Nota 19.2

Questo campo della statistica si occupa della costruzione di raggruppamenti (in inglese *cluster* significa grappolo, gruppetto) da un insieme di dati ed è particolarmente adatto per l'uso degli algoritmi genetici (cfr. nota 19.3), sia perché mutazioni e incroci sono definibili in modo molto naturale, sia perché nella cluster analysis viene utilizzata una molteplicità di criteri di ottimalità per le partizioni che negli approcci tradizionali richiedono ogni volta algoritmi di ottimizzazione diversi e spesso computazionalmente difficili e quindi non applicabili per insiemi grandi (e spesso anche solo medi) di dati, mentre gli algoritmi genetici non dipendono dalle proprietà matematiche delle funzioni utilizzate e hanno una complessità che cresce solo in modo lineare con il numero dei dati. Siccome l'algoritmo non dipende dalla funzione di ottimalità scelta, lo stesso algoritmo può essere usato per un criterio di ottimalità qualsiasi. Nella letteratura è descritta una grande varietà di misure di somiglianza o di diversità, tra le quali in un'applicazione concreta si può scegliere per definire l'ottimalità delle partizioni, ma il modo in cui viene usato l'algoritmo genetico è sempre uguale.

È per esempio piuttosto difficile trovare algoritmi tradizionali per il caso che l'omogeneità e la diversità dei gruppi non siano descritte mediante misure di somiglianza o diversità tra gli individui ma direttamente da misure per i gruppi, mentre ciò non causa problemi per l'algoritmo genetico.

Elenchiamo alcuni campi di applicazione del raggruppamento automatico: classificazione di specie in botanica e zoologia (*tassonomia numerica*) o di aree agricole o biogeografiche, classificazione di specie virali o batteriche, definizione di gruppi di persone con comportamento (istruzione, attitudini, ambizioni, livello di vita, professione) simile in studi sociologici o psicologici, creazione di gruppi di dati omogenei nell'elaborazione dei dati (per banche dati o grandi biblioteche), elaborazione di immagini (ad esempio messa in evidenza di formazioni patologiche in radiografie mediche), individuazione di gruppi di pazienti con forme diverse di una malattia o riguardo alla risposta a un tipo di trattamento, classificazione di malattie in base a sintomi e test di laboratorio, studi linguistici, raggruppamento di regioni (province, comuni) relativamente a caratteristiche economiche (o livello generale di vita o qualità dei servizi sanitari), individuazione di gruppi di località con frequenza simile per quanto riguarda una determinata malattia, reperti archeologici o paleontologici o mineralogici (descritti ad esempio mediante la loro composizione chimica) o antropologici, dati criminalistici (impronte digitali, caratteristiche genetiche, forme di criminalità e loro distribuzione geografica o temporale), confronto tra molecole organiche, classificazione di scuole pittoriche, indagini di mercato (in cui si cerca di individuare gruppi omogenei di consumenti), raggruppamenti dei clienti di un assicurazione in gruppi per definire il prezzo delle polizze, classificazione di strumenti di lavoro o di prodotti nell'industria oppure dei posti di lavoro in una grande azienda, confronto del costo della vita nei paesi europei, divisione dei componenti di un computer in gruppi per poterli disporre in modo da minimizzare la lunghezza di cavi e circuiti.

In queste applicazioni, che si differenziano fortemente per la quantità degli oggetti da classificare (poche decine nel caso di oggetti archeologici, milioni di pixel nell'elaborazione di immagini) e per la natura dei dati, spesso non è facile scegliere un criterio di ottimalità robusto (cambi di scala possono ad esempio influenzare l'esito della classificazione, quando si usano distanze euclidee) e superare la spesso notevole complessità computazionale.

Nota 19.3

Gli algoritmi genetici sono una famiglia di tecniche di ottimizzazione che si ispirano all'evoluzione naturale. I sistemi biologici sono il risultato di processi evolutivi basati sulla riproduzione selettiva degli individui migliori di una popolazione sottoposta a mutazioni e ricombinazione genetica. L'ambiente svolge un ruolo determinante nella selezione naturale in quanto solo gli individui più adatti tendono a riprodursi, mentre quelli le cui caratteristiche sono meno compatibili con l'ambiente tendono a scomparire.

L'ottimizzazione genetica può essere applicata a problemi le cui soluzioni sono descrivibili mediante parametri codificabili capaci di rappresentarne le caratteristiche essenziali. Il ruolo dell'ambiente viene assunto dalla funzione obiettivo che deve essere ottimizzata.

Questo metodo presenta due grandi vantaggi: non dipende da particolari proprietà matematiche e soprattutto la complessità è in generale praticamente lineare. Negli algoritmi genetici, dopo la generazione iniziale di un insieme di possibili soluzioni (individui), alcuni individui sono sottoposti a mutazioni e a scambi di materiale genetico. La funzione di valutazione determina quali dei nuovi individui possono sostituire quelli originali.

Questa tecnica viene applicata con successo a problemi di ricerca operativa, al raggruppamento automatico (un campo della statistica che si occupa di problemi di raggruppamento e classificazione di dati), al problema del commesso viaggiatore, all'approssimazione di serie temporali, alla previsione della conformazione spaziale di proteine a partire dalla sequenza degli aminoacidi, all'ottimizzazione di reti neuronali e di sistemi di Lindenmayer, a modelli di vita artificiale (sociologi tentano invece di simulare l'evoluzione di comportamenti, ad esempio tra gruppi sociali o nazioni).

Nell'applicazione di questi metodi il matematico può intervenire in vari modi: nello sviluppo e nel controllo degli algoritmi (generazione di numeri casuali per la ricerca di conformazioni ottimali in uno spazio multidimensionale di conformazioni, grafica al calcolatore), nella codifica dei dati, nell'organizzazione delle informazioni.

Nota 19.6

Per fare bene il suo lavoro, lo statistico che lavora in un'azienda, nell'amministrazione pubblica o nella ricerca clinica, deve comprendere i compiti che gli vengono posti e deve essere in grado di interagire con i committenti. Nonostante ciò la statistica è di sua natura una disciplina matematica che si basa sul calcolo delle probabilità, una teoria astratta e difficile, e richiede conoscenze tecniche in altri campi della matematica come analisi reale e complessa, analisi armonica, calcolo combinatorio (ad esempio per la pianificazione di esperimenti). Nell'analisi delle componenti principali e nella ricerca di raggruppamenti sarà compito dello statistico scegliere la rappresentazione dei dati e le misure per la somiglianza o diversità di individui e gruppi. In questo corso abbiamo potuto accennare solo ad alcune delle difficoltà concettuali e tecniche che si incontrano.

Nella statistica multivariata in particolare probabilmente molte tecniche sono ancora da scoprire e i metodi più efficienti si baseranno forse su metodi geometrici avanzati, ad esempio della geometria algebrica reale e della teoria delle rappresentazioni di gruppi.

Ci sono tanti campi di applicazione della statistica in medicina, bioinformatica, farmacologia, matematica finanziaria, linguistica, demografia, che uno studente che intraprende questa professione dopo aver acquisito una solida formazione matematica può sperare in un'attività interessante e gratificante.

L'abitudine ai dati e alla loro interpretazione formerà le sue capacità di giudicare situazioni complesse in modo razionale oltre a fornirgli un ricco patrimonio di informazioni, quindi potrà anche aspirare a una carriera amministrativa o manageriale.

Nel suo lavoro giornaliero potrà, nei contatti con ricercatori clinici o amministratori o con l'opinione pubblica utilizzare le proprie conoscenze teoriche per chiarire il significato di risultati di test clinici o di rilievi statistici o per proporre nuovi esperimenti o indagini.