Procés de mineria de dades

PID_00284563

Julià Minguillón Alfonso
Ramon Caihuelas Quiles

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

Julià Minguillón Alfonso

Llicenciat en Enginyeria Informàtica el 1995, màster en Combinatòria i Comunicació Digital el 1997 i doctor enginyer d'Informàtica el 2002 (Universitat Autònoma de Barcelona). Des del 2001 exerceix com a professor als Estudis d'Informàtica, Multimèdia i Telecomunicació de la Universitat Oberta de Catalunya. Pertany al grup de recerca Learning Analytics for Innovation and Knowledge Application (LAIKA), on desenvolupa projectes d'investigació relacionats amb l'anàlisi i la visualització del comportament dels usuaris d'entorns virtuals d'aprenentatge i xarxes socials.

Ramon Caihuelas Quiles

Llicenciat en Ciències de la Informació per la Universitat Autònoma de Barcelona. Postgraduat en Disseny d'Aplicacions per la Universitat Politècnica de Catalunya. Màster en Gestió de Tecnologies de la Informació per Enginyeria La Salle (Universitat Ramon Llull). Doctorand en Informàtica per Enginyeria La Salle (Universitat Ramon Llull). Actualment treballa com a responsable del grup de bases de dades i suport al desenvolupament de l'Àrea de Tecnologies de la Universitat de Barcelona, i com a professor col·laborador dels estudis d'Informàtica d'Enginyeria La Salle (Universitat Ramon Llull) i de la Universitat Oberta de Catalunya.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Julià Minguillón Alfonso

Primera edició: setembre 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Julià Minguillón Alfonso, Ramon Caihuelas Quiles
Producció: FUOC

Índex

Introducció

La mineria de dades ha d'inscriure's dins d'un procés d'abast més ampli: el descobriment de coneixement dins de grans bases de dades o KDD, que és la sigla del terme anglès *knowledge discovery in databases*.

Nota

Aquest mòdul és una revisió del mòdul previ escrit per Ramon Sangüesa i Solé.

En aquest mòdul definim els objectius del procés de descobriment de coneixement a partir de dades, en delimitem les diferents fases, ens centrem en la fase de mineria de dades i comentem els problemes més freqüents i importants inherents a tot el procés.

Objectius

Els objectius d'aquest mòdul són els 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. Anticipar-se als problemes concrets que es presenten en cadascuna de les fases del projecte.

1. Descobriment de coneixement en grans volums de dades

Des d'un punt de vista acadèmic, es considera la mineria de dades com a part d'un procés més gran anomenat *descobriment de coneixement a partir de dades*. No obstant això, actualment alguns autors fan servir indistintament les dues denominacions (*descobriment de coneixement* i *mineria de dades*). Nosaltres també emprarem tots dos termes de manera indiferent.

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

«El procés de KDD (knowledge discovery in databases) és el procés no trivial consistent a descobrir patrons vàlids, nous, potencialment útils i comprensibles dins d'un conjunt de dades».

Referència bibliogràfica

L'article original en què apareix aquesta definició és: G. Piatetsky-Shapiro; C. Mateus; P. Smyth; R. Uthurusamy (1993). «KDD-93: Progress and Challenges in Knowledge Discovery in Databases». *AI Magazine* (vol. 15, núm. 3, pàg. 77-87).

Remarquem aquí alguns aspectes que poden passar per alt en una primera lectura de la definició que acabem d'apuntar:

1) «Patrons vàlids»: cal entendre aquest concepte com a coneixement correcte contrastable amb la realitat.

2) «Potencialment útils»: la utilitat està en relació amb l'objectiu que ens proposem a l'hora de dur a terme 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, tot i que és completament inútil per decidir si un client serà fidel a l'empresa durant els pròxims sis mesos.

3) «Comprensibles»: la comprensibilitat està relacionada amb l'usuari que maneja el coneixement, els patrons, etc., extrets de les dades. De manera que 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, els paràmetres d'un model de regressió són prou comprensibles; per a un usuari menys preparat, potser és més comprensible veure un gràfic en pantalla.

Una altra definició que posa èmfasi en la part que correspon a la mineria de dades és la que van presentar Holsheimer i Siebes. Aquests autors defineixen la mineria de dades de la manera següent:

«La mineria de dades és el procés consistent a trobar models comprensibles a partir de grans volums de dades».

En aquesta definició el realment important és la paraula *model*. Un model és una descripció articulada i abstracta d'una realitat.

Referència bibliogràfica

El treball original en què apareix aquesta definició és: M. Holsheimer; A. Siebes (1994). *Data mining: the Search for Knowledge in Databases*. Report tècnic CS-R9406 (gener). Amsterdam: Centrum voor Winskunde en Informatica (CWI).

Per descriure un edifici, podem tenir un model que es basi únicament en conceptes estructurals, o que només tingui en compte la distribució de conductes elèctrics. El que sí que és important i hem de tenir en compte és que hi ha uns models que s'avenen millor que uns altres a la mena d'utilitat que volem obtenir de les dades. Ja tornarem a tractar aquest tema més endavant.

L'extracció del coneixement a partir de bases de dades és un procés complex que es dirigeix a l'obtenció de models de coneixement a partir de dades recollides en una o més bases de dades en virtut d'uns objectius determinats. Aquest procés consistent a passar de dades a coneixement suposa un canvi de llenguatge d'expressió i implica diverses fases.

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 proposat.

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

- Comprensible per a qui duu a terme el projecte quan analitza el coneixement extret.

- Directament integrable en les operacions de l'empresa, tot i 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 extregui aquesta ocurrència representa un canvi de llenguatge poc potent. En canvi, un que ens doni com a resultat una regla del tipus:

Si codi postal = 08012, llavors compra (margarina)

ha efectuat un canvi cap a un llenguatge més comprensible. El coneixement que els clients que compren margarina viuen en la perifèria de la ciutat i els que compren mantega, al centre, pot ser una relació significativa o no depenent de l'objectiu que ens hàgim proposat per dur a terme el procés de descobriment.

1.1. Què entenem per *coneixement*?

No entrarem en discussions molt profundes sobre 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 coneixement nou. Noteu que en aquesta accepció, el coneixement sempre està sotmès a revisió a la llum de noves informacions.

Podem pensar en el coneixement com a creença justificada. Podríem tenir la creença intuïtiva que la majoria dels clients d'una empresa són de Barcelona, però si les dades ens diuen que 15,000 clients són de Barcelona i 30,000 de la resta de Catalunya, haurem d'interpretar el fet com que la majoria dels clients d'una empresa són de fora de Barcelona. La nostra creença inicial no queda, doncs, justificada, i no és un coneixement vàlid. Les opinions i els prejudicis solen estar equivocats si no es basen en dades reals.

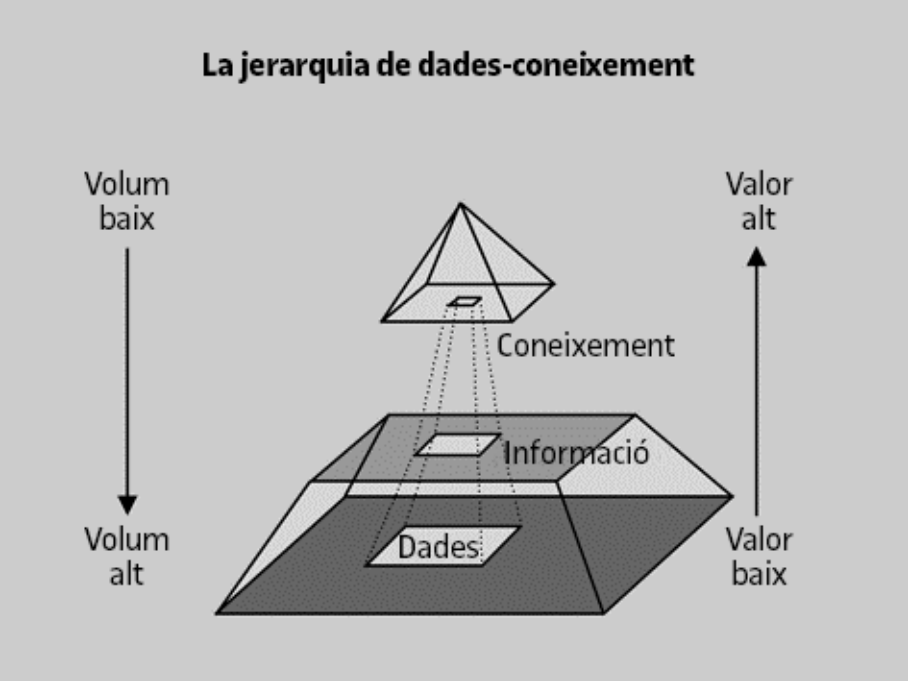
Només podem arribar a una justificació de la nostra creença quan relacionem les dades en brut (el nombre de clients de cada zona) amb una altra dada procedent del nostre coneixement previ (per exemple, que tenim 45,000 clients). Faríem una altra interpretació si el nombre de clients fos 250,000: hauríem de relativitzar la creença, ja contrastada, que la majoria dels clients (30,000) eren de fora de Barcelona.

El que es demana al procés de mineria de dades és que aporti un primer nivell d'interpretació mitjançant l'extracció de relacions amb dades en brut. D'aquest primer nivell se'n pot extreure coneixement encara més elaborat.

Hi ha 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 1 reflecteix molt bé la jerarquia d'importància que es desenvolupa en el procés d'extracció de coneixement.

Figura 1. Piràmide del coneixement



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 expressar el coneixement que volem obtenir són la renda i el tipus d'activitat esportiva:

Taula 1

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*	Pesos	Mitjana	23	H
1	3	Tarda	Pesos	*Stretch*	Mitjana	35	D
2	1	Matí	Pesos	Pesos	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 pesos; dos, estiraments (*stretch*); un, aeròbic, i dos, *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 dir que hi ha una correlació de 0.8 entre la renda i l'activitat principal, cosa que ja és un primer canvi de llenguatge.

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

```
If Act1 is Steps Then
   Renta is Media
   Rule's probability: 0.981.
   The rule exists in 52 records.
   Significance Level: Error probability < 0.2.
```

que ens indica que si l'activitat principal d'un client és *steps*, llavors podem assegurar amb una probabilitat del 98.1 % que la seva renda és mitjana, que aquest aspecte ha estat observat en cinquanta-dos casos originals, i que la significació d'aquesta regla és menor del 20 % (ara no entrem en el significat d'això), cosa que és un nou canvi de llenguatge, i fins i tot més comprensible per a algú amb coneixements d'estadística.

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

2. Les fases del procés d'extracció de coneixement

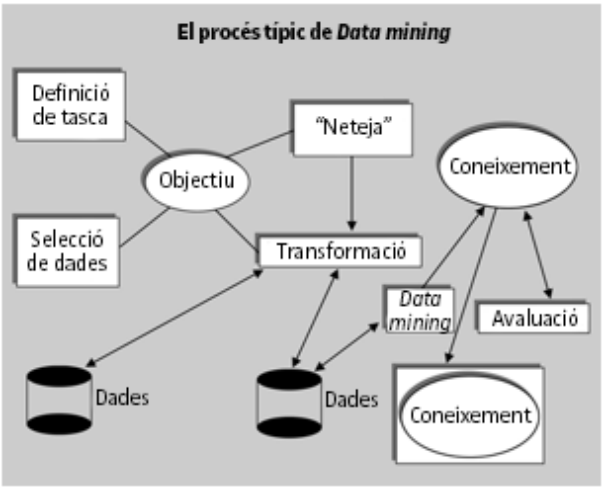
Com hem dit, la mineria de dades no és un procés que es faci en una sola fase. No és que decidim expressar un tipus d'objectiu determinat (predir la continuïtat dels clients, per exemple) i automàticament es generi un model que ens resolgui l'objectiu (que ens digui si un nou client serà fidel durant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

Hi ha diferents formulacions de les fases per dividir el procés. Seguirem Fayyad i esquematitzarem el procés de descobriment segons les fases que expressem gràficament a la figura 2.

Referència bibliogràfica

Trobareu l'article de les fases del procés d'extracció de coneixement en l'obra següent:
U. Fayyad; G. Piatetsky-Shapiro; P. Smyth (1996). «The KDD process for extracting useful knowledge from volumes of data». *Communications of the ACM* (vol. 39, núm. 11, pàg. 27-34).

Figura 2. Procés típic de *data mining*



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 assenyalat abans, 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, per la qual cosa caldrà tornar a extreure coneixement nou.

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. Així doncs, és el moment de decidir si es tracta, per exemple, de trobar dependències entre variables (la renda i l'activitat, per exemple), si volem sa-

ber què distingeix una mena d'usuari d'un altre, si volem conèixer tendències o detectar patrons, etc.

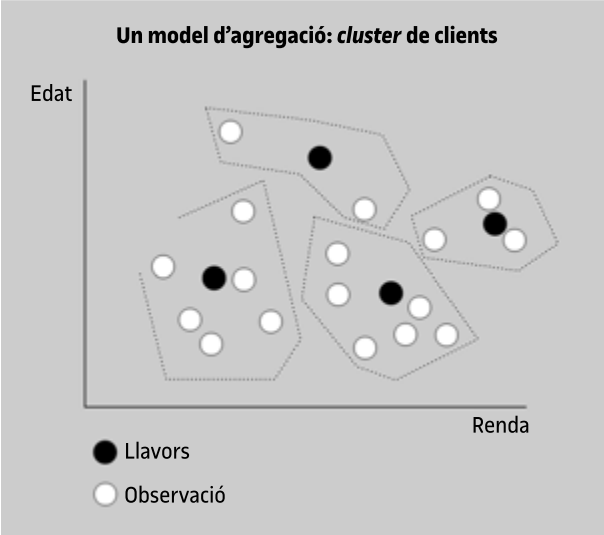
La tasca principal de cadascun d'aquests projectes en general pot assimilar-se a alguna de les següents:

1) Trobar similituds i agrupar objectes semblants. Correspon a un projecte en què tenim «poca» informació del domini i volem començar a tenir una idea més clara sobre aquest tema. Els models típics per aconseguir 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 trobar similituds consisteix a trobar grups de clients semblants.

Figura 3. Model d'agregació: *cluster* de clients



2) Classificar objectes. La tasca d'aquests projectes de mineria de dades no és exactament igual a la del punt anterior. Aquí el més habitual és partir d'una situació més informada, sabent que hi ha grups ja definits. El que ens interessa en aquest cas és estudiar millor les diferències existents entre un grup i un altre, les seves característiques peculiars. Per norma general, la classificació d'objectes és el pas previ a la realització de prediccions; és a dir, per saber alguna conducta d'interès a partir d'una sèrie de dades. Tot això ho podem refinar obtenint coneixement predictiu.

Lectura recomanada

Podeu trobar més informació sobre l'evolució d'aquests models classificatoris a l'article següent:
X. Wu; V. Kumar; J. Ross Quinlan i altres (2008). «Top 10 algorithms in data mining». *Knowledge Information Systems* (vol. 14, pàg. 1-37). https://doi.org/10.1007/s10115-007-0114-2

Alguns models classificatoris típics són els arbres de decisió com ara CART, ID3, C4.5 i C5.0; també es poden esmentar les xarxes neuronals per a classificació i els sistemes basats en regles de classificació.

Els arbres de decisió ofereixen una estructura en què en cada node s'hi 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. Si seguim les branques des de l'arrel fins a les fulles, s'obté una sèrie de condicions que permeten classificar les noves observacions. Per exemple, a la figura 4 hi podem veure l'estructura d'un arbre de decisió l'objectiu de la qual és predir si un client d'un gimnàs sol·licitarà els serveis d'un entrenador personal o no.

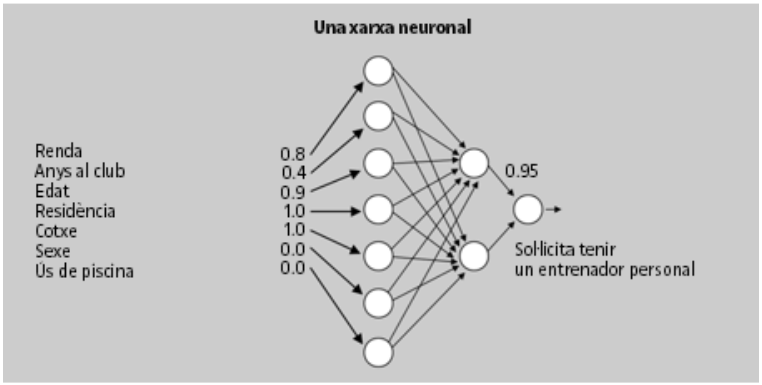
Figura 4. Arbre de decisió

Normalment, els mètodes de construcció d'arbres de decisió venen acompanyats de més informació que permet 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. Aquestes xarxes presenten certes analogies amb la manera com estan connectades les neurones cerebrals i s'organitzen en forma de molts nodes de procés connectats que donen una o més sortides. Les diferents capes de nodes estan connectades entre si amb més o menys força mitjançant uns factors o pesos que indiquen la importància de les sortides que han tingut lloc per a cada node. En conjunt, el que fan és aprendre a ajustar els valors d'aquests pesos amb la finalitat de ser tan predictives com sigui possible. En l'actualitat, la capacitat de còmput disponible ha fet que ressorgeixin amb força, i han donat lloc al que es coneix com *deep learning*.

A la figura 5 hi podem veure una xarxa neuronal senzilla utilitzada també per predir si un client sol·licitarà entrenador personal o no.

Figura 5. Xarxa neuronal



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

Per part seva, les regles de classificació tenen una expressió com aquesta:

$$\text{Antecedent} \Rightarrow \text{Conseqüent}$$

Aquestes regles imposen una sèrie de condicions sobre els valors que prenen els atributs d'entrada amb la finalitat d'indicar a quina classe poden pertànyer. En el nostre exemple, la classe estava determinada mitjançant l'atribut Entrenador personal. Hi ha dues classes, la dels clients que sol·liciten entrenador i la dels que no. Així doncs, la forma que té la regla ens indica sota quines condicions un client sol·licitaria tenir 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, en poden adoptar d'altres. Per exemple, les regles de classificació que s'obtenen amb mètodes com el CN2 o les que s'obtenen a partir de mètodes de lògica inductiva, que són una conjunció de condicions lògiques sobre els valors que poden adoptar els atributs (igualtat, comparació, etc.) i no solen tenir cap indicació respecte al grau de validesa, significació o error. La regla anàloga a l'anterior, que es pot obtenir fent servir CN2, té la forma següent:

```
If (Act1 = Steps) => (Trainer = no)
```

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

Ara bé, de vegades el que ens interessa demanar no és un atribut que adopti valors en un conjunt finit de valors (numèrics o no), com acabem de veure. Per

exemple, pot interessar-nos predir la durada de les demandes que efectua un departament a partir d'altres dades. En aquest cas, estarem intentant 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 entre el valor mínim i el màxim detectats.

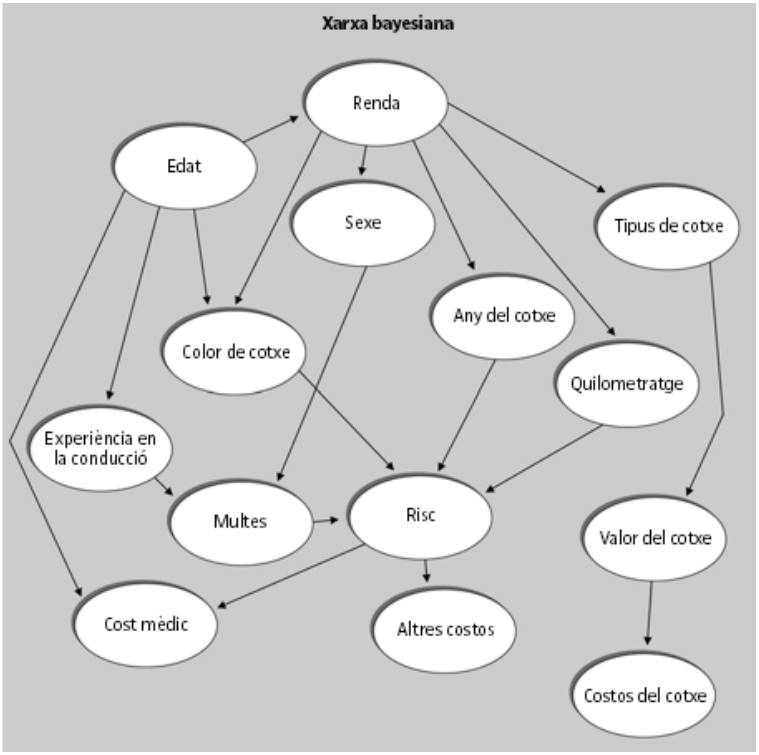
Hi ha molts exemples de models predictius: alguns exemples de models que ens permeten arribar a aquest objectiu són els arbres de decisió i els models predictius clàssics de l'estadística, per exemple, els models de regressió. Aquí hi podríem incloure també els models que no detecten valors concrets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

4) Descriure. Tot i que obtenir agrupacions d'objectes és un primer nivell de descripció, amb *coneixement descriptiu* ens solem referir a trobar i expressar associacions significatives o causals entre diferents variables.

Un exemple típic d'aquesta mena de models és el de les xarxes bayesianes i, malgrat que amb menys potència, les regles d'associació. Les presentem breument a continuació.

Considerem primer una xarxa bayesiana d'exemple, com la que veiem a la figura 6.

Figura 6. Xarxa bayesiana



Aquesta xarxa bayesiana indica coses interessants. Ha estat extreta d'un conjunt de dades simplificades que relaciona diverses variables utilitzades en assegurances per predir si un client és de risc o no. Tan sols inspeccionant visualment el model, ja ens indica que les variables que influeixen més directament en la classe de risc d'un client són el color del cotxe, el quilometratge, les multes i l'any de matriculació. Així mateix, podem veure que altres relacions, com per exemple l'edat del conductor, no influeixen directament en la classe, sinó indirectament per via del color i en combinació amb altres variables, com la renda del client. A més, per a cada enllaç podem saber la probabilitat condicional que hi ha entre les diferents variables connectades. En aquest cas, podem veure els valors de la taula 2.

Taula 2. Taula de probabilitats condicionals

Sexe	Edat	Quilometratge alt	Quilometratge baix
Dona	Jove	0.708	0.292
Dona	Gran	0.582	0.418
Home	Jove	0.714	0.296
Home	Gran	0.255	0.645

Per tant, tenim una idea de la influència mútua entre les variables, quines són realment rellevants per 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 les variables relacionades.

Per part seva, les regles d'associació miren de trobar coocurrències prou significatives entre grups de variables. L'únic requisit 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 ens interessa 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 dona el cas següent:

```
(Renda > 80,000) & (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 80,000 euros, una edat superior als quaranta anys i viu en la zona que hem designat com a C, llavors també resulta que ha demanat entrenador personal i fa servir la piscina.

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 l'edat, el color del cotxe que compri, l'any de matriculació i el valor dels atestats presentats l'últim any. A més, ens indicarà el tipus o nivell de força d'associació existent entre totes aquestes variables. Els models descriptius són molt fàcils d'interpretar si es presenten en un llenguatge pròxim a l'expert en l'àmbit (en aquest cas, les assegurances).

5) Explicar. Aquí es tracta d'obtenir models que puguin donar-nos les raons de per què s'ha produït un comportament determinat. Per exemple, si havíem predit que un client determinat adquiriria una sèrie de productes determinats i no ho ha fet, a què es pot deure? Exemples de models d'aquesta mena són les xarxes bayesianes, que permeten determinar 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.

Vegem algunes consideracions globals per a tota mena de tasques de mineria de dades. És important que tinguem clar que per a cadascuna d'aquestes tasques es pot fer servir més d'un model. A més, per construir cada mena de model tenim a la nostra disposició diversos mètodes que estaran inclosos, o no, en les eines comercials existents en el mercat, i cadascuna d'aquestes es pot avenir, o no, amb els condicionants tècnics del sistema de base de dades del qual disposem i del sistema d'informació en què se situa. Per exemple, les xarxes bayesianes es poden emprar com a models descriptius, predictius i explicatius.

Lectura recomanada

Trobareu més informació sobre els models explicatius a 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.: IEEE Press.
Un treball més recent sobre regles és el següent:
M. Almutairi; F. Stahl; M. Bramer (2017). «Improving Modular Classification Rule Induction with G-Prism Using Dynamic Rule Term Boundaries». A: M. Bramer; M. Petridis (eds.). *Artificial Intelligence XXXIV*. SGAI 2017. Springer, Cham («Lecture Notes in Computer Science», 10630). https://doi.org/10.1007/978-3-319-71078-5_9

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

- 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, per exemple, que hem de conèixer millor els nostres clients és un objectiu massa general que necessita precisar-lo. 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. No obstant això, sigui com sigui, sempre queden objectius que requereixen un esforç combinat. Potser primer haurem d'agrupar, després classificar i més tard extreure un model predictiu.

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

- Seleccionar el mètode necessari per construir-lo. Aquest tercer punt està molt relacionat amb el segon i ens permetrà avaluar les possibilitats i les característiques de les diferents eines existents: complexitat del mètode i cost computacional.

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

Una vegada definit l'objectiu, i quan l'hem posat en relació amb la tasca principal del projecte, quins models ens interessen més i quins mètodes i eines ens fan falta, hem de passar a trobar la matèria primera: les dades. No és tan senzill com sembla.

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

Origen de les dades

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

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

Amb aquesta tecnologia s'assegura, per exemple, que per als diferents registres el camp nivell de risc del client té el mateix significat, tot i que ha resultat de fusionar les diferents interpretacions que li dona 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 predir tendències. En principi, són una tecnologia dirigida a la presa de decisions. Llavors, quin problema presenten? Doncs que la majoria de les empreses no tenen aquesta tecnologia instal·lada! I aquesta afirmació es fa més certa com més petita és l'empresa. En conseqüència, el més normal és que, com a primera fase, calgui localitzar les fonts de dades a partir de:

- Les bases de dades disperses pels diferents departaments que considerem que poden ser rellevants per al projecte que duem a terme.

- Les bases de dades transaccionals, que registren les operacions dia a dia i que acumulen informació històrica que pot ser rellevant per obtenir el model que estem buscant.

Cap d'aquestes dues operacions és senzilla. De fet, cal que l'empresa –si no té un *data warehouse*– faci una bona política de gestió de dades de cara a la presa

de decisions. És molt probable que calgui crear i introduir nous processos per obtenir les dades que necessitem.

Exemple sobre la 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 cadascun més enllà d'un cert nombre de mesos, per la qual cosa va haver d'introduir un nou procés i modificar la base de dades corresponent per disposar de dades anteriors en el temps, necessàries per poder extreure coneixement més fiable i analitzar tendències i patrons estacionals.

2.2. Preparació de les dades

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

Una vegada localitzades les fonts de dades, hem de procedir a preparar-les perquè se'ls puguin aplicar els mètodes o les eines que construiran el model desitjat. Aquesta fase, encara que sembli senzilla, juntament amb la de selecció de dades, consumeix el 70 % de l'esforç (o més!) en els projectes de mineria de dades de nova implantació.

En aquest punt, cal assegurar-se d'unes quantes coses. Vegem-les:

1) Que les dades tinguin prou qualitat. És a dir, que no continguin errors, redundàncies o que presentin altres menes de problemes. També s'entén la qualitat de les dades com aquella propietat que asseguri la qualitat del model resultant (per exemple, que asseguri que serà prou predictiu, si es tracta de crear un model d'aquesta mena). No cal dir que aquesta última accepció de la paraula *qualitat* és encara més problemàtica de garantir.

2) Que les dades siguin les necessàries. Potser n'hi haurà que no ens caldran, i potser n'hem d'afegir d'altres. Sol ser molt estrany que les dades que necessitem realment ja hagin estat recollides pel sistema amb el propòsit de dur a terme justament la mena d'estudi de mineria de dades que volem emprendre. Això, normalment, suposarà afegir camps nous a les diferents relacions d'una base de dades procedents d'altres relacions o d'altres bases de dades.

3) 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 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.

Una pràctica comuna és generar noves variables derivades d'expressions més o menys complexes a partir de les dades originals. Això permet originar indicadors que en si mateixos ja agregaran informació i seran molt útils en el procés de conèixer la informació que guarden les dades i en el procés posterior de creació del model.

Lectura recomanada

Trobareu més informació sobre la fase de preparació de dades a 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.
Un treball més recent és el següent:
S. García; J. Luengo; F. Herrera (2015). *Data preprocessing in data mining* (vol. 72). Cham, Suïssa: Springer International Publishing.

2.2.1. Neteja de les dades

La neteja de dades consisteix a processar les dades per eliminar les que siguin errònies o redundants. També hi ha qui inclou en aquesta fase el pas consistent a eliminar alguns dels atributs de les dades, cosa que es coneix com a *selecció d'atributs*, i la intenció dels quals és assegurar que amb menys dades es puguin obtenir models de la mateixa qualitat. Comentarem aquest aspecte més endavant.

Fins i tot estant en la forma adequada, sol passar que les dades no són perfectes en un 100 %. Les dades introduïdes a mà o procedents de la fusió de diverses bases de dades solen mostrar factors de distorsió importants. Revisem quins són els més freqüents:

1) Dades incompletes. Pot passar, especialment en aquells atributs en què quan es va dissenyar el procés corresponent d'entrada de dades es va decidir que no eren obligatòries o que tenien format lliure, 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 estem considerant. En anglès això es coneix com a *missing values*.

Normalment, el que es fa és «completar» els valors que no apareixen. Per exemple, per als valors numèrics es complementa calculant el valor mitjà observat per a un atribut, tot i que també és possible fer servir tècniques d'imputació més sofisticades.

Exemple de dades incompletes

Un exemple de dades incompletes és el que tenim quan emplenem el camp corresponent al carrer, però oblidem, o no es representa correctament, el corresponent al número o al pis. Una manera de solucionar les dades que falten consisteix a substituir-les per un «valor raonable», tot i que això dependrà del context, òbviament, i no sempre serà possible.

Per exemple, completar valors numèrics: si, per exemple, ens falta informació sobre el salari d'un treballador, podem adscriure-li la mitjana dels valors dels salaris de la resta dels treballadors de la companyia o de la seva secció. No cal dir que aquest mètode presenta alguns problemes i pot induir a errors o rebaixar la qualitat del model resultant. Així i tot, a vegades és tot el que es pot fer.

2) Dades redundants. Algun cop es repeteixen tuples que corresponen al mateix objecte.

Exemple de repetició de tuples

Sovint ens trobem amb 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 de dades diferents. Una variació són aquells 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 adquireix els productes dins d'una cadena de botigues.

Taula 3

Identificador	Nom	Direcció	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 sigui la mateixa persona. Necessitem eines per decidir quina de les dues interpretacions és la correcta, o si es tracta d'un problema d'assignació d'identificadors. Aquest és un procés difícil i moltes vegades cal fer servir eines estadístiques només per poder detectar-ho. És molt normal, no obstant això, 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, els que volen «despistar» o ocultar els diferents canvis de domicili.

3) Dades incorrectes o inconsistents. Cas molt comú quan el tipus de valors que pot rebre un atribut no està controlat perquè ha estat declarat com a «text lliure», o bé està definit com un tipus determinat (cadena alfanumèrica, posem per cas), però no s'han mantingut els processos de control d'errors necessaris. Per exemple, un client amb una edat superior a cinquanta anys que rebi descomptes per Carnet Jove; o el 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

Un exemple de dades inconsistents és el cas següent, si suposem que només hi ha deu botigues en una determinada cadena comercial, és evident que passa una cosa estranya amb el client «Martín».

Taula 4

Identificador	Nom	Direcció	Centre
24,567	Poch	Roca, 33-1	1
32,456	Martínez	Travessera, 222	2
33,345	Martín	Roca, 33-1	144
33,400	Sala	Diagonal, 556	1
33,441	Arregui	Diagonal, 222	2

4) Errors de transcripció. Molt típics i que poden donar lloc a algun dels problemes anteriors. Per exemple, majúscules/minúscules, accents i altres caràcters especials, etc.

Exemple típic d'error de transcripció

El mètode que explori les dades pot decidir que «Barcelona» i «BARCELONA» són dues poblacions diferents. I ja no diguem si s'ha introduït «BCN».

5) Dades envellides. Certes dades es converteixen en incorrectes perquè no han estat actualitzades de la manera adequada. Aquest cas pot ser molt complex de detectar perquè depèn molt del context.

Vegem uns exemples de dades envellides: un cas típic d'aquesta categoria és el domicili o la domiciliació bancària quan no es notifiquen els canvis corresponents. Un altre cas podria donar-se quan es tracta de treballar amb rangs d'edats, i a les dades cada persona apareix amb el rang d'edat corresponent. Suposem, per exemple, que en lloc de guardar la data de naixement es guarda l'edat del client quan es dona d'alta. Si no hi ha un procediment d'actualització 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 2014 tenien cinquanta-nou anys, l'any 2020 en tenen o en tindran seixanta-cinc. Han passat de la categoria de clients «veterans» a «jubilats», però a la base de dades es continuen considerant clients de la primera categoria.

6) Variacions a les referències als mateixos conceptes. Per exemple, un advocat pot ser considerat com a «professional liberal», mentre que un altre client que també ho sigui pot estar categoritzat com a «autònom». És més probable que això passi si la mateixa informació es guarda en bases de dades diferents.

Exemple de variacions a 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 què 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 ràpida fusió dels sistemes d'informació de totes dues empreses va provocar, entre altres conseqüències, que durant un llarg període de temps les dades dels clients comuns es guardessin per duplicat en dues bases de dades diferents, en què, a més, alguns atributs, com el de professió, adoptaven valors de conjunts diferents. Aquí ho teniu. Fixeu-vos en els clients «Martínez» i «Sala»: sabem que tots dos són professors universitaris. No obstant això, per al banc «Martínez» és professor i per a l'asseguradora «Sala» era mestre. Com que el banc també recollia la categoria professional Mestre entre les que podien ser assignades als seus clients, va passar el que va passar...

Taula 5

Identificador	Nom	Professió	Risc
24,567	Poch	Advocat	Tot risc
32,456	Martínez	Professor	Tercers
33,345	Martín	Construcció	Tot risc
33,400	Sala	Mestre	Tot risc
33,441	Arregui	Cuiner	Tercers

Aquí tenim les dades corresponents als tipus de préstec sol·licitats per cada client i la quantitat corresponent. Què li passa a «Martínez»? I a «Sala»? Quina és la seva veritable professió? Com ho detectem?

Taula 6

Identificador	Nom	Professió	Préstec	Muntant	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	Martín	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

Doncs bé, per a aquesta mena de problemes és per als quals els *data warehousing* i altres sistemes de gestió de bases de dades intenten aportar solucions.

7) **Dades esbiaixades**. Aquesta mena de problema pot donar-se amb dades que compleixen tots els altres requisits de qualitat esmentats fins al moment. Es tracta d'aquells tipus 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'aquesta mena de conjunts de dades.

Exemple de dades esbiaixades

És possible que hàgim triat sense adonar-nos un conjunt de clients que són majoritàriament joves o de determinada mena de professió. Segons quin sigui l'objectiu del nostre estudi, pot no interessar aquest tipus de biaix. Els models generalment intenten explicar o predir pensant en les majories, i poden ser molt incorrectes per a les categories que no s'han tingut en compte. A més, els biaixos en les dades i els models construïts plantegen problemes ètics i fins i tot legals quan afecten persones en funció del seu gènere, ètnia, etc. Hi ha mecanismes per intentar balancejar conjunts de dades que presenten un biaix, malgrat que seria millor no tenir-lo present d'origen.

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

2.2.2. Transformació de les dades

Les dades no sempre estan en la forma més adequada per poder aplicar els mètodes necessaris per a la tasca que cal dur a terme i el model que es vol. En general, ens trobarem que haurem de fer 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 bastant bé com a Gran o Jove perquè són aquests els grups d'edat que interessa distingir i estudiar. Ara bé, pot passar que a la base de dades o en diverses bases de dades utilitzades aquest camp tingui un valor numèric (edat entre els valors numèrics de divuit a cent anys, per exemple). La solució consisteix a assignar una categoria a cada rang de valors que necessitem, fixant bé una correspondència entre

els valors numèrics i la categoria (per exemple, categoria Gran pot correspondre als valors més grans de seixanta anys), o bé intercedint automàticament procediments de discretització.

2) Dades categòriques a numèriques. Disposem de dades que apareixen descrites mitjançant valors categòrics, i el que necessitem realment és disposar dels valors numèrics corresponents. Hem d'efectuar el procés invers, adscrivint una traducció a cada categoria en el conjunt corresponent de valors numèrics. Per exemple, fent que la categoria Jove de l'atribut Edat equivalgui al rang de valors 18-25 anys. El problema és que per cada aparició a la base de dades, potser no podríem posar un interval, sinó un valor únic. En aquest cas, cal efectuar noves transformacions.

3) Altres transformacions. Moltes vegades, per simplificar la representació cal efectuar una altra mena de transformacions, ja sigui d'escala, o d'unitats (com en el cas de quantitats relatives a diners en moneda de diferents països). Per exemple:

a) Simplificació de valors: dividir els sous per mil o un milió.

b) Agrupació de valors continus en franges: totes les compres entre les vuit i les deu corresponen al valor 1; les de deu a dotze, al valor 2, etc.

c) Normalització de dades: posar els valors numèrics en un interval determinat. Per exemple, moltes xarxes neuronals i algorismes d'agrupació obliguen (o prefereixen) que els valors numèrics estiguin entre 0.0 i 1.0.

d) Addició d'una etiqueta que indiqui 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 (cosa que es pot haver derivat de l'experiència o bé a partir d'un altre mètode de mineria de dades).

e) Expansió d'un atribut: pel fet que el valor d'un atribut pot adoptar valors en un conjunt limitat de categories. Per exemple, l'atribut Risc d'incendi pot prendre valors en les categories Alt, Baix i Mitjà; i pel fet que calgui expressar les dades en forma numèrica podem disgregar l'atribut Risc d'incendi en els atributs Risc-alt, Risc-mitjà i Risc-baix, cadascun dels quals pot prendre el valor 0 o 1 indicant l'existència o no de cada tipus de risc. En anglès, això és el que es coneix com a *dummy variables*.

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

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

Taula 7

Identificador	Nom	Professió	Préstec	Muntant	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	Martín	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:

Taula 8

Identificador	Nom	Professió	Personal	Hipoteca	Muntant	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	Martín	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 necessitem sigui un conjunt diferent. Comentem a continuació els problemes i les solucions més habituals:

1) Derivació de dades: podem fer servir els atributs de les dades existents per 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 s'estigui duent a terme.

Exemple de derivació de dades

Típicament, es pot derivar l'atribut Edat de la diferència existent entre la data actual i la data de naixement declarada. O, per a un estudi mèdic sobre l'obesitat, pot passar que disposem de les dades següents: edat, altura (en m), pes (en kg), sexe i professió.

Taula 9

Identificador	Altura	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

No obstant això, ens interessa obtenir l'índex de massa corporal (IMC), que correspon a aquesta senzilla fórmula:

$$IMC = \frac{P}{A^2}$$

en què P és el pes en quilograms i A és l'altura en metres.

Taula 10

Identificador	Altura	Pes	Sexe	Professió	IMC
24,567	1.90	88	Dona	Advocat	24.4
32,456	1.85	92	Home	Mestre	26.9
33,345	1.78	73	Home	Construcció	23.0
33,400	1.70	65	Dona	Representant	22.5
33,441	1.78	110	Home	Cuiner	34.7

De fet, quan es fa 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'altura. Normalment, no obstant això, les derivacions solen ser més complexes, involucrar més d'un atribut i generar també més d'un resultat.

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

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

Exemple d'enriquiment

Podem afegir a la informació que tenim dels nostres clients el resultat d'una enquesta en què els haguéssim preguntat quin cotxe voldrien comprar i si pensen canviar de cotxe l'any que ve.

Aquí tenim la taula de clients original:

Taula 11

Identificador	Nom	Professió	Préstec	Muntant	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	Martín	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

A continuació, el que van contestar a una enquesta telefònica:

Taula 12

Identificador	Tipus de cotxe	Any que ve
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, operació que en una base de dades relacional es pot fer amb una operació de Join, obtenim:

Taula 13

Identificador	Nom	Professió	Préstec	Muntant	Saldo actual	Saldo mitjà	Tipus de cotxe	Any que ve
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	Martín	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

Llavors, podem respondre a preguntes com ara qui ens demanarà un crèdit l'any que ve?

2.2.3. Reducció de la dimensionalitat

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 el dona tant la quantitat de registres que té com el nombre d'atributs que es manegen. 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. És l'anomenada *maledicció de la dimensionalitat*. Per tant, si és possible treballar amb menys dades i obtenir els mateixos resultats, seria millor des d'un punt de vista d'eficiència.

Els mètodes de reducció de dimensionalitat volen justament treballar amb menys dades i obtenir els mateixos resultats. A continuació, presentem breument les tècniques habituals de reducció de la dimensionalitat.

1) Reducció del nombre de registres per tractar. La reducció del nombre de registres per tractar consisteix a trobar un conjunt de dades de dimensions més petites per construir el tipus de model que necessitem amb el nivell de qualitat necessari.

L'estadística ha desenvolupat eines per triar prou conjunts de dades de cara a la construcció de models. Així doncs, és bo recórrer a les seves tècniques per obtenir un conjunt més reduït, però igualment potent, de dades inicials. Els problemes que hem d'evitar aquí són principalment que el conjunt triat no sigui massa esbiaixat cap a un conjunt d'objectes amb característiques molt concretes i poc representatives, com ja hem comentat abans. També passa que, si el conjunt és massa reduït, les conclusions que s'extrauran del model resultant final no seran prou significatives. Per tant, de vegades és necessari conservar un conjunt alt de registres per mantenir prou qualitat.

Una diferència que cal destacar de la manera com 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 poden ser els que més interessen. En conseqüència, els mètodes per reduir el nombre de registres que s'utilitzarien en l'anàlisi de dades tradicional i en mineria de dades són lleugerament diferents.

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

2) Reducció del nombre d'atributs per tractar. La reducció del nombre d'atributs per tractar també rep el nom de *selecció d'atributs*.

Aquest fet representa haver de detectar atributs irrellevants que manquen d'efecte sobre la qualitat del model final (és a dir, que el model ens permet contestar les mateixes preguntes amb aquests atributs o sense), i atributs o combinacions d'atributs equivalents (atributs que permeten 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 introduirem de la manera adequada amb un parell d'exemples.

Exemple

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

Taula 14

Identificador	Altura	Pes	Sexe	Professió	IMC
24,567	1.90	88	Dona	Advocat	24.4
32,456	1.85	92	Home	Mestre	26.9
33,345	1.78	73	Home	Construcció	23.0
33,400	1.70	65	Dona	Representant	22.5
33,441	1.78	110	Home	Cuiner	34.7

Com hem vist, sembla raonable pensar que la combinació d'atributs (Altura, Pes) és equivalent a l'IMC, ja que aquest últim atribut es calcula a partir dels altres dos. Podem dir, per tant, que la combinació d'atributs (Altura, Pes) aporta la mateixa informació que l'atribut IMC, amb la qual cosa podem substituir els dos atributs Altura i Pes per l'atribut IMC amb una senzilla operació SELECT típica de les bases de dades relacionals.

Taula 15

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 a fi de trobar aquesta mena d'equivalències. Cada atribut menys pot representar milions de bytes necessaris per representar el conjunt de dades.

Exemple

Ara veurem 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 només pot pertànyer a una sola classe. La classe 0 és la dels clients de poca morositat; la classe 1, la d'alt risc de morositat. No és necessari que ens preocupem ara de com s'ha obtingut aquesta classificació. A més, hem introduït alguns canvis en els valors per deixar més clar què volem dir:

Taula 16

Ident.	Nom	Professió	Préstec	Muntant	Saldo actual	Saldo mitjà	Tipus de cotxe	Any pròxim	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
33,345	Martín	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, fins i tot semblant interessants, ens presenten un problema. Tothom que té un préstec personal ha respost afirmativament la pregunta de si es comprarà un cotxe l'any que ve; tothom que té una hipoteca l'ha respost negativament. Per tant, aquests dos atributs són molt poc informatius. Evidentment, si hi hagués més tipus de préstecs, això no seria així, tot i que 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 la mena de préstec i la intenció de compra, cosa que ja és bastant significativa. Si volem agrupar els clients per la similitud que tenen, potser aquests dos atributs no ens fan cap falta.

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

A la fase de mineria de dades tenim les dades amb la qualitat adequada, en el format adequat i hem seleccionat els atributs i els registres aparentment necessaris i rellevants. Tenim decidit quina mena de model volem obtenir. Per tant, ara cal triar un mètode de construcció de models entre la multitud de mètodes que permeten obtenir el model que ens interessa.

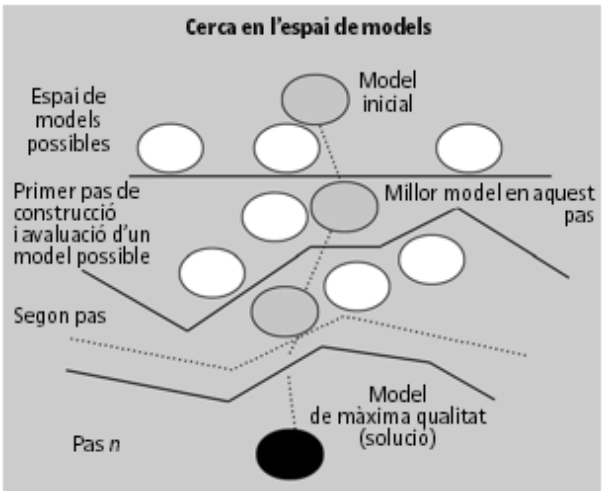
De fet, per a una mateixa tasca, ens poden ser útils diversos models. Igual que l'aparent diversitat de models ens pot ocultar les similituds amb la tasca per a la qual s'ha d'aplicar, tampoc podem deixar de remarcar la gran similitud que hi ha en el procés de construcció dels diferents tipus de models.

En efecte, el procés de construcció de models consisteix a trobar el model (el coneixement) que respon millor a les característiques implícites en les dades. Aquest tipus de problema se sol conceptualitzar com un procés de cerca.

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

Podem representar el procés de cerca amb l'esquema de la figura 7.

Figura 7. Cerca en l'espai de models



2.3.1. Mecànica general del procés de cerca

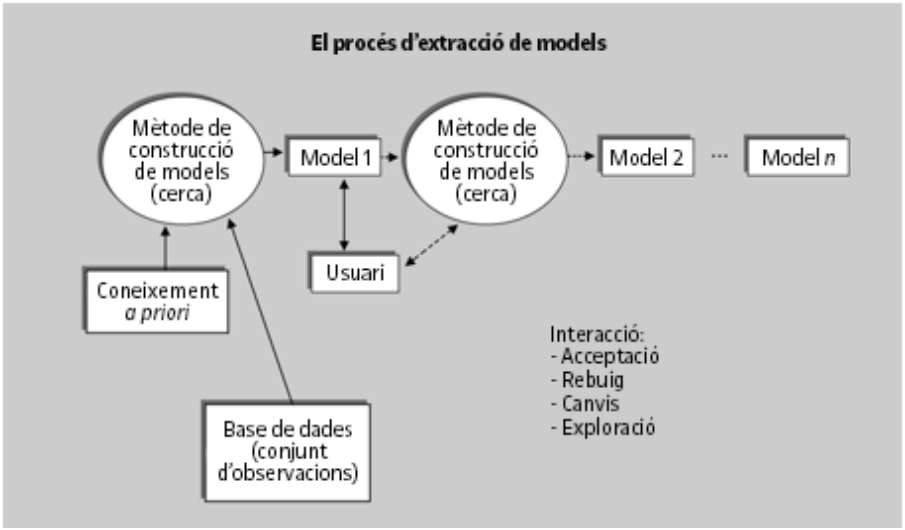
En què consisteix un procés de cerca, llavors? Tot procés de cerca parteix d'un model inicial (que pot ser el model buit, inexistent) i a cada pas modifica aquest model mitjançant un conjunt d'operadors de modificació de models. A cada pas és possible aplicar més d'un operador, i també es poden generar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

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

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

Com que la major part de les vegades el conjunt de combinacions possibles és tan gran que és impossible generar-les totes en un temps raonable, cal triar encertadament a cada pas un subconjunt de models parcials. Així, seleccionarem models que sembli que asseguren un model final. Per saber quin és el millor model, farem servir alguna mena de funció d'avaluació, que dona a un model parcial un valor més alt com més probable sigui que es trobi en el camí cap a un model final bo, de qualitat alta.

Figura 8. El procés d'extracció de models



No cal amagar que els problemes principals d'aquesta mena de conceptualització són els següents:

1) Disposar d'una mesura de qualitat que qualifica com el millor possible un model que només és relativament bo respecte als models que té prop seu, però que no és el model òptim dins de tot l'espai de solucions possibles. Aquest problema es denomina *problema de l'obtenció de mínims locals*. Seria com triar la poma més gran d'un fruiter, sense tenir en compte que pot haver-hi pomes encara més grans en altres fruiters.

2) Obtenir un model que sigui bo tenint en compte només les dades de què disposem. Per posar un exemple, si estem modelitzant el comportament dels clients d'una empresa i disposem d'una base de dades de tres mil clients per fer proves, de la qual obtenim un model classificatori, no volem que, en veure el cas tres mil un, el model s'equivoqui i l'assigni a 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 que consisteix a obtenir models que generalitzen malament i són massa específics respecte a les dades que han vist es coneix com el *problema de la sobreespecialització* (en anglès, *overfitting*).

3) En el transcurs del temps és possible que el model obtingut es degeneri perquè comencen a aparèixer observacions noves per a les quals no generalitza bé. Aquest és el problema de l'envelliment de models. Cal actualitzar-los, ja sigui des de zero o incrementalment, depenent de la mena de model de què es tracti.

Veiem, llavors, 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. És a dir, com es van construint els models.

- La funció d'avaluació. Avaluen que el model que s'està construint és millor o pitjor, i permeten comparar models.

2.3.2. Varietat de models de cerca

Pel que acabem de dir, és 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. Bé, això és veritat només fins a 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 del procés de cerca pugui avaluar com ha anat funcionant el procés i fer que la cerca adopti una certa direcció.

En el fons, es pot considerar que un procés de cerca parteix de més informació que les dades. Vegem-ho esquematitzat a la figura 8. Els elements per considerar en aquest esquema són els següents:

1) Base de dades. Conjunt d'observacions o exemples de què disposem.

2) Coneixement *a priori*. Tot aquell coneixement de què es disposa i es vol fer servir. Per exemple, saber que determinades pautes de conducta sempre estan associades. En el cas del banc de què hem estat parlant, 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*, què hem de tenir en compte? Les dades o l'usuari humà que ha de donar aquesta indicació inicial? Meitat i meitat? Com alterem el model que s'està construint amb la finalitat d'equilibrar la influència de les dues fonts d'informació, l'usuari i les dades?

Com podem expressar aquest coneixement *a priori*? Aquí exposem algunes de les tècniques d'expressió d'aquest coneixement més típiques utilitzades en alguns dels mètodes, tot i que no sempre aplicables a tots:

- Distribució *a priori* de la probabilitat dels valors de les dades: normalment, és algun paràmetre que permet 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 determinar els paràmetres que creiem importants.

- Relacions entre les dades: indicar que considerem que dos atributs han de ser tractats com a dependents, per exemple. 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 es trobin en el mateix grup d'observacions final.

- Relacions de coocurrència: semblants a les de relacions de dependència que hem comentat abans. Per exemple, «sempre que algú compra cervesa, també compra cacauets».

3) 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 per una mesura de qualitat que és una combinació del grau amb què el model s'ajusta a les dades i les preferències de l'usuari. Això no vol dir que el procés de cerca procedeixi automàticament fins a trobar un model de prou qualitat. L'usuari pot decidir detenir el procés de cerca i retrocedir fins a una situació anterior en què comença a explorar introduint una altra informació. Aquest era el sistema seguit, per exemple, per Data Surveyor.

Lectura recomanada

Trobareu informació sobre la proposta original de Data Surveyor a l'obra següent: M. Holsheimer; A. Siebes (1994, gener). *Data mining: the Search for Knowledge in Databases*. Report tècnic CS-R9406. Amsterdam: Centrum voor Winskunde en Informàtica (CWI).

Segons aquest esquema, doncs, podem dividir els mètodes de mineria de dades en el que concerneix coneixement *a priori*, mena de dades i procés de construcció, en els termes que anotem a continuació:

1) Quant a coneixement *a priori*:

- Tipus de coneixement *a priori* que permeten expressar.

- Llenguatge en què permeten expressar aquest coneixement.

- Reutilització del coneixement extret per un altre sistema de mineria de dades amb la finalitat de guiar el procés de cerca. En anglès, *multistrategy learning*.

2) Pel que fa a la mena de dades:

- Mètodes que només fan servir observacions, és a dir, només informació respecte als valors que s'han recollit, sense més informació afegida. Els mètodes que parteixen d'aquesta base es coneixen com a *mètodes no supervisats*. Un exemple típic d'això són els mètodes d'agregació. En anglès, *clustering*.

- Mètodes que afegeixen informació discriminant les observacions. Per exemple, a cada observació se li assigna el nom o l'indicatiu de la classe a què pertany. Parlem pròpiament, llavors, d'exemples i contraexemples, més que d'observacions. Els mètodes que parteixen d'aquesta situació inicial es denominen *mètodes supervisats*. Un exemple típic d'això és la classificació.

3) En el que concerneix el procés de construcció:

- Mètodes *batch*. Mètodes que estan pensats per utilitzar totes les dades existents com a únic conjunt de dades per obtenir un model «en un sol pas».

- Mètodes incrementals. Pensats per 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. Útils per anar modificant els models a mesura que van canviant les dades en el temps (envelliment de dades o deriva de models).

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

Lectura recomanada

El treball següent descriu aspectes metodològics sobre el procés de mineria de dades:
A. Feelders; H. Daniels; M. Holsheimer (2000). «Methodological and practical aspects of data mining». *Information and Management* (vol. 37, núm. 5, pàg. 271-281).

2.3.3. 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 estàvem estudiant. Però quina qualitat presenta 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 tota mena de models, ja que, com vam dir, cada model es dirigeix a un objectiu diferent.

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 es fa servir per construir el model i un altre, per avaluar-lo. De vegades, s'introdueix un tercer conjunt entre aquests dos: el conjunt de validació. Utilitzem un conjunt de dades per construir el model; un altre, per donar-lo per bo (validar-lo), i un tercer, per avaluar-lo. Normalment, aquests tres conjunts, fins i tot reflectint informacions sobre els mateixos atributs, procedeixen de conjunts de dades diferents.

Amb tot, hem de tenir en compte que no es poden comparar models predictius amb models descriptius. Per als primers, hem d'emprar mesures que ens indiquin fins a quin punt són bons fent prediccions. Per als segons, hem de mesurar fins a quin punt s'ajusten al domini descrit.

Però quines mesures es fan servir més sovint per avaluar la qualitat del model? I com es pot efectuar l'avaluació d'una manera més metòdica que la que acabem d'apuntar?

Vegem un exemple de procés d'avaluació: presentarem un cas del món de les assegurances. Si volem extreure un model predictiu que ens indiqui 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 –quins sabem que són de risc i quins no– en dues bases de dades: una per construir el model i una altra per validar-lo.

Sobre la primera sèrie de dades, apliquem un mètode de predicció (per exemple, una classificació que ens indiqui si un client pertany a la classe de risc o a la de no risc), i a partir d'aquí extraiem les combinacions de valors que prediuen 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 permís de conduir amb més de dos anys d'antiguitat són els més propensos a pertànyer a la classe de risc. Podem emprar aquesta regla de predicció sobre el conjunt de dades d'avaluació per saber si aquest «minimodel» és correcte:

- Si amb la citada regla es classifica correctament un percentatge de clients significatiu (posem el 95 %), llavors podem considerar que tenim un model de bona qualitat. És a dir, si el model classifica correctament clients que el mètode no havia utilitzat per construir-lo. Doncs bé, això voldrà 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 classificats com no risc, quan en realitat sí que són de risc) és alta, llavors tenim un model predictiu de baixa qualitat.

En aquest últim 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 esmentar una possible mesura de qualitat: l'error en la predicció.

Encara ens queda la interpretació final. Quan ja 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 el significat del coneixement que ens està mostrant. És aquí on correm el risc de les falses interpretacions.

Un exemple molt senzill de falses interpretacions és el d'aquell sistema de predicció que va determinar que el factor més important per calcular si una persona resident a Londres podia quedar-se embarassada era el sexe, ja que en

el 99.99 % dels casos les persones que es quedaven embarassades eren dones, mentre que el 0.01 % restant no s'identificava com a tals, indicant que una variable binària potser no és adequada.

Són nombrosos els exemples de models que donen conclusions evidents, i cal estar a l'aguait en aquesta fase d'interpretació per no recollir com una gran troballa alguna cosa que ja se sap. D'aquí la importància dels mètodes que permeten integrar coneixement *a priori*, especialment coneixement negatiu, del que s'està segur que no pot passar o que no és rellevant. No obstant això, hi ha problemes que no són evidents.

Vegem un altre exemple, conegut com 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 (el Canadà) entre 1981 i 1990.

Wen es va concentrar en els casos de pacients sotmesos a una colecistectomia primària oberta. Alguns d'aquests pacients també havien estat sotmesos a una apendicectomia en el procés de colecistectomia, cosa que es coneix com una *apendicectomia incidental o discrecional*.

En la taula següent podem veure els resultats que reflecteixen les morts que van tenir lloc a l'hospital comparant els pacients que havien sofert apendicectomia durant l'operació de colecistectomia primària oberta i els que no:

Taula 17. Cas de l'apendicectomia beneficiosa

	Amb apendicectomia	Sense apendicectomia
Percentatge de morts 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 per comparar els resultats dels dos grups i esbrinar si mostraven una diferència significativa. Es van trobar que, segons el test, la diferència era, efectivament, significativa.

Aquest «descobriment» del coneixement del fet que una apendicectomia incidental durant l'operació de colecistectomia pot «millorar» les probabilitats de sobreviure cal prendre-se'l amb una mica de calma. Com és possible que una apendicectomia incidental pugui millorar els resultats?

Wen va considerar per separat un grup de pacients de baix risc. Aquest grup de pacients, en canvi, mostrava que l'efecte de l'apendicectomia discrecional presentava resultats bastant insatisfactoris. Paradoxalment, podria ser que l'apendicectomia casual afectés negativament tant els pacients de baix risc com els d'alt risc, però que, considerats junts, fes l'efecte d'un efecte positiu. Això es coneix com a *paradoxa de Simpson*, i cal estar-hi molt a l'aguait. Vegem com acaba de funcionar.

A la taula següent es mostren dades fictícies que permetrien interpretar aquesta paradoxa:

Taula 18. Divisió en pacients de baix risc 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

A la taula següent hi trobarem les proporcions corresponents a les morts dins de l'hospital classificades com a apendicectomia incidental i pacients de risc corresponents amb les dades de la taula anterior:

Taula 19. Efecte combinat

	Amb apendicectomia	Sense apendicectomia
Baix risc	0.0009	0.0006
Alt risc	0.1000	0.0500
Combinat	0.003	0.0070

Podem dir que les categories de risc i les morts estan altament correlacionades. Era més probable que les apendicectomies anessin aplicades a pacients de baix risc que 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, llavors podem dir que és més probable que pertanyi a la categoria de «Baix risc» (joves). Ara bé, aquest fet no implica de cap de les maneres que passar per una apendicectomia disminueixi el risc d'alguns pacients. Si la informació sobre el risc no apareix a la taula, es pot extreure aquesta conclusió purament il·lusòria.

Lectura recomanada

Trobareu el cas estudiat per Wen de l'apendicectomia beneficiosa a l'obra 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 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 ha cap manera d'afirmar que la millora a curt termini es pugui considerar fruit de l'apendicectomia.

Així doncs, la interpretació requereix molta precaució i més d'una mirada sobre els resultats. Per això, conceptes com l'esmentat AutoML estan lluny de ser realment una eina per construir models de mineria de dades de manera automàtica, sempre es requerirà la validació dels experts de l'àmbit de coneixement.

Lectura recomanada

Un treball recent que descriu els errors típics que es poden cometre en la interpretació de models és el següent:
C. Molnar; G. König; J. Herbinger; T. Freiesleben; S. Dandl; C. A. Scholbeck i altres (2020). «Pitfalls to avoid when interpreting machine learning models». arXiv preprint arXiv:2007.04131

2.4. Integració dels resultats en el procés

L'últim pas consisteix a integrar els resultats de la mineria de dades en el procés típic del sistema d'informació en què s'estigui aplicant.

Un exemple senzill és el del procés de documentació textual emprat 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 diferents paraules clau. És clar que darrere de tot això hi ha un esforç previ de classificació. Aplicant algorismes de mineria de dades textuals, és possible construir un procediment de classificació que assigni automàtica-

ment les paraules clau a les diferents 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 en la base de dades documental del diari.

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

Lectura recomanada

Trobareu més informació sobre els processos de documentació textual que es feien servir en alguns grans diaris a l'obra següent:
J. Schmitz; G. I. Armstrong; J. D. C. Little (1990). «CoverStory-Automated News Finding in Màrqueting». *DSS Transactions*.
Actualment, l'àmbit del *text mining* ha evolucionat moltíssim, com mostra el treball següent:
C. C. Aggarwal; C. Zhai (2012). «A survey of text classification algorithms». A: *Mining text data* (pàg. 163-222). Boston, DT.: Springer.

2.5. Observacions finals

Posarem èmfasi en el fet que, malgrat aquesta presentació lineal, el descobriment avança de manera iterativa. No acaba amb la construcció d'un model i amb la generació d'informació resumida. Una vegada que es disposa d'un model, cal treballar-hi, fer preguntes noves, preguntar-se què passa si en lloc de disposar de les relacions que mostra el model n'hi hagués unes altres, etc. Aquest fet pot representar, alhora, l'aportació de dades que no necessitàvem 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.

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

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

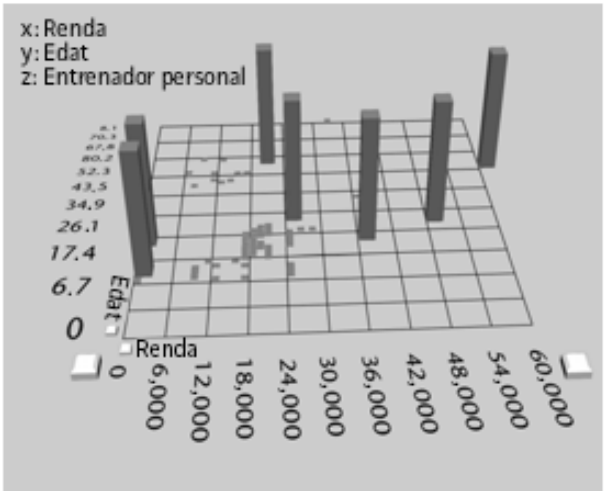
El nucli d'aquest material docent se centra en els models de mineria de dades, els mètodes per construir-los i els algorismes sobre els quals es basen. No obstant això, normalment hi ha una sèrie d'eines que també guarden relació amb la mineria de dades i que almenys hem de tenir present.

3.1. Eines de visualització

Una manera molt potent i intuïtiva d'obtenir coneixement a partir de dades és mitjançant la inspecció visual, aprofitant les capacitats del sistema visual humà.

Vegem un exemple senzill de la potència de les eines d'inspecció visual: a la figura 9 hi apareix representada la relació entre el nivell d'ingressos (Renda, eix X) dels socis del club esportiu que farem servir com a exemple al llarg del programa, la seva edat (Edat, eix Y) i si sol·liciten un entrenador personal (Entrenador personal, eix Z).

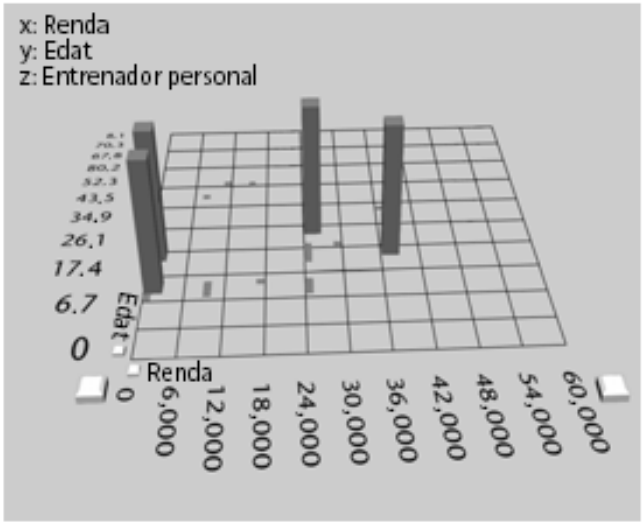
Figura 9



Podem extreure algun coneixement d'aquesta visualització: sembla que la major part dels socis que sol·liciten el servei d'entrenador personal es concentren en les rendes superiors a 24,000 euros, fins i tot una visualització tan poc elegant com aquesta ja ens permet identificar patrons i tendències. Fem unes quantes comprovacions més:

1) Fem un filtratge de les dades per sexe, de manera que al gràfic hi apareguin només els homes:

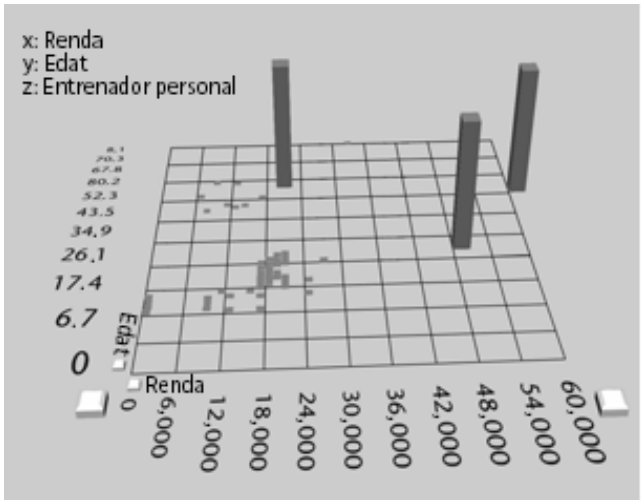
Figura 10



Podem extreure la conclusió que els qui sol·liciten predominantment el servei d'entrenador són homes joves amb rendes mitjanes-baixes.

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

Figura 11. Filtrant segons sexe: dones

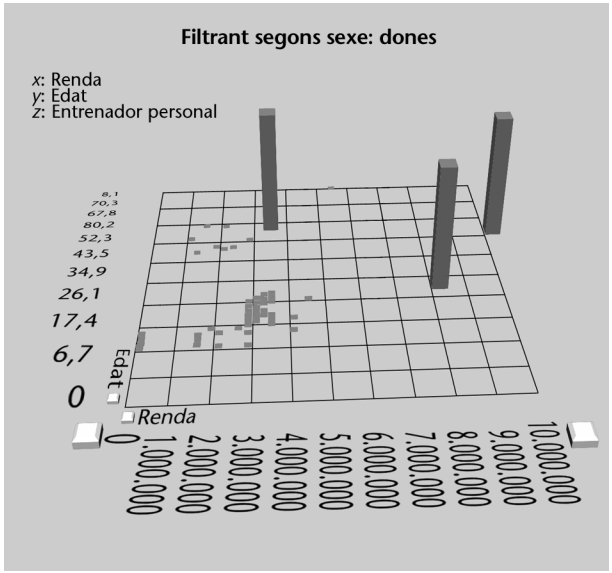


Sembla, doncs, que es tracta de dones grans i amb rendes més altes, oi? Si ara filtrem el que tenim segons el districte de residència i només s'eliminen els socis que procedeixen de la zona A (un barri de classe alta) resulta que gairebé no apareixen socis a la figura 12.

Així doncs, sembla que fins ara arribarem a conèixer bastant bé els clients que sol·liciten el servei d'entrenador personal en aquest gimnàs: són homes joves de renda baixa-mitjana, però que viuen en el districte alt de la ciutat, o bé dones grans amb renda mitjana-alta que també viuen en el mateix barri. Ara, quan ja tenim una primera idea de les dades, podríem aplicar altres mètodes

que ens donessin un resultat numèric més precís i un model de predicció més detallat que el que ens permet la inspecció visual.

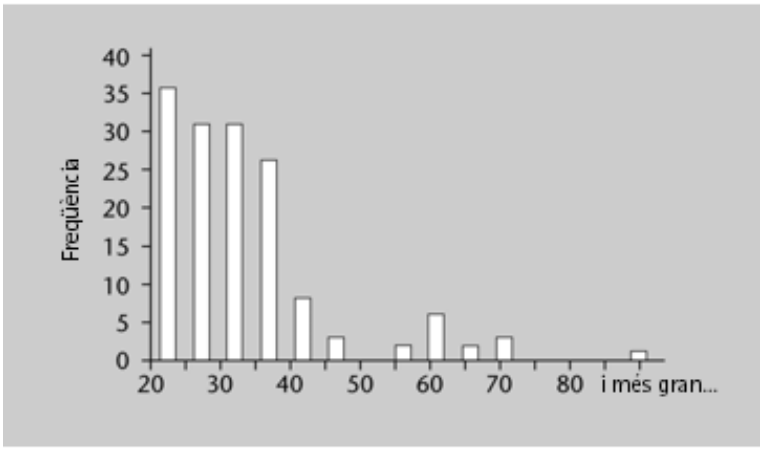
Figura 12



Les eines de visualització són bastant útils en la fase de preparació de dades i la d'interpretació dels models resultants. En la primera fase, aquestes eines permeten ajudar-nos a conèixer millor les dades, i es complementen amb les eines típiques d'estadística descriptiva, que ens permeten 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.

Una de les eines visuals més tradicional en aquesta part són els histogrames. Vegem un exemple d'utilitat dels histogrames. Aquí podem veure la distribució dels valors de les edats entre els clients del nostre club:

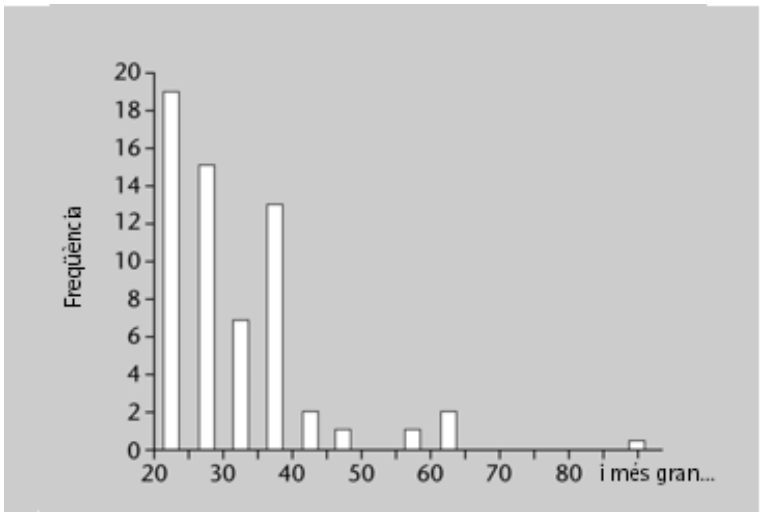
Figura 13



Triant una variable discriminant (o una variable de classe) poden donar-se comparacions entre les distribucions de valors per diverses classes o combi-

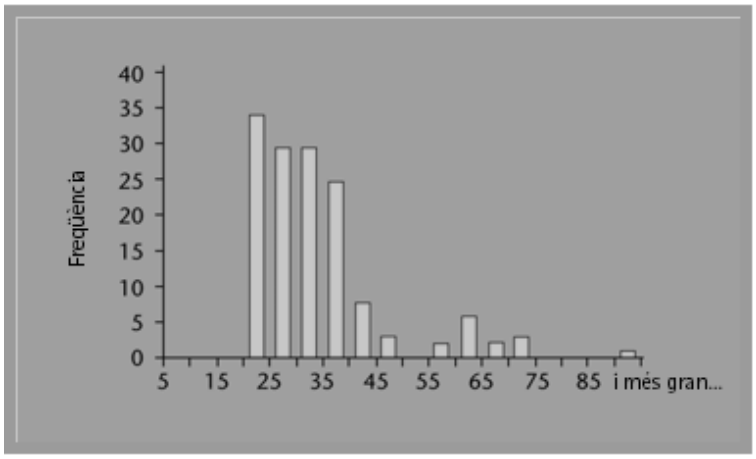
nacions dels valors dels altres atributs. Aquí tenim la distribució de les edats entre les dones que van al club:

Figura 14



I aquí, entre els homes:

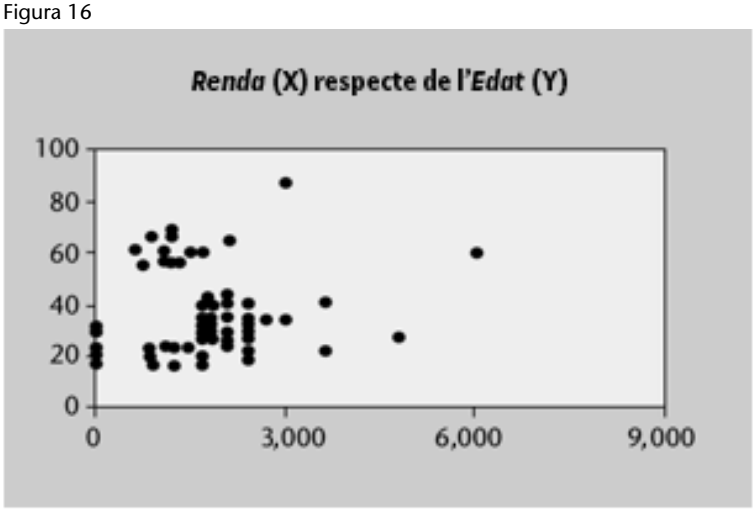
Figura 15



A partir d'aquestes gràfiques, podem dir que hi ha diferències significatives? Malgrat que totes dues gràfiques no poden superposar-se i comparar-se directament, perquè, per exemple, no comparteixen la mateixa escala, sí que permeten observar alguns fets que poden ser rellevants (per exemple, per als usuaris de més edat).

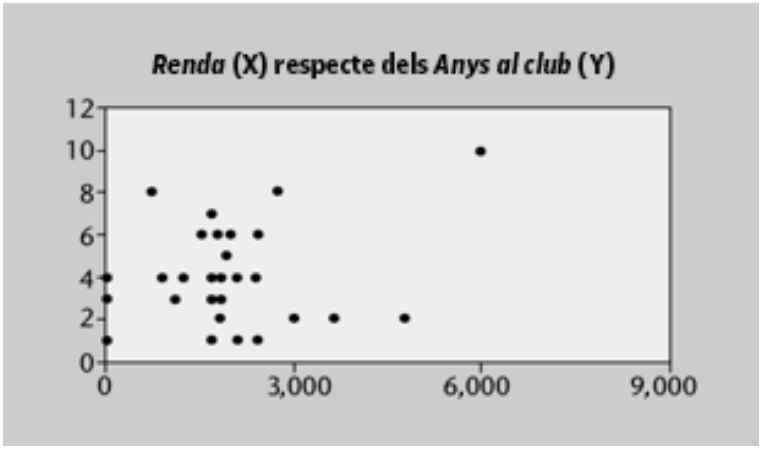
Una altra eina útil són els diagrames de dispersió, en anglès, *scatterplots*, que donen una idea de la relació que hi ha entre els valors de dues variables. Vegem un exemple d'utilitat dels diagrames de dispersió. Aquí podem veure la relació amb els valors de les variables Renda i Edat:

Figura 16



I aquí, entre Renda i Anys al club:

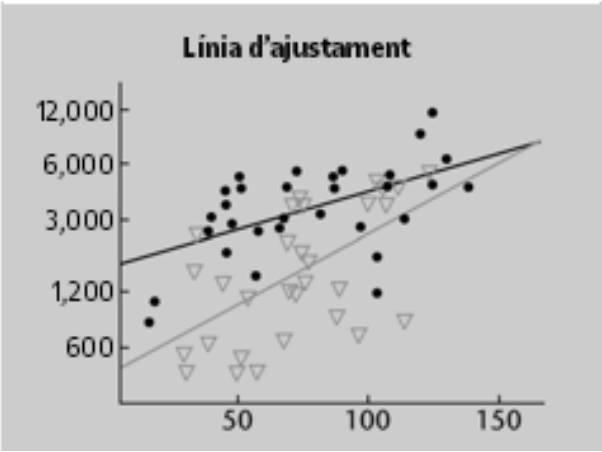
Figura 17



Normalment, l'existència de relació entre els valors de dues variables podem estudiar-la construint una funció que dibuixi una línia que expliqui els valors d'una variable en funció dels de l'altra. Així mateix, es pot derivar el coeficient de correlació entre totes dues variables, o de recta d'ajust o de regressió. Ara no entrarem a definir ni explicar aquests conceptes; només ens centrem en l'aspecte de visualització.

Vegem un 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 per a dos grups diferents.

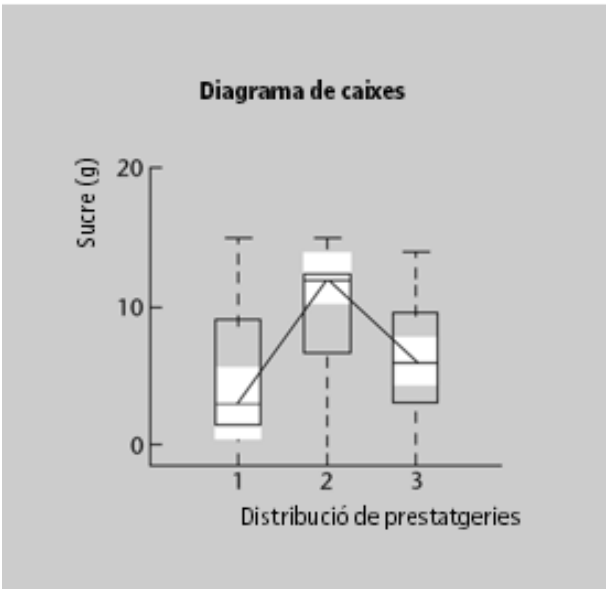
Figura 18. Línia d'ajustament



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

Vegem un exemple d'utilitat d'un diagrama de caixes. Aquí tenim una mostra d'un diagrama de caixes que relaciona determinats productes amb la disposició a les prestatgeries del supermercat:

Figura 19. Diagrama de caixes



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

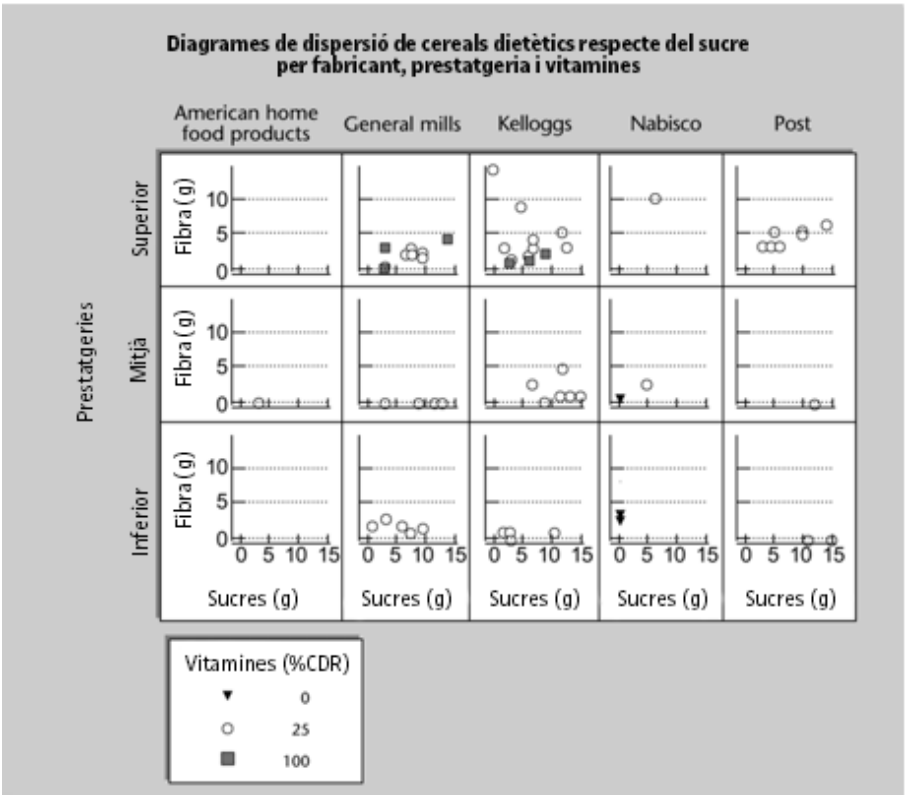
És a dir, si treballem amb observacions que tenen vint atributs, conceptualment estem treballant en un espai de vint dimensions que no podem visua-

litzar de cap manera. Ara bé, projectant part d'aquests atributs en representacions tridimensionals, o més aviat bidimensionals, podem extreure alguna mena d'intuïció que després podem confirmar o refutar amb una altra mena d'eines –procedents de l'estadística i de l'aprenentatge automàtic–, i començar 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 diverses variables i diferents fonts. Aquestes tècniques són bastant comunes i útils per mirar de comparar rendiments de centres dispersos geogràficament.

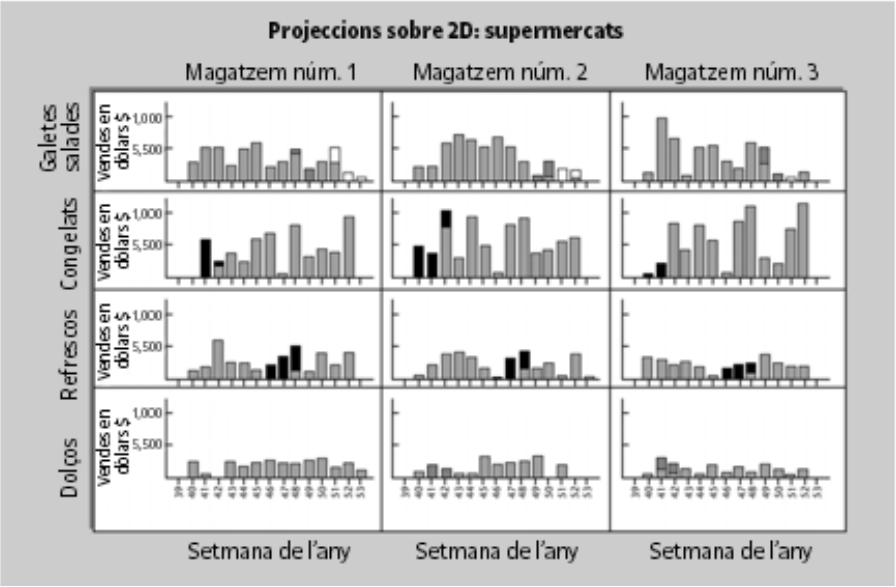
Vegem un exemple de tècniques de projecció sobre 2D: sucre i disposició a les prestatgeries. Aquí tenim un diagrama de dispersió que relaciona el contingut de fibra i sucre en diverses marques de cereals en relació amb la posició que ocupen a les prestatgeries d'un supermercat.

Figura 20. Diagrames de dispersió de cereals dietètics respecte al sucre per fabricant, prestatgeria i vitamines



El camp d'eines de visualització de dades té una gran activitat i hi ha moltes eines per explorar dades. Aquí tenim un altre exemple de projecció sobre 2D:

Figura 21. Projeccions sobre 2D: supermercats



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 aconseguit de diversos productes (caramels, begudes no alcohòliques, aliments congelats i galetes salades) en tres supermercats diferents.

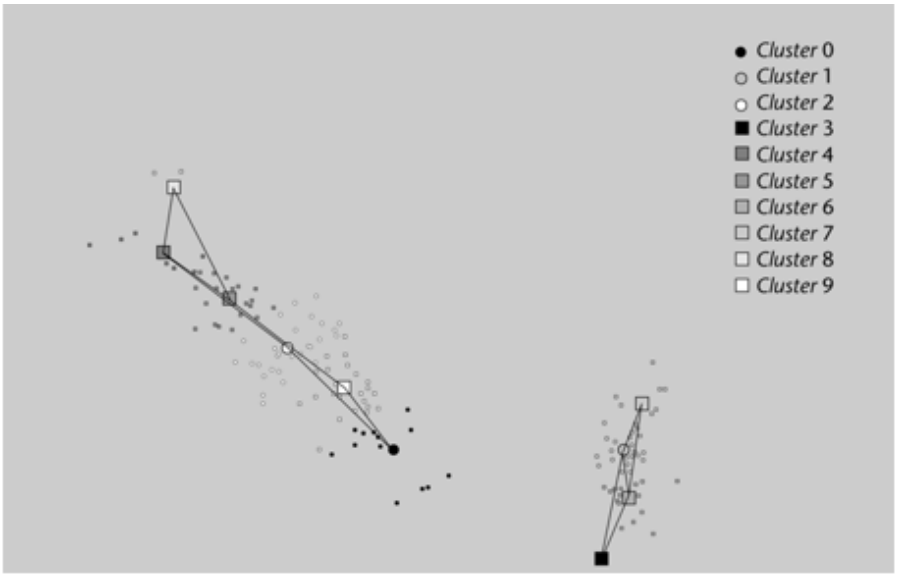
Altres eines permeten combinar diverses formes de visualització en una sola. A la figura següent podem veure com es combina l'estructura d'un arbre de decisió en tres dimensions amb els histogrames que reflecteixen la distribució dels valors de la partició que s'indueix a escala de node:

Figura 22. Visualització en 3D d'histograma en la participació corresponent a un node

Per al cas d'agregació de dades, en anglès *clusters*, en què s'intenta trobar grups de dades semblants, la representació de núvols de punts sobre espais bidimensionals permet estudiar cada grup d'objectes segons les característiques triades.

Vegem un 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 per a aquesta tasca consisteix a representar gràficament el camí que connecta els grups d'objectes (*clusters*) més pròxims o fortament relacionats:

Figura 23

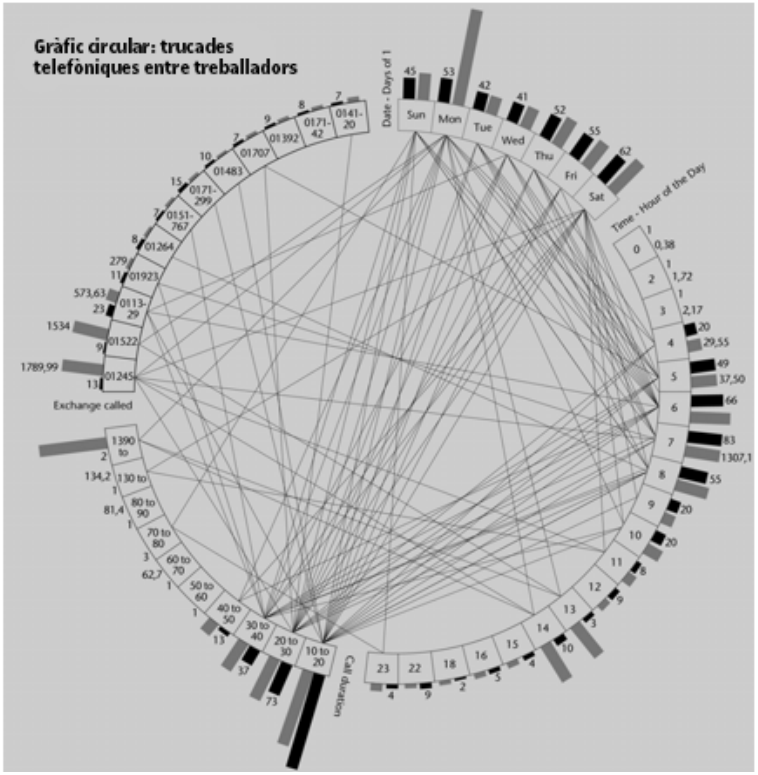


En aquest cas, el gràfic indica que hi ha deu *clusters* i assigna un color diferent als elements de cada *cluster*. Si llegim els colors, podrem veure que hi ha fortes relacions (línies contínues) entre els *clusters* 1, 3, 6 i 7, i entre els grups 0, 2, 4, 5, 8 i 9.

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

Vegem finalment un exemple d'utilitat de gràfic circular. El que veiem aquí relaciona trucades entre diversos números de telèfon dels treballadors d'una empresa:

Figura 24. Gràfic circular: trucades telefòniques entre treballadors



A la vista del gràfic, endevineu qui treballa amb qui? Aquest tipus de visualitzacions permet extreure un coneixement «ràpid» del conjunt de dades que pot ser explotat després.

La potència de les diferents eines de visualització es posa de manifest per la seva correcta connexió amb la resta de les utilitats per a 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. És important tenir en compte que aquestes visualitzacions mostrades com a exemples no són òptimes des del punt de vista de la visualització de dades, són simplement una mostra de la capacitat exploratòria que tenen com a eines d'anàlisi visual.

3.2. *Data warehouse*

El concepte original de *data warehouse* va ser presentat per William Inmon i comercialitzat per IBM amb el terme *information warehousing*, establint l'analogia entre els magatzems físics de les empreses on es podien localitzar de manera flexible els materials segons la necessitat i l'equivalent quant a les dades d'interès de les seves diferents àrees de l'empresa.

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

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

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

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

4) Mantenir la independència entre els procediments dirigits 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 requisits diferents i maneres de veure les dades també diferents i canviants. El *data warehousing* agrupa diversos tipus d'eines i tecnologies.

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

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

2) Eines de conversió automàtica de dades.

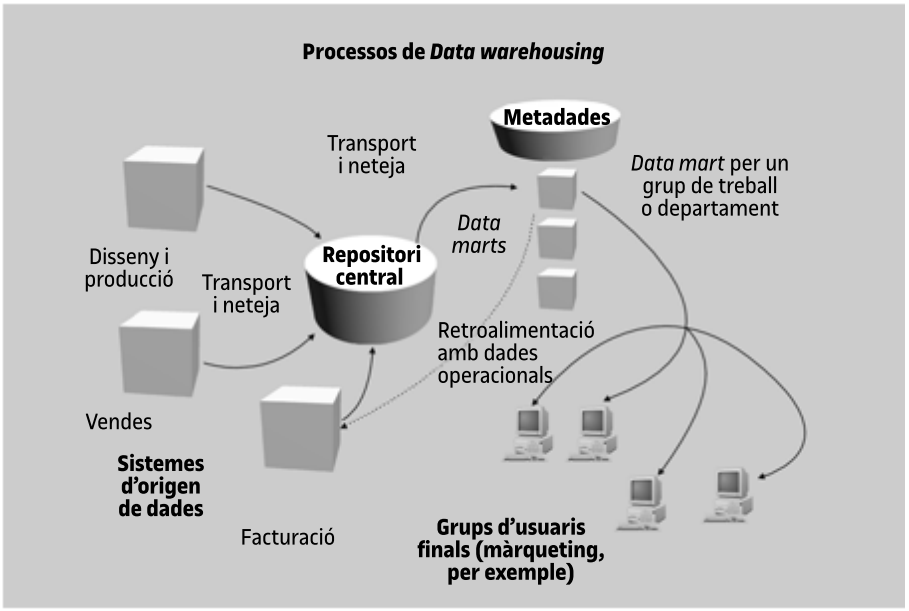
3) Tecnologies client/servidor per accedir a dades distribuïdes en plataformes diferents.

4) Integració d'eines d'anàlisis i relació amb sistemes de presa de decisions, sistemes de presa de decisions en grup i sistemes d'informació per a executius.

L'aspecte més crític del *data warehousing* és probablement el modelatge dels diferents usos i perspectives que es té de les dades, així com la integració transparent dels diferents productes de programari i la seva actualització i millora contínues i el més automàtiques possible.

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

Figura 25. Processos de *data warehousing*



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

1) Sistemes d'origen de dades. Sistemes que recullen les dades en el 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* consisteix a poder integrar les informacions procedents de diferents 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 a altres enclavaments on es guardaran en la forma o formes adequades. Tradicionalment, aquest tipus de procediment era tasca de programació i era difícilment ampliable. Els productes de transport i neteja disponibles avui dia adopten una òptica més d'especificació, en què s'indica d'on procedeixen les dades i què els ha de passar sense arribar a processar-les. Per norma general, aquesta part implica una descripció de les dades en un altre llenguatge de descripció de dades, i crea el que es coneix com a *metadades*.

3) Repositori 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 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 repositori central estan especialment pensades per millorar la construcció dinàmica 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 diferents components de l'empresa i la mantenibilitat del repositori.

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*, intenten millorar la comprensió i l'accés per part dels usuaris finals.

5) *Data marts*. Es tracta de «personalitzar» la visió, els components i els continguts del *data warehouse* segons les necessitats dels diferents 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 decisions integrant-les en el repositori central. Aquesta és una desviació notable respecte de l'ús tradicional de les eines de presa de decisions 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 treballar amb clients. Aquest aspecte del *data warehouse* permet, per exemple, oferir suggeriments a un client després que hagi contestat una sèrie de preguntes directament al personal d'atenció al client. És un dels aspectes en què les eines de mineria de dades ofereixen més resultats.

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*.

En el concepte *data warehousing* poden agrupar-se plataformes de programari, eines d'extracció i conversió de dades, bases de dades preparades per a consultes complexes i dinàmiques, eines d'anàlisis de dades i eines de gestió de bases de dades.

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

Lectura recomanada

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.
I un treball més recent del mateix autor seria el següent:
W. Inmon; D. Strauss; G. Neushloss (2010). *DW 2.0: The architecture for the next generation of data warehousing*. Elsevier.

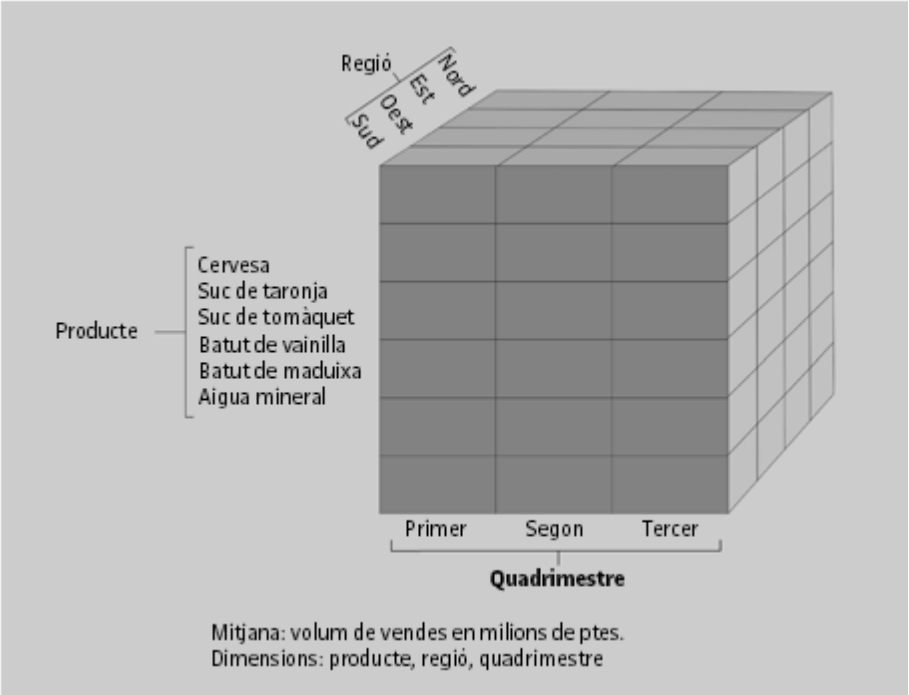
3.3. Mètodes OLAP

OLAP és la sigla de l'expressió anglesa *on-line analytical processing*. Els mètodes OLAP van aparèixer per analitzar les dades de vendes i màrqueting, però també per processar dades administratives i consolidar dades procedents de diverses fonts de cara 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, dirigides a respondre consultes bastant prefixades i rutinàries).

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

Figura 26



La unitat de dades d'OLAP és el «cub», una representació de les dades que permet «tallar-les» a llesques i veure-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 es denominen *columnes de dimensió*, i que faciliten i preveuen dades per resums i informes. Les columnes anomenades *columnes agregades* permeten precalcular quantitats com a recomptes, sumes i mitjanes.

Construir un cub requereix una anàlisi detallada de les necessitats de dades del grup d'usuaris al qual va dirigit, i pot requerir, així mateix, bastant temps, tant de disseny com d'instal·lació per primera vegada. Compensa pel fet que facilita extraordinàriament les tasques d'anàlisis de dades dels diferents grups d'usuaris i, una vegada establert, és més senzill de modificar que les taules relacionals tradicionals.

3.4. Sistemes OLTP

OLTP és la sigla de l'expressió anglesa *on-line transactional processing*. Els sistemes de processament de transaccions en línia (OLTP) tenen com a objectiu guardar la integritat de les dades necessàries per administrar una organització de manera eficient.

Així doncs, els sistemes OLTP volen mantenir models de dades que corresponguin a la visió que cada treballador (o tipus de treballador) té de l'organització. En lloc de veure l'organització com una estructura de dades organitzada de taules i relacions, les eines d'OLTP la presenten en forma de jerarquies i dimensions, de manera que podem observar les mateixes dades des de perspectives diferents.

Els sistemes tradicionals són dinàmics, en la mesura en què sempre estan sent actualitzats amb noves dades. Per a analitzar-los és necessari fer una «fotografia» del seu estat en un moment donat i aplicar les eines d'anàlisis corresponents. Dur a terme aquest tipus de treball sol 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. Així mateix, el sistema de bases de dades pot no estar preparat per guardar el resultat d'aquestes anàlisis.

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

3.5. Estadística

La tasca consistent a analitzar grans volums de dades ha estat i continua sent el regnat de l'estadística, en concret, l'anàlisi de dades. L'enfocament tradicional de l'estadística es dirigeix a la recopilació 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 situat les tècniques estadístiques clàssiques davant una gran oportunitat pràctica i també davant la necessitat de crear eines que, fins i tot mantenint la sòlida fonamentació teòrica aportada per aquesta disciplina, donin respostes fàcilment comprensibles a usuaris no sempre ben preparats estadísticament dins dels límits de temps imposat per la velocitat que requereixen els nous entorns de treball.

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 noves a partir de la seva experiència.

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

Resum

Podem dir, a manera de 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 d'assenyalar 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 arribar a proporcionar la qualitat de resultat exigida pels nous objectius.

En mineria de dades es busca l'obtenció 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 tan comprensible com sigui possible. És important que es pugui 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 el desenvolupament en bases de dades, estadística, aprenentatge automàtic i visualització. Hem d'entendre la mineria de dades com un procés continu que integra els aspectes següents:

1) Definició de l'objectiu del projecte de mineria de dades, precisant la tasca principal que cal fer i triant el mètode més adequat segons les circumstàncies.

2) Selecció de les dades rellevants.

3) Preparació de les dades de cara a assegurar que siguin vàlides i es trobin en condicions de ser emprades pel mètode seleccionat.

4) Mineria de dades pròpiament dita, és a dir, aplicació sobre les dades ja preparades del mètode triat i construcció del model corresponent.

5) Interpretació del model obtingut, que pot provocar la revisió d'algunes de les fases anteriors.

6) 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.