METODOLOGIE STATISTICHE DI CLASSIFICAZIONE A CONFRONTO: ANALISI DISCRIMINANTE E CART

Maddalena Agonigi, Silvia Calamai, Irene Ricci

SNS, Pisa

RIASSUNTO

Un problema di classificazione si verifica quando è necessario dover attribuire ad un 'dato' una classe di appartenenza sulla base di misurazioni condotte e quando si vuole stabilire una regola che consenta di catalogare nuovi dati sulla base di quelli già esistenti: in entrambi i casi ciò di cui si ha bisogno è un classificatore o una regola di classificazione. Le metodologie che possono essere impiegate a questo scopo sono molteplici, quelle più comunemente utilizzate per la risoluzione di simili questioni sono l'analisi discriminante, la tecnica del *Kernel density estimation* e quella del *Kth nearest neighbor*. Sono invece più recenti le applicazioni di metodologie ad albero a problemi di classificazione, come ad esempio gli algoritmi CART (*Classification and Regression Trees*), CHAID (*Chi-squared Automatic Interaction Detector*), QUEST (*Quick, Unbiased, Efficient Statistical Tree*).

CART è un algoritmo a segmentazione binaria, la cui peculiarità è quella di presentare i risultati nella forma di diagrammi ad albero rovesciato costituito da nodi che vengono suddivisi in altri due livelli, e così via: in questo modo la classificazione finale, ottenuta in maniera automatica, ha una forma molto semplice e compatta che consente una agile interpretazione della struttura predittiva dei dati e quindi una veloce classificazione di nuove entrate. Uno dei suoi principali vantaggi è nel non fare alcun tipo di assunzione sul set di dati oggetto di analisi, né sul tipo di variabile, né sulla distribuzione statistica seguita dalle variabili stesse; la metodologia può essere applicata a dati di tipo categorico e/o numerico, a dati con struttura non standard, a set di dati ampi e ad alta dimensionalità.

Un database di fonemi vocalici provenienti da soggetti di area toscana è stato già sottoposto all'analisi discriminante lineare per valutare quali parametri e quali scale siano più efficaci nella individuazione delle diverse categorie vocaliche (Calamai 2003). L'analisi discriminante è un test statistico peraltro largamente utilizzato nell'analisi dei sistemi vocalici delle lingue naturali (vedi, tra gli altri, Syrdal & Gopal 1986, Cosi et al. 1995, Jassem et al. 2001); risulta essere scarsamente utilizzata, al contrario, la metodologia CART (un'applicazione recente della metodologia in ricerche di tipo prosodico è Bouzon & Hirst 2002): il suo utilizzo permetterà di valutare lo stesso corpus di dati in un'ottica differente e di porre a confronto i risultati ottenuti mediante le due diverse procedure statistiche, evidenziandone debolezze e potenzialità.

Riferimenti bibliografici

- Bouzon C. & D. Hirst 2002 "The influence of Prosodic Factors on the Duration of Words in British English", *Speech Prosody 2002. An International Conference*, Aix-en-Provence.
- Breiman L., J.H. Friedman, R.A. Olshen & C.J. Stone 1984 *Classification and Regression Trees*, Boca Raton, Chapman & Hall/CRC.
- Calamai S. 2003 "Spazi acustici e spazi uditivi di vocali pisane", in G. Marotta & N. Nocchi (acd), *La coarticolazione. Atti delle XIII giornate di studio del gruppo di fonetica sperimentale*, Pisa, ETS: 265-275.
- Cosi P., F.E. Ferrero & K. Vaggies 1995 "Rappresentazioni acustiche e uditive delle vocali italiane", in A. Cocchi (acd), *Atti del XXIII Convegno Nazionale AIA*, Bologna: 151-156.
- Syrdal A.K. & H.S. Gopal 1986 "A Perceptual Model of Vowel Recognition based on the Auditory Representation of American English Vowels", *The Journal of the Acoustical Society of America*, 79: 1086-1100.
- Jassem W., M. Krzysko & W. Wolynski 2001 "Reduction of Speaker-related Variability on Polish Vowel Spectra", *Journal of the IPA*, 31: 187-202.
- Guida per l'utente di AnswerTree 3.0. 2001. SPSS

Sessione a cui il lavoro fa riferimento: **1. Linguistica, Fonologia e Fonetica**