarem i els diversos models que voldrem obtenir amb mètodes diferents, és el de la cadena de centres esportius Hyper-Gym. Es tracta d'una simplificació de diversos estudis reals dins el mateix camp. Evidentment, Hyper-Gym no existeix com a empresa.

L'activitat principal dels diversos centres esportius de la cadena en qüestió correspon a l'oferiment de les seves instal·lacions als seus clients perquè desenvolupin activitats esportives lliures o dirigides. D'aquests centres, n'hi ha un total de tres. Dos es troben en barris de classe mitjana-alta i l'altre en un barri de classe alta. Tot i així, les activitats que s'ofereixen en cadascun són pràcticament les mateixes. Tots els centres disposen d'una sala de musculació on es poden fer diversos exercicis amb peses i màquines multiexercici; d'una a quatre sales lliures destinades a fer classes de diversos tipus (ioga, TBC*, aeròbic, estiraments**, gimnàstica sueca, etc.); d'una a tres sales amb màquines automàtiques d'exercicis cardiovasculars (bicicletes, màquines de rem, cintes rodants, etc.) i una piscina de grandària diversa depenent de quin centre es tracti. També ofereix altres serveis: saunes, banys turcs, cabina d'estètica, etc.

*** Inicials de *total body conditioning*.**

**** També anomenat *stretching* o *stretch* en els fullets publicitaris d'Hyper-Gym.**

Els clients d'Hyper-Gym han d'escollir la modalitat d'ús a la qual es volen adherir. Hi ha diverses tarifes mensuals que corresponen a diversos horaris. Per exemple, hi ha un horari d'hora que va de 7 a 9 del matí, un altre de 9 a 12, de 12 a 17, de 17 a 21 i de 21 a 11. Per simplificar, però, les dades que utilitzarem només distingiran entre matí i tarda. Quan un client s'associa a un centre d'Hyper-Gym, escull un horari en què hi pot entrar. Evidentment també hi ha la tarifa d'entrar-hi en qualsevol horari, que és més cara que les altres. Finalment, amb un altre suplement és possible accedir a qualsevol hora a qualsevol dels centres d'Hyper-Gym, encara que siguin diferents d'aquell al qual inicialment s'havia associat un client.

Cada client rep una targeta magnètica identificativa amb la seva foto digitalitzada, que li permet d'acreditar-se i entrar a les instal·lacions i sortir-ne. Aquesta targeta obre una barrera de pas a l'entrada de cada centre. Si es comprova que es tracta d'entrar en un horari que no és el cobert per la modalitat d'inscripció del client, la barrera no s'obre fins que no s'abona una quantitat suplementària. En el moment de sortir també cal fer ús de la targeta per a poder desbloquejar la barrera de sortida. Així, es poden recollir dades que els gestors de la cadena consideren de possible utilitat.

Si el client no indica res més, es dóna per suposat que vol fer servir lliurement les instal·lacions d'Hyper-Gym. Ara bé, té altres alternatives:

a) Es pot matricular a diverses activitats dirigides que tenen el seu propi horari: musculació, gimnàstica de manteniment, aeròbic, gimnàstica aquàtica o *aquagym*, dansa aeròbica, ioga, estiraments, etc.

b) Igualment pot decidir l'ús dels diversos estris d'entrenament amb ajuda d'un entrenador personal. En aquest cas, cal fer un estudi mèdic que reflecteixi l'estat general del client. Tenint en compte resultats de les proves mèdiques i els objectius d'entrenament declarats pel client, l'entrenador confecciona un pla d'entrenament. Aquests plans són prou estàndard i corresponen a diversos objectius. En principi, es descriuen sota les denominacions de "Millora cardiovascular", "Reducció de pes", "Elasticitat", "Manteniment" i "Alt rendiment". Un client pot iniciar el seu pla d'entrenament personal en qualsevol moment, deixar-lo, reprendre'l i canviar d'entrenador personal tantes vegades com vulgui.

Igualment, un client pot abandonar la relació amb Hyper-Gym en qualsevol moment per poder-la reiniciar en un futur. Alguns dels problemes als quals es volen adreçar els administradors d'Hyper-Gym corresponen a saber quina mena de clients són més duradors i per què. Dit d'una altra manera, volen saber les causes que fan que una persona abandoni la pràctica esportiva i així poder emprendre les accions (promocions, descomptes, canvis d'activitats, etc.) que els permetin de mantenir el client més temps. Aquesta és una de les possibilitats. Una altra cosa que també els interessa saber és si els clients, tot i tenir un horari assignat, fan ús d'altres hores, fins i tot pagant. Això pot permetre de redissenyar els horaris de manera que s'assegurin més beneficis i més permanència de clients. També volen saber si realment l'objectiu que han d'assolir és mantenir els clients fidels molt de temps o si els interessa més que hi hagi una alta rotació de clients.

Aquestes són algunes de les qüestions que es plantegen i a algunes ens hi adreçarem amb els mètodes i tècniques que aprendrem durant l'assignatura. D'altres, les deixarem de banda, però us proporcionaran un bon punt de partida per jutjar mètodes nous que semblin adreçats a resoldre-les.

Els responsables d'Hyper-Gym saben que tenen moltes dades i en volen extreure la màxima utilitat possible.

La base de dades d'Hyper-Gym és força senzilla. Com veureu, però, a l'hora de plantejar-se extreure'n determinades informacions ens trobarem amb més d'un problema. Succeeix que, com en la majoria de casos, els administradors de la cadena, en crear-la, no van anticipar moltes de les utilitats que se'n podien extreure. De fet, reflecteix una preponderància de l'ús administratiu de les dades que recull.

La relació principal és la de *Clients*, que manté els clients actuals i els dels darrers dos anys. Les dades que es mantenen en aquesta relació es comenten a continuació.

Atribut	Valors possibles		Descripció
Número de soci	0000-99999		Identificador dins Hyper-Gym
NIF	00000000-9999999		A efectes de facturació, la lletra es calcula automàticament
Nom	20 caràcters		Nom del soci
Cognom 1	20 caràcters		Primer cognom
Cognom 2	20 caràcters		Segon cognom
Carrer	20 caràcters		Adreça
Núm.	3 números		Adreça
Població	20 caràcters		Adreça
Codi postal	6 caràcters		Adreça
Professió	1	Assalariat baix	Simplificació de la classificació oficial d'activitats professionals
	2	Assalariat mitjà	
	3	Assalariat alt	
	4	Professional liberal	
	5	Autònom-comerç	
	6	Autònom-d'altres	
	7	Funcionari	
	8	D'altres	
Data d'alta	6 números		Format DD/MM/AA
Data-de baixa	6 números		Format DD/MM/AA
Horari actual	1	7 a 9	
	2	9 a 12	
	3	12-17	
	4	17-21	
	5	21-11	
	6	Lliure	
Entrenador actual	3 nombres		Identificador de l'entrenador
Pla d'entrenament actual	C	Cardiovascular	
	M	Manteniment	
	H	Alt rendiment	
	P	Reducció pes	
	E	Elasticitat	
Centre	1-3		On està matriculat
Activitat principal	Ioga, *stretch*, TBC, etc.		Activitat que fa preferentment

Atribut	Valors possibles	Descripció
Activitats secundàries	loga, *stretch*, TBC, etc.	Segona activitat en ordre de preferència
Sexe	H-D	
Renda	0-10.000.000	
Edat	0-99	
Ús piscina	Sí-No	
Propietari de pis	Sí-No	
Descompte	0%-10%	

Per cada client es disposa d'un històric en què s'indiquen la data, l'hora i el minut tant de l'entrada com de la sortida de cada centre on ha efectuat alguna activitat. Fixeu-vos que no hi ha cap limitació en aquest sentit. Un client pot anar a tants centres com vulgui durant el dia i fer tantes activitats lliures com vulgui.

Les altres taules són relativament menys rellevants. Com veureu, en conjunt, hi ha un munt de redundàncies i descriptors que poden generar inconsistències:

a) Aquí tenim l'històric d'entrades i sortides.

Atribut	Valors	Descripció
DNI soci	00000000-99999999	Identificador soci
Moviment	Entrada/Sortida	Tipus de moviment
Data	DD/MM/AA	Dia
Hora	HH:MM	Hora i minut
Centre	1-3	Centre on es recull la dada

b) Històric d'activitats dirigides:

Atribut	Valors	Descripció
DNI soci	00000000-99999999	Identificador soci
Moviment	Alta/Baixa	Inici o final activitat
Data	DD/MM/AA	Dia
Centre	1-3	Centre on ha iniciat/finalitzat l'activitat

c) Històric d'horaris:

Atribut	Valors	Descripció
DNI soci	00000000-99999999	Identificador soci
Horari	1-6	Franja horària
Data-d'alta	DD/MM/AA	Dia inici
Data-de baixa	DD/MM/AA	Dia final

Insistim...

... en el fet que, en els arxius per a dur a terme les pràctiques, les dades queden molt simplificades, però mantenim aquesta diversitat per tal que cadascú proposi possibilitats d'exploració noves.

Les dates d'alta i baixa indiquen en quin moment el client decideix accedir al gimnàs en una determinada franja horària i en quin moment canvia d'opinió.

d) Històric d'entrenaments:

Atribut	Valors	Descripció
DNI soci	00000000-99999999	Identificador soci
Entrenament	1-6	Tipus d'entrenament
Data-d'alta	DD/MM/AA	Inici
Data de-baixa	DD/MM/AA	Final

Aquesta relació recull els diversos tipus d'entrenament que ha pogut seguir un client.

e) Històric d'entrenadors:

Atribut	Valors	Descripció
DNI soci	00000000-99999999	Identificador soci
Entrenador	00000000-99999999	Identificador entrenador
Data-d'alta	DD/MM/AA	Inici entrenament
Data-de baixa	DD/MM/AA	Final entrenament

En aquesta relació es guarda informació respecte a amb quins entrenadors ha practicat un determinat client.

Els administradors d'Hyper-Gym tenen diverses preguntes al pensament:

- Determinar quines són les característiques dels clients que demanen el servei d'entrenador personal.
- Determinar quant de temps es manté un client en una activitat determinada.
- Veure si en aquest darrer sentit hi ha diferències entre activitats.
- Veure si hi ha entrenadors "conflictius": que mantenen els seus clients per poc temps.
- Determinar què vol dir "per poc temps".
- Intentar veure amb les dades de què es disposa si hi ha cap raó relacionada amb la tipologia dels clients que determini que certs entrenadors mantinguin poc temps els seus clients.
- Relacionar les característiques dels clients i les diverses franges horàries.

- Conèixer quina mena de clients tenen; en concret, si hi ha cap mena de similitud entre clients i, si n'hi ha, en quants grups es podrien dividir els clients. Això podria permetre, per exemple, de dissenyar ofertes de descomptes per activitat segons els tipus de clients.

Més endavant anirem presentant nous problemes, qüestions i modificacions d'aquestes informacions inicials per tal de poder-hi aplicar els diversos mètodes que anirem descrivint.

Resum

La mineria de dades cerca l'extracció de coneixement nou, vàlid i útil per als objectius que es plantegi qui emprengui aquest procés. El resultat d'un procés de mineria de dades és un model que ha de ser el màxim de *comprensible* possible. És important poder interactuar amb aquest procés i aprofitar el coneixement *a priori* de què es disposi.

Els processos de mineria de dades es basen en resultats procedents de la recerca i desenvolupament en bases de dades, estadística, aprenentatge automàtic i visualització.

La mineria de dades s'ha d'entendre com un procés continu que integra els aspectes següents:

a) Definició de l'objectiu del projecte de mineria de dades, precisar la tasca principal que cal fer i escollir el mètode més adient segons les circumstàncies.

b) Selecció de les dades rellevants.

c) Preparació de les dades per a assegurar-se que siguin vàlides i estiguin en condicions de ser utilitzades pel mètode escollit.

d) Mineria de dades pròpiament dit; és a dir, aplicació sobre les dades ja preparades del mètode escollit i construcció del model corresponent.

e) Interpretació del model obtingut, que pot fer revisar algunes de les fases anteriors.

f) Integració en el sistema de tractament d'informació, que comprèn l'observació del rendiment i, en cas de canvi de l'entorn o "envelliment" del model, inici d'un procés de mineria de dades nou.

Activitats

1. Proposeu un possible projecte de mineria de dades que es correspongui amb la vostra àrea d'activitat professional.
a) Expliciteu clarament quins són els objectius del projecte.
b) Intenteu veure amb quina tasca té més semblances (agrupació, predicció, classificació, etc.).
c) Identifiqueu fonts d'informació possibles.
d) Definiu el format de les dades de què disposeu.
e) Identifiqueu quines operacions (fusió, separació, derivació, etc.) caldria fer sobre les dades.
f) Digueu quins atributs penseu, *a priori*, que no han de ser necessàriament rellevants per a la tasca que us heu proposat.
g) Proposeu alguna forma d'avaluació del model obtingut.

2. Proposeu un possible projecte de mineria de dades que no es correspongui amb la vostra àrea d'activitat professional.
a) Indiqueu sobre quin sector d'activitat (empresarial, enginyeria, científica, etc.) heu pensat que es poden aplicar tècniques de mineria de dades.
b) Expliciteu clarament quins són els objectius del projecte.
c) Intenteu veure amb quina tasca té més semblances (agrupació, predicció, classificació, etc.).
d) Identifiqueu fonts d'informació possibles.
e) Definiu el format de les dades de què voldríeu disposar.
f) Identifiqueu quines operacions (fusió, separació, derivació, etc.) caldria fer sobre les dades.
g) Digueu quins atributs penseu, *a priori*, que no han de ser necessàriament rellevants per a la tasca que us heu proposat.
h) Proposeu una manera d'avaluar la qualitat del model obtingut.

3. Accediu a la pàgina web indicada al marge i cerqueu informació sobre projectes que s'adrecin a resoldre el problema que us plantejàveu en l'activitat 1 i projectes que ho facin respecte a l'activitat 2.

WEB
http://www.kdnuggets.com
Accediu a aquesta pàgina web per fer les activitats 3 i 5.

4. Escrigueu un breu resum en què compararéu allò que heu trobat aplicat i allò que vosaltres proposàveu en les activitats 1 i 3.

5. Accediu a la pàgina web indicada al marge i cerqueu dins de "Software" per a cada tipus de tasca que hem identificat (agregació, classificació, etc.) una eina que sembli que permeti, efectivament, de construir els models adequats a la tasca.

Glossari

avaluació de models *f* Assignació de valors de qualitat als models resultants d'un procés de mineria de dades. La mesura de qualitat depèn del tipus de model que es vol obtenir.

coneixement *m* Creença justificada per un procés d'interpretació i validació. L'entenem com a sotmès sempre a revisió segons com canviïn les circumstàncies.

conjunt d'entrenament *m* Subconjunt del conjunt original de dades emprat per a construir un primer model.

conjunt de prova *m* Subconjunt del conjunt original de dades emprat per a avaluar la qualitat del model.

dada categòrica *f* Dada que pren valor en un conjunt finit, normalment no numèric.

dada numèrica *f* Dada que pren valor en un conjunt numèric continu.

data warehouse *m* Tecnologia que estén les bases de dades tradicionals i que afegeix noves utilitats que permetin de definir diverses perspectives sobre les dades per a cada grup de treball de l'empresa i diversos nivells d'abstracció de les dades, facilitar l'anàlisi de les dades per a cada tipus de grup i facilitar la presa de decisions, inclosa la que es fa des del punt de vista operatiu.

derivació de dades *f* Procés d'obtenció d'atributs nous a partir dels ja existents.

discretització *f* Transformació de dades numèriques a categòriques.

extracció de coneixement *f* Procés d'extracció de coneixements vàlids, útils i comprensibles a partir de grans volums de dades.

integració *f* Incorporació del model obtingut en un procés de mineria de dades al sistema d'informació de l'empresa.

interpretació *f* Fase que involucra el judici humà per treure sentit al model resultant d'un procés de mineria de dades.

mètode no supervisat *m* Mètode que es caracteritza pel fet que només s'utilitzen les observacions recollides sense afegir-hi cap informació de categorització o classificació que permeti d'establir diferències entre allò que són exemples d'una classe i d'una altra. Terme procedent de l'aprenentatge automàtic. // Mètode que actua sense la supervisió d'un tutor humà.

mètode supervisat *m* Mètode que es caracteritza pel fet utilitza les observacions recollides i una o més informacions addicionals que indiquen informació de classificació. En cas que la classificació sigui binària, es parla d'*exemples* (observacions que pertanyen a una classe) i *contraexemples* (observacions que no hi pertanyen). Terme procedent de l'aprenentatge automàtic. // Mètode en què l'aprenentatge és guiat per un tutor humà.

mineria de dades *f* Part de l'extracció de coneixement a partir de bases de dades que s'encarrega de la construcció de models.

model *m* Descripció d'una realitat utilitzant un determinat nivell d'abstracció; és a dir, determinats conceptes i relacions.

neteja de dades *f* Procés d'eliminació de dades errònies. Inclou eliminació de redundàncies, valors impossibles, etc.

reducció de dimensionalitat *f* Procés mitjançant el qual es limiten o bé el conjunt total de dades que cal tractar, o bé el conjunt d'atributs, o bé el conjunt de valors que pren cada atribut. D'aquesta manera s'intenta reduir la complexitat del procés de mineria de dades sense afectar la qualitat del resultat final.

selecció d'atributs *f* Part del procés de reducció de dimensionalitat adreçada a reduir el nombre d'atributs del conjunt original de dades.

selecció de dades *f* Detecció de les fonts de dades necessàries per a portar endavant un procés d'extracció de coneixement.

transformació de dades *f* Cada un dels tipus de procediments emprats per a deixar les dades preparades per a la seva utilització per un mètode de mineria de dades.

Bibliografia

Adriaans, P.; Zantinge, D. (1996). *Data mining*. Boston: Addison-Wesley.

Anand, T.; Kahn, G. (1992). "SPOTLIGHT: A Data Explanation System". *Proceedings of the Eighth IEEE Conference on Applied Artificial Intelligence* (pàg. 2-8). Washington D.C.: IEE Computer Society.

Baum, D. (1996). "*Data warehouse*: Building Blocks for the Next Milennium". *Oracle Magazine* (núm. 2, pàg. 34-43).

Berson, A.; Smith, S.J. (1997). *Data warehousing, Data mining, and OLAP*. Nova York: McGraw-Hill.

Breiman, L.; Friedman, H.; Olshen, R.; Stone, C. (1984). *Classification and Regression Tress*. Wadsworth: Chapman & Hall.

Cendrowska, J. (1987). "PRISM: An algorithm for inducing modular rules". *International Journal of Man-Machine Studies* (vol. 27, núm. 4, pàg. 349-370).

Clark, P.; Boswell, R. (1991). "Rule Induction with CN2: Some Recent Improvements". A: Y. Kodratoff (ed.). *Proceedings of the European Working Session on Learning: Machine Learning* (pàg. 151-163). Berlín: Springer Verlag.

Codd, E.F. Codd, S.B.; Salley, C.T. (1993). *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*. E.F. Codd Associates.

Fayyad, U.; Uthurusamy, R. (ed.) (1995). *Proceedings of the First International Conference On Knowledge Discovey and Data mining (KDD-95)*; Cambridge: MIT Press.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (ed.) (1996). "Advances in Knowledge Discovery and Data mining". Cambridge: MIT Press.

Holland, J.H.; Holyoak, K.J., Nisbett, R.E.; Thaggard, P.R. (ed.) (1986). *Induction: Processes of Inference, Learning and Discovery*. Cambridge: MIT Press.

Holsheimer, M.; Siebes, A. (1994, gener). "Data mining: the Search for Knowledge in Databases". *MReport Tecnic CS-R9406*. Amsterdam: Centrum voor Winskunde & Informatica, CWI.

Huan Liu, H.; Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data mining*. Boston: Kluwer Academic Publishers.

Inmon, W. (1996). *Building the Data warehouse* (2a. ed.). Nova York: John Wiley & Sons.

Kimball, R. (1996). *The Data warehouse Toolkit: Practical Techniques for Building Dimensional Data warehouses*. Nova York: John Wiley & Sons.

Michalski, R.S. (1983). "A Theory and a Methodology of Inductive Learning". A: R.S. Michalski; J.G. Carbonell; T.M. Mitchell (ed.). *Machine Learning: An Artificial Intelligence Approach*. Los Altos: Morgan Kaufmann Publishers.

Mitchell, T.M. (1997). *Machine Learning*. Nova York: McGraw-Hill.

Molina, L.C. (1998, agost). *Data Mining no Processo de Extraçao de Conhecimiento de Bases de Dados. Master Thesis*. São Paulo: Instituto de Ciências Matemáticas de Sao Carlos, Universidade de Sao Paulo.

Muggleton, S. (ed.) (1992). *Inductive Logic Programming*. Londres: Academic Press.

Muggleton, S.; De Raedt, L. (1994). "Inductive Logic Programming: Theory and Methods". *Journal of Logic Programming* (núm. 19, pàg. 629-579).

Piatetsky-Shapiro, G.; Mateus, C.; Smyth, P.; Uthurusamy, R. (1993). "KDD-93: Progress and Challenges in Knowledge Discovery in Databases". *AI Magazine* (núm. 15, vol. 3, pàg. 77-87).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann Publishers.

Pyle, D. (1999). *Data Preparation For Data mining*. Morgan Kaufmann Publishers.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

Rezende, S.O.; Oliveira, R.B.T.; Félix, L.C.M.; Rocha, C.A.J. (1998, setembre). "Visualization for Knowledge Discovery in Databases". *Proceedings of the International Conference on Data mining*. Rio de Janeiro.

Ripley, B.D. (1994). "Neural Networks and Related Methods for Classification". *Journal of the Royal Statistical Society* (vol. 3, núm. 56, pàg. 409-437).

Schmitz, J.; Armstrong, G.; Little, J.D.C. (1990). "CoverStory-Automated News Finding in Marketing". *DSS Transactions*.

Simoudis, E.; Fayyad, U.M. (1997, març). "Data mining Tutorial". *First International Conference on the Practical Applications for Knowledge Discovery in Data Bases*. Londres.

Weiss S.; Kulikowski, C. (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. San Francisco: Morgan Kaufmann.

Wen, S.W.; Hernández, R.; Naylor, C.D. (1995). "Pitfalls in Nonrandomized Studies: The case of incidental Appendectomy with Open Cholecystectomy". Journal of the American Medical Association (núm. 275, pàg. 1687-1691).

Westphal, C.I.; Blaxton, T. (1998). *Data mining Solutions. Methods and tools for Solving Real-World Problems*. Nova York: John Wiley & Sons.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Nova York: John Wiley & Sons.

Dipòsits de dades per a practicar i comparar mètodes de mineria de dades

Department of Information and Computer Science. Universitat de Califòrnia a Irvine, Estats Units. *UCI ML Database Repository*. http://www.ics.uci.edu/~mlearn/MLRepository.html.
Dipòsit de bases dades per a aprenentatge automàtic. Lloc web on s'apleguen diversos conjunts de dades d'ús públic per a poder fer proves i comparacions amb nous mètodes d'aprenentatge automàtic. Alguns dels conjunts de dades que es tracten en les activitats s'han baixat d'aquí i vosaltres mateixos n'haureu de baixar algun altre. Els més coneguts són: la *SoyaBean*, que recull dades sobre les malalties de les plantes de soja i que va ser la base per a desenvolupar un sistema expert de diagnosi utilitzat en agricultura amb molt d'èxit; la base de dades *Iris* per a la classificació de plantes; la base de dades *Moviegoers database*, sobre assistència i preferències del públic de cine; la *Credit Screening Database* sobre dades per a la concessió de crèdits; la base de dades *WWTP* (*Wastewater Treatment Plants*), amb les dades de la depuradora d'aigües urbanes de Manresa i Girona). Són conjunts que presenten alguna dificultat especial (manca de valors, dades poc precises, etc.). Periòdicament se n'afegeixen més.

GMD (Gesellschaft für Matemathik und Datenverarbeitung). http://www.gmd.de/ml-archive/frames/datasets/datasets-frames.html.
Dipòsit del GMD, un dels centres de recerca europeus més importants en informàtica i intel·ligència artificial. Replica alguns dels conjunts de dades del dipòsit d'UCI Irvine, però també n'aporta de molts altres, resultat del fet que GMD també és el nucli dels dipòsits de materials al World Wide Web de la Xarxa Europea d'Excel·lència en recerca en Aprenentatge automàtic (MLNET). També té moltes publicacions sobre nous mètodes i les seves aplicacions.

Universitat Carnegie Mellon. *Monk's problems*. http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/thrun/papers/thrun.MONK.html.
Tot i que el conjunt de dades conegut com a *Monk's problem* també es pot trobar a l'UCI ML Database Repository, la ubicació original és la Universitat Carnegie Mellon.
Podeu trobar més de cent pàgines en què es discuteix el rendiment de diversos mètodes d'aprenentatge sobre conjunts de dades de referència. Resultat de l'escola d'estiu sobre aprenentatge automàtic que va tenir lloc a Brussel·les l'any 1991. Alguns dels mètodes posats a prova són (agrupats per famílies de mètodes) els següents:

- AQ17-DCL, AQ17-HCI, AQ17- AQ14-NT, AQ15-GA, AQR.
- Assistant Professional.
- mFOIL.
- ID5R, IDL,ID5R-hat, TDIDT, ID3.
- CN2.
- CLASSWEB.
- ECOBWEB.
- PRISM.
- *Backpropagation* (xarxes neuronals).