Urrezko datuak

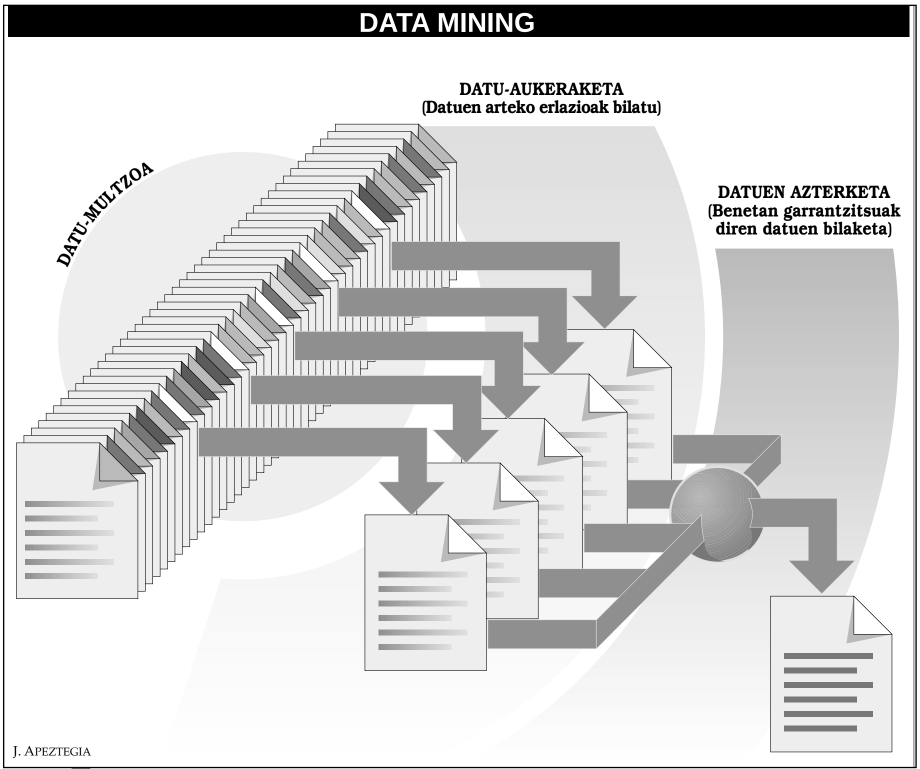
Josu Waliño*

Informazioaren garaian bizi omen gara eta gure inguruan pilatzen diren datuen artean, urrea aurkitu dute zenbaitzuk. Ez da harritzekoa. Gero eta datu gehiago biltegiratzen dute erakundeek eta lastoaren artean orratza bilatzea zaila bada ere, ahalegin horrek sor ditzakeen abantailak era askotakoak izan daitezke. Datuak eskuratzeko teknika berriek boterea lortzeko bidea urratzen lagun dezakete.

Gaur egun, edozein erakundek gordetzen duen datu-multzoa izugarri handia da. Egia da datu horietarik asko alferrik gordetzen direla, baina multzo handi hori aztertzea eta bertan dagoen informazio garrantzitsua ateratzea, kopuru txikia bada ere, lasto artean orratz bat bilatzea bezain zaila da. Hala ere, lan horrek enpresa bati ekar diezazkiokeen onura ekonomikoak oso handiak dira. Adibide soil batez hobeto ulertuko duzu adierazi nahi duguna; demagun jatetxe bateko kudeatzailea zarela eta zure bezeroek eskatu ohi dituzten menu-konbinazio ezberdinetan oinarriturik, zein plater berri eskaintzea komeni zaizuen jakiterik baduzula. Abantailak, bada, begibistakoak dira. Eta adibide berak balio digu supermerkatuek, burtsek, enpresa handiek edo poliziak berak erabil ditzaketen informazioa eskuratzeko. Eta hau hasiera besterik ez da.
Hori guztia posible da gaur egun *data mining* deituriko teknika berri bati esker. Teknika hau garatzen ari diren zientzilariak datu-multzo osoa hartu eta, hainbat azterketa estatistiko erabiliz, datuen artean egon daitezkeen erlazioak aurkitzen saiatzen dira; datu-multzo horretatik zaborra baztertzen dute eta benetan garrantzitsua den informazioa jasotzen dute. Urre bilatzaileen moduan egiten dute lan, ibaietako lurra bahetuz, urre zati txikien bila ibiliko bailiran, alegia.

Informazioa eskuratzea: oztopo-lasterketa

Aipatu emaitzetara iristeko, hainbat teknika erabil daitezke. Horietako bat "arau-zuhaitzen indukzio" izenaz ezagutzen dena da: konbinazio desberdinen bidez, arau egokienak azalduko dizkigu metodo honek. Adibidez,



***Data mining* deituriko teknika garatzen ari diren zientzilariak datu-multzo osoa hartu eta, hainbat azterketa estatistiko erabiliz, datuen artean egon daitezkeen erlazioak aurkitzen saiatzen dira; datu-multzo horretatik zaborra baztertzen dute eta benetan garrantzitsua den informazioa jasotzen dute.**

"BALDIN entsalada ETA makailu tortila ORDUAN txuleta piperrekin". Huskeria badirudi ere, eskura dezakegun datu-multzoekin era horretako konbinazioak egitea arazo bihur daiteke, prozedura horren bidez sortzen diren konbinazioak oso konplexuak direlako: *"BALDIN patatak ETA(EZ xerra ETA piperrak ETA(EZ izozkia ETA kafea) ETA ......"* Arazo horiek gainditzeko, teknika aurreratuagoak garatu dira eta horien artean sare neuronalen erabileran oinarriturikoa. Sistema honen ekarpena funtzionamendurako duen oinarrian datza: giza pentsamenduaren logika imitatzen saiatzen dira datuen artean dauden erlazioak bilatzeko. Sare neuronalek indukzioak baino emaitza hobeak ematen dituzte (% 75 inguruko asmatze-tasa), baina datuak erlazionatzeko erabiltzen duen arau-multzoa oso konplexua eta askotan ulertezina gerta daiteke. Horrek bi arazo sortzen ditu: batetik, datuen azterketa eskatu duten bezeroei prozesua zertan oinarritu den azaldu ezina —adibidez, bere menpe dauden zenbait enpresek duten porrot arriskua esaterako orduan—, eta bestetik, sarean matxurarik sortuz gero, oinarrizko arauak berraztertu ezina.

Baina arazo horiei ere aurki jarri ko zaie konponbidea. Gaur egun "algoritmo genetikoak" erabiltzen hasi dira: finantza-arloko neurri-arauetan oinarrituriko printzipioak aplikatzen zaizkie datuei bahetzeko orduan. Teknika hau oso eraginkorra ez bada ere, bezeroentzat ulergarriagoa da. Beste aukera egokia da logikako metodo sinpleak erabiltzea datuen arteko erlazioak azalduko dizkiguten arauak aurkitzeko, edo forma normal disjuntiboan oinarriturikoa, honek emaitza oso onak ematen baititu.

Baina zalantzarik gabe gaur egun arrakasta eta etorkizun gehien duten tekniken artean lengoaia naturalean oinarriturikoak nagusitzen ari dira: ordenadorea kontrolatzeko, lengoaiaren hitz arruntak erabiltzen dituzten teknikak. Hauen arrakastaren zergatia bistakoa da: gaur egun, munduan dauden datu gehienak testu arruntean daude oraindik, paperean, mikrofitxa edo testu-prozesadoreko orritan gordeta eta beraz, datu horiek irakurtzea zaila da datu-bilatzaileentzat. Horrela, teknika berri honek izan ditzakeen abantailak nabariak dira, testu arruntetan datuak aztertzen dituzten software-paketeak sortu baitira beraiei esker. Horrela uler daiteke beraz, zenbait sektoretan teknika hauek garatzeko azaltzen den interesa. Esate baterako, nahikoa da poliziak izan ditzakeen pertsona susmagarrien zerrendak testu moduan izatea, programa honek erabiltzen dituen multzoen teoria eta analisi linguistikoa erabiliz, zuzenean "Nor da susmagarririk handiena?" galdetuta erantzun azkar bat lortzeko.

Harrigarria gerta daiteke hori, baina prozesu luze baten lehen urratsa baino ez da izango; sortu berria den teknika honek emango digu oraindik harritzeko aukera gehiago izan ditzakeen aplikazio desberdinetan.

*** Elhuyar - ZETIAZ**