Arbres de classificació i regressió en recerca biomèdica

M.Luz Calle[1] i Josep Anton Sánchez[2]

Grup de Recerca en Bioinformàtica i Estadística Mèdica
http://www.uvic.cat/eps/recerca/bioinformatica/ca/inici.html

GRASS. Grup de Recerca en Anàlisi Estadística de la Supervivència

1. Dept. de Biologia de Sistemes
Universitat de Vic
malu.calle@uvic.cat

2.Dept. d'Estadística i I.O.
Universitat Politècnica de Catalunya
josep.a.sanchez@upc.edu

Un objectiu important en la investigació clínica és l'obtenció de regles de decisió fiables que puguin utilitzar-se en la pràctica clínica. Per exemple, regles que facilitin la presa de decisions adequades davant d'una situació d'urgència o, més en general, regles que permetin classificar els pacients en diferents grups de risc i, conseqüentment, assignar-los-hi els tractaments més apropiats. L'obtenció d'aquest tipus de regla de decisió no es una tasca senzilla ja que requereix un coneixement profund dels diferents factors que intervenen, de les seves possibles interaccions i de l'efecte sobre la resposta del pacient.

Els mètodes estadístics tradicionals, como la regressió logística, no són sempre apropiats per a tractar aquest tipus de problema de classificació. Això és així quan el grup de pacients que s'està estudiant és molt heterogeni, quan hi ha una gran quantitat de factors o variables predictores possibles i quan hi ha indicis de possibles interaccions complexes entre els diferents factors. Aquestes interaccions són generalment difícils de modelar i pràcticament impossible de modelar quan el nombre de variables i interaccions és molt gran. Per aquest motiu és molt important disposar de tècniques alternatives que permetin revelar l'estructura oculta de les dades i reduir el nombre de possibles predictors sense necessitat d'explicitar un model concret.

Una d'aquestes tècniques són els arbres de classificació i regressió (CART). En els darrers anys hi ha hagut un interès creixent en l'ús d'aquesta metodologia en molts camps diferents i, especialment, en estudis d'epidemiologia genètica. Aquest tipus d'estudi habitualment involucra una gran quantitat de predictors, marcadors genètics com els SNPs o informació sobre expressió genètica, que presenten patrons complexes d'interacció. Els arbres de classificació i regressió han demostrat ser efectius en aquest tipus de situació permetent identificar o seleccionar els factors més rellevants, explorar les interaccions entre les variables clíniques, genètiques i ambientals i analitzar l'impacte d'aquestes interaccions sobre la susceptibilitat a desenvolupar una determinada malaltia o sobre la progressió d'aquesta malaltia. En aquest sentit, aquesta tècnica permet analitzar interaccions complexes associades a respostes continues, binàries (cas/control), nominals, comptatges i de temps fins a esdeveniments en l'àmbit de la supervivència.