POLITECNICO DI MILANO
Facoltà di Ingegneria
Dipartimento di Elettronica e Informazione

Sistemi informatici per il riconoscimento automatico di timbriche musicali in segnali monofonici mediante tecniche di statistica multivariata

Relatore Interno: **Prof. C. Ghezzi**
Relatore Esterno: **Prof. G. Haus**
Correlatore: **Ing. E. Pollastri**

Tesi di Laurea di
Giulio Agostini

Anno Accademico 1998–1999

alla Sig.ra Bonzanni

Ringraziamenti

Desidero ringraziare il Prof. Ghezzi per la disponibilità, il Prof. Haus per i suoi indispensabili insegnamenti e per la fiducia.

Ringrazio ardentemente Emanuele Pollastri, che mi ha continuamente incoraggiato e concretamente aiutato anche nei momenti più difficili, dando vita ad una collaborazione fattiva e ricca di soddisfazioni. A lui sono dovuti gli algoritmi di segmentazione ed estrazione delle caratteristiche.

Alessandro Contenti ha fornito un contributo inestimabile nelle fasi di configurazione dei compilatori e di stampa. La mia riconoscenza, pur nella sua finitezza, va ben oltre, e comprende cinque anni di profonda amicizia.

Grazie al Prof. Flury, a Keith Martin e al Prof. Battistini per l'interesse dimostrato e per la collaborazione. Grazie a Giorgio Zoia per aver letto i miei deliri ed avermi incoraggiato nella fase iniziale.

Grazie a mia madre e a mia sorella Elena, per aver sopportato l'insopportabile. A mio padre per il sostegno a distanza. A Sara per la rassegnazione.

Grazie a Marco Abramo per avermi dato una mossa.

Grazie ai pachidermi Alessandro Agustoni e Giovanna Ferrara per essermi stati molto vicini in questo periodo travagliato. Grazie a tutti gli altri compagni di avventura.

Grazie ai miei amici nella musica: Damiano Rota, Andrea Carnevali, Paolo Ravasio, Matteo Perego, Giordano Donadoni e Gianmarco Colombo.

Grazie ai ragazzi di `it.comp.lang.c++` e di `it.comp.software.tex` per aver fornito soluzioni ai miei problemi in tempi ridottissimi, a Fabrizio Fantasia per il compilatore, e al Prof. Wicks per la solerzia.

Grazie alla musica che ho ascoltato e che ho suonato. In particolare grazie a Giusseppe Jelasi e a Michele Sala, per avermi prestato tanti bei CD. E al mio masterizzatore, per averne fatto una copia di sicurezza.

Grazie a tutti coloro che mi hanno fatto ridere.

Ricordo con affetto Galbu, lo zio Antonio, il Prof. Flury e la casa di Pontida.

Indice

capitolo 1

Introduzione

L'inserimento dei linguaggi di *audio strutturato* (*Structured Audio* [87, 90, 91]) all'interno del neonato standard MPEG-4 ha riacceso nella comunità scientifica, se mai si fosse affievolito, il desiderio di realizzare uno strumento di trascrizione automatica, in grado di separare tutte le sorgenti sonore all'interno di un brano musicale complesso e scriverne lo spartito, a dire il Santo Graal dell'Informatica Musicale. Le applicazioni commerciali a livello *consumer* di un *unmixer* di questo tipo sarebbero innumerevoli, e questo è sufficiente a giustificare gli sforzi della ricerca in questa direzione.

Volendo grossolanamente e semplicisticamente suddividere questa impresa in sotto-obiettivi, funzionalmente interrelati, si enucleano due blocchi principali: il *pitch-tracking*, che consente di estrarre da un frammento audio musicale l'informazione relativa alla partitura, e il riconoscimento delle sorgenti sonore (*sound-source recognition*), che consente di etichettare con un nome, o con una struttura dati complessa a piacere, le timbriche impiegate nel brano, associando le parti ai singoli strumenti. Ciò rispecchia la netta separazione delle funzioni descrittive attribuite ai due linguaggi dello Structured Audio: SASL (*Structured Audio Score Language*, linguaggio di partitura) e SAOL (*Structured Audio Orchestra Language*, linguaggio di orchestrazione [89, 93]).

Il presente lavoro si colloca nella seconda area di ricerca (*sound-source recognition*), e si propone uno scopo necessariamente più modesto, ma pure stimolante: quello di realizzare un sistema informatico in grado di riconoscere, a partire dal segnale audio relativo all'esecuzione di alcune note, lo strumento musicale che le ha generate, interrogando, attraverso tecniche di statistica multivariata e di analisi dei *cluster*, una base di dati contenente le caratteristiche salienti di un numero finito di strumenti (figura 1.1). Per

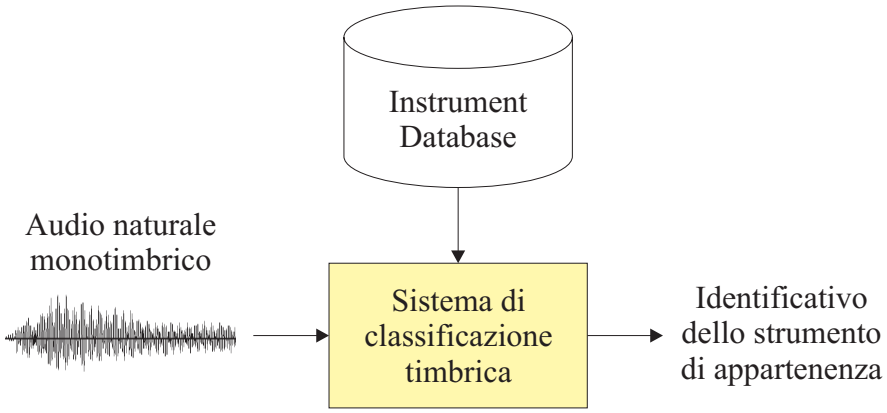


Figura 1.1. Sistema di riconoscimento timbrico per esecuzioni monotimbriche.

fare questo, si pongono ipotesi di lavoro piuttosto restrittive, quali la monotimbricità e monofonicità dei dati in ingresso. Si considerano, cioè, semplici *linee melodiche* eseguite da un solo strumento alla volta.

Parallelamente, si intendono fornire alcuni strumenti atti a investigare la rilevanza delle caratteristiche significative estratte dai brani audio al fine della classificazione.

1.1 Principali aree di ricerca nel campo dell'*audio content analysis* e relative applicazioni

In questa sezione si effettuerà un'analisi più generale e accurata delle aree di ricerca attive nell'ambito dell'analisi automatica del contenuto dell'audio (*audio content analysis*). In un secondo momento, si considerano le principali applicazioni che ne possono scaturire.

Audio segmentation and classification

La *segmentazione e classificazione audio* si occupa di individuare, all'interno di brani audio, le regioni che presentano contenuto omogeneo al più alto livello di astrazione: si distinguerà, tra musica, parlato, rumore e silenzio. Tuttavia, il numero e il genere di contenuti da separare varia a seconda del tipo di applicazione. Ad esempio, ai fini del monitoraggio del palinsesto di una emittente radiofonica, si può pensare di sviluppare un sistema automatico che conteggi i minuti di pubblicità durante l'arco della giornata. Per un sistema di recente realizzazione, si veda [109].

Content retrieval

Il proliferare di basi di dati audio, grazie anche alla diffusione di efficienti tecnologie di compressione, ha fornito una spinta notevole verso la ricerca di tecnologie in grado di effettuare interrogazioni volte al reperimento di informazioni multimediali (*content retrieval*), con la stessa facilità ed efficienza a cui si è abituati per le ricerche in documenti testuali. Alcuni motori di ricerca (ad esempio AltaVista) hanno recentemente recepito questa necessità, e forniscono un semplice sistema di ricerca testuale basata sul nome del file e del titolo della pagina che lo contiene. Si possono individuare tre approcci a questo problema.

1. Ridurre il problema ad una ricerca testuale. I brani audio che si prestano maggiormente a questa soluzione sono quelli contenenti puro parlato, come i telegiornali o le conferenze stampa. Già l'estrazione dei testi da un brano musicale si rivela molto più problematica. Tuttavia, attraverso una indicizzazione automatica dei contenuti (vedi sotto), questa tecnica è applicabile a qualsiasi tipo di sorgente.
2. Confrontare due brani audio (*audio-audio matching*). La tecnica è particolarmente adatta qualora non sussistano problemi di banda passante e di spazio di memorizzazione, e ci siano stringenti requisiti di tempo reale. La tecnica di base più utilizzata in questi casi è il calcolo della correlazione tra i due segnali. Una possibile applicazione è, ancora, il monitoraggio di un flusso audio radiofonico per individuare tutte le istanze di trasmissione di un determinato brano o spot pubblicitario. Per una tecnica euristica molto efficace, si veda [32].
3. Confrontare i due brani sulla base di una rappresentazione ad un più alto livello di astrazione. Un esempio in campo musicale viene dal *melody retrieval*, in cui ogni brano contenuto nella base di dati è rappresentato da una o più melodie (ad esempio in formato MIDI), e la *query* in ingresso, che consiste in un file audio canticchiato dall'utente, viene convertito in MIDI e confrontato.

Automatic audio indexing

Con l'*indicizzazione automatica dei contenuti* [29] si intende la trascrizione e codifica di descrizioni relative ad uno *stream*

multimediale e ai singoli eventi in esso contenuti. Il nascituro standard MPEG-7 propone numerosi e ricchissimi schemi di codifica dei diversi media, comprendente un insieme di descrittori ed un linguaggio di definizione per estenderlo a seconda delle necessità della particolare applicazione. In linea con la filosofia MPEG, gli *algoritmi* di codifica e decodifica non faranno parte dello standard.

Speech analysis

Il parlato è oggetto di interesse fin dagli anni 60, e si possono individuare due principali aree di ricerca. Il *riconoscimento del parlato* (*speech recognition*), si propone di trascrivere il testo relativo ad un discorso o ad una conversazione, mentre l'*identificazione del parlatore* (*speaker id*) si occupa di discriminare l'impronta vocale caratteristica di ogni essere umano. I risultati in questo campo sono molto incoraggianti, arrivando a tassi di errore del 17,4% per parlato non espressamente registrato per *speech recognition*, tanto che numerosi pacchetti commerciali consentono la dettatura automatica in diversi linguaggi, con tassi di errore soddisfacenti. Si noti, tuttavia, che in presenza di grosse quantità di audio preregistrato da convertire in testo (ad esempio un archivio di servizi giornalistici), le prestazioni di una conversione in tempo reale può non essere sufficiente.

Genre recognition

Per quanto riguarda il segnale musicale, sono stati messi a punto diversi sistemi in grado di determinarne il genere. Similmente agli altri sistemi di segmentazione e identificazione[1], essi prevedono una prima fase di estrazione di caratteristiche (*features*), e il successivo raffronto con caratteristiche estratte da un insieme di brani rappresentativi, attraverso le più svariate tecniche di *pattern-recognition*. Il sistema descritto in [40] è in grado di distinguere brani di musica classica, jazz, e musica pop con un tasso di errore medio pari al 45%, contro una *chance performance* del 67%, ad indicare che in questo campo c'è ancora molto spazio per i miglioramenti.

[1] Sebbene nell'ambito dell'Intelligenza Artificiale il riconoscimento, l'identificazione e la classificazione siano tre problemi distinti, nel presente lavoro questi termini verranno usati intercambiabilmente, ove non generino confusione.

Pitch tracking

Il problema della determinazione dell'altezza percepita di un singolo suono (*pitch detection*), di una melodia (*pitch tracking*), o di un brano di complessità arbitraria (*score tracking*) è studiato da tempo, ed ha come obiettivo, irraggiungibile nella sua accezione più generale persino da un essere umano, la trascrizione automatica di una partitura a partire dall'esecuzione. Strettamente collegato a questa applicazione, è il tentativo di separare le diverse sorgenti sonore (*sound source separation*, o *unmixing*) in altrettanti canali sonori. Come sarà precisato nella sezione 2.1, un traguardo così ambizioso, allo stato attuale, è da considerarsi purtroppo utopico. Simile, ma sicuramente più facilmente risolvibile, è il problema del *riconoscimento* delle sorgenti sonore (*sound source recognition*), eventualmente sovrapposte, in un brano audio. Il prodotto commerciale SoundFisher [105, 106], ad esempio, consente di classificare suoni di varia natura, come risate, versi di animali, applausi e squilli di telefono. Sia per le tecniche utilizzate che per le applicazioni conseguenti, questa tematica si mescola agli altri problemi di classificazione già presentati e con la più specifica classificazione di timbriche musicali, che è oggetto del presente lavoro, e che verrà approfondito nel capitolo 2.

Beat tracking

Inversamente ai sistemi di trascrizione automatica, in un certo senso, i *beat tracker*, nota la partitura, indicizzano il file audio di una sua esecuzione, individuando al suo interno il maggior numero di eventi possibile: inizio di battuta, ingressi di altri strumenti, singole note o accordi.

Onset e *offset detection*

Ad un livello di astrazione più basso si collocano le problematiche di *onset* e *offset detection*: determinare quando inizia e quando finisce un evento sonoro, può essere un compito non banale, e non solo in un contesto rumoroso o di audio complesso. Un metodo basato sulla sola intensità relativa può facilmente essere tratto in inganno dai cosiddetti *ghost onset* derivanti da tremolo e glissati. Per un sistema che utilizza le trasformate *wavelet*, si veda [49].

La ricerca nel campo dell'*audio content analysis* è ricca di applicazioni in diversi settori [26]: dall'intrattenimento all'ambito medicale-protesico, dalla sorveglianza e controllo ambientale al restauro di registrazioni, dalla comprensione di comandi vocali al giornalismo.

L'indicizzazione e successivo reperimento di multimedialità è forse l'aspetto più attuale e su cui le grandi aziende del settore stanno investendo, anche in vista di un ulteriore miglioramento e diffusione delle infrastrutture telematiche in tutto il mondo. Il *media annotation* si può rivelare utile anche in fase di *playback*: posizionarsi nel punto "dove entrano i fiati" non richiederà più ricerche estenuanti e imprecise. Si tenga presente inoltre che l'audio è parte inscindibile della totalità del materiale video in circolazione, e le relative tecniche di segmentazione ed indicizzazione, ad esempio, potrebbero basarsi anche sulla colonna sonora, che in alcuni casi (sparatorie, inseguimenti, etc.) è più facilmente analizzabile.

La comprensione della percezione umana delle varie forme di audio, può notoriamente portare ad una sua sensibile *compressione*, come già accade per il formato MPEG-1/layer III. Riuscire a trascrivere e codificare un brano musicale in un formato come quello dello Structured Audio ne consentirebbe la trasmissione a bassissimi *bitrate* [101]. La codifica automatica della partitura può essere facilmente ottenuta con i *pitch-tracker* dell'ultima generazione, con una polifonia di tre o quattro voci. Per quanto riguarda l'informazione relativa all'orchestrazione, con il linguaggio SAOL è possibile la realizzazione di tutti algoritmi di sintesi conosciuti [89, 93]. Sarà quindi sufficiente codificare, per ogni strumento, il particolare modello adottato e i relativi parametri. Il problema è che ad un algoritmo di sintesi non sempre corrisponde un algoritmo di analisi, o di stima dei parametri. Quand'anche questo sia vero, come per esempio nella sintesi additiva (FFT) o nella sintesi FM (coefficienti di Bessel), la stima sarebbe realizzabile solo nel semplice caso monofonico, e la resa finale non sarebbe eccezionale: un'orchestrazione di strumenti di sola sintesi additiva risulterebbe perlomeno monocorde... La sintesi *wavetable*, che associa ad ogni MIDI *key* un suono in formato PCM, è ampiamente supportata dallo standard MPEG-4 [92], e sembra essere un'idea migliore, specie nel caso in cui l'*encoder* e il *decoder* condividano gli stessi banchi di suoni. In questo modo, è infatti sufficiente inviare gli identificativi associati agli strumenti che compongono l'orchestra, senza necessariamente trasmettere anche i banchi, che comporterebbero un fastidioso costo fisso iniziale.

L'interesse commerciale verso questa tecnica di sintesi è stato recentemente sancito dall'armonizzazione degli standard di due istituzioni del settore (la MIDI Manufacturer's Association e la Creative Technology Ltd.) nel nuovo *open standard* Downloadable Samples-2 (DLS-2). Fra i parametri descrittivi dei suoni collegati alle MIDI *key*, si trovano quelli relativi al *looping*. La fase di sostegno degli strumenti, infatti, viene modellata mandando in *loop* un segmento accuratamente selezionato del suono; in seguito all'evento di *note off* viene eseguito il segmento finale. Questo modello della fase di sostegno è piuttosto povero, in quanto periodico, e costituisce un punto debole della sintesi *wavetable*.

Lo schema a blocchi di un sistema avveniristico, ma plausibile, di uno Structured Audio *encoder* basato su analisi e resintesi *wavetable* è riportato nella figura 1.2.

La trasmissione di audio strutturato ha inoltre il non indifferente vantaggio che il *rendering* della scena sonora viene effettuato dal ricevitore, permettendo all'utente un controllo più fine sulla riproduzione, che per la prima volta non si limita più alla manopola del volume e a un equalizzatore grafico.

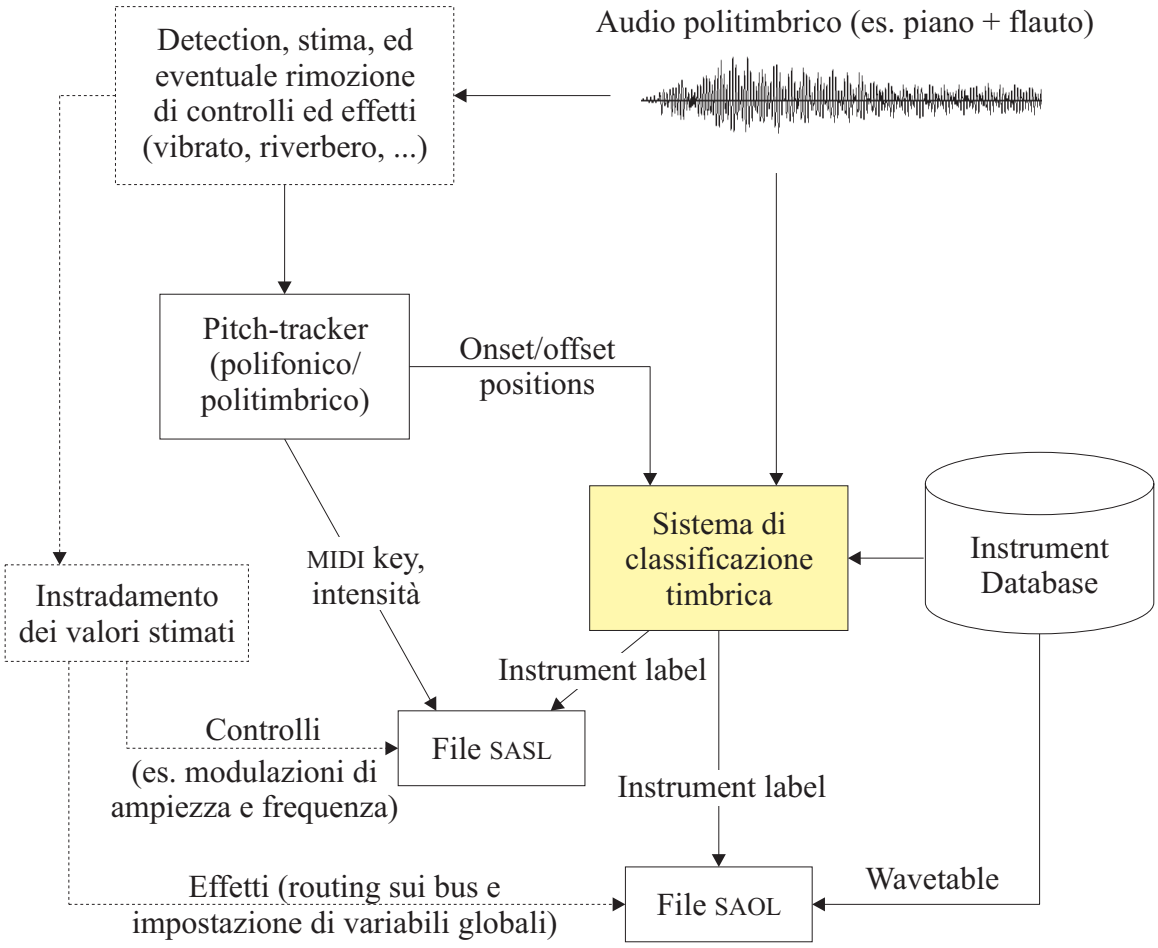


Figura 1.2. Una possibile integrazione del sistema di riconoscimento timbrico all'interno di un *encoder* per flussi Structured Audio. I blocchi tratteggiati rappresentano moduli di futuribile realizzazione.

Da ultimo, si vuole porre l'accento sulle potenzialità che sistemi di questo tipo hanno in ambiti più settoriali, ma altrettanto importanti. Ad esempio, un compositore potrebbe trarre vantaggio da strumenti di *authoring* in grado di scegliere un determinato timbro musicale attraverso interrogazioni in linguaggio naturale, o per somiglianza con altri, o in modo da creare un maggiore contrasto. Anche gli strumenti di *audio editing* trarrebbero notevole giovamento dall'introduzione di indicizzazioni basate sul contenuto.

1.2 Organizzazione del lavoro

Nel capitolo 2 vengono esaminati i principali studi sulla percezione del timbro, le varie scuole di pensiero e le diverse realizzazioni di sistemi automatici di riconoscimento del timbro. Nel capitolo 3 saranno analizzate le tecniche statistiche di classificazione di base che sono state adottate nella realizzazione del sistema, la cui architettura è presentata in dettaglio nel capitolo 4. Il capitolo 5 tratta più da vicino i vari aspetti dell'ingegnerizzazione dell'applicazione realizzata, e nel capitolo 6 vengono esposti i risultati, le conclusioni, e le estensioni future del sistema e dell'intero progetto.

capitolo 2

Analisi critica della letteratura

2.1 Il timbro, un concetto sfumato

L'ente di standardizzazione americano ANSI, definisce in questo modo [2] i tre principali attributi che caratterizzano la percezione di un suono:

Acutezza (*pitch*)	Attributo della sensazione uditiva grazie al quale i suoni possono essere ordinati su una scala da acuto (*high*) a grave (*low*)
Livello sonoro (*loudness*)	Attributo dell'intensità della sensazione uditiva grazie al quale i suoni possono essere ordinati su una scala da piano (*soft*) a forte (*loud*)
Timbro	Attributo della sensazione uditiva grazie al quale un ascoltatore può giudicare che due suoni, presentati nello stesso modo, e aventi pari *loudness* e *pitch*, sono diversi

Il fatto che un istituto di standardizzazione del calibro dell'ANSI sia dovuto ricorrere ad una definizione "per complemento," dà un'idea della vaghezza del concetto di timbro[1].

Il *pitch* di un suono dipende principalmente dalla sua frequenza[2], e il livello sonoro dall'ampiezza della sollecitazione. Il timbro, invece, è caratte-

[1] In [65] si legge che il timbro è il "cestino multidimensionale" degli psicoacustici.

[2] Persino il *pitch* dei segnali sinusoidali (*simple tone*) dipende da altri fattori, quali l'intensità, la durata e inviluppo [46].

Attributo percettivo	Grandezza fisica
Durata	Tempo (ms)
Pitch	Frequenza (Hz)
Loudness	Ampiezza (dB)
Timbro	??

Tabella 2.1. Principali grandezze fisiche associate agli attributi percettivi.

rizzato da una moltitudine di fattori fisici, tra cui il contenuto spettrale e il relativo inviluppo (che Helmholtz già nel 1863 [43] aveva individuato come determinanti), ma anche la differenza di fase tra le diverse armoniche, o il grado di inarmonicità. Le percezioni di queste tre caratteristiche del suono, a cui si aggiunge la durata (tabella 2.1), si influenzano vicendevolmente: basta modulare periodicamente il *pitch* o modificare l'inviluppo di un suono per renderne il timbro irriconoscibile; il livello sonoro è influenzato anche dal contenuto spettrale e dalla durata dello stimolo.

Per comprendere quanto sia complesso e mal posto il problema del riconoscimento del timbro, e più in generale del problema della separazione delle sorgenti sonore, basti pensare alle quotidiane esperienze di ascolto, specie in ambito musicale. A volte gli strumenti sono così amalgamati che anche un "orecchio esperto" fatica a distinguerli: una sezione di fiati che intona una triade maggiore è facilmente percepito come unico coerente evento sonoro [5], per non parlare di un "tutto" orchestrale. In [85] vengono approfonditamente analizzati i fattori che determinano questo effetto di "fusione" timbrica (*blend*).

La parola "timbro," quindi, assume diversi significati per diverse persone in diversi contesti, tant'è che per descriverne le proprietà si utilizzano parole colorite[3] come "metallico," "nasale," "brillante," e c'è chi, addirittura, vorrebbe che fosse "espunta dal vocabolario dell'acustica" [62].

Si potrebbe continuare citando paradossi percettivi [96], o sconfinando in considerazioni ermeneutiche che mettono in discussione il concetto stesso di suono, o di musica (si ricordi il celebre pezzo 4'33" di John Cage, per qualsiasi strumento o insieme di strumenti), ma sarebbe oltre gli scopi di questo scritto.

[3] In tedesco "timbro" si traduce con "*Klangfarbe*," letteralmente "colore del suono."

2.2 Computational Auditory Scene Analysis

La letteratura è ricca di ricerche basate sull'ambiguo concetto di "evento sonoro," che porta ai problemi riportati nel paragrafo precedente, e si traducono inevitabilmente in sistemi il cui tasso di errore si alza drasticamente quando i test vengono effettuati su brani di audio complesso e affetto da rumore, ad esempio una registrazione radiofonica in modulazione d'ampiezza.

Per far fronte a queste difficoltà, un gruppo di ricercatori del Massachussets Institute of Technology, sotto la guida del Prof. Barry L. Vercoe, ha creato un nuovo modo di intendere l'esperienza di ascolto, basato sul manifesto "Auditory Scene Analysis," scritto nel 1990 da Albert Bregman [5], dove sono raccolte numerose considerazioni di carattere psicoacustico relative a questo processo percettivo. L'ascolto non è più visto riduzionisticamente come una semplice "rappresentazione interna" dello spartito, ma come un processo olistico e "d'insieme." Il punto focale si sposta sul concetto di *stream*, ovvero "un'organizzazione psicologica rappresentante una sequenza di eventi acustici internamente coerenti proveniente da una o più fonti sonore" [65].

Si è così formata la disciplina denominata *Computational Auditory Scene Analysis* CASA, che cerca di simulare all'elaboratore i modelli psicoacustici suggeriti in letteratura.

Un apporto fondamentale alla comunità scientifica è rappresentato dalla realizzazione di modelli dell'apparato uditivo umano sempre più fedeli, tanto da costituire per le ricerche attuali una valida alternativa a sistemi di analisi e rappresentazione del segnale più tradizionali come le trasformate ondine e di Fourier. L'argomento verrà trattato più approfonditamente nel paragrafo 2.4.

Il lavoro di Ellis [18] è centrale nell'ambito CASA. Viene introdotto l'approccio *prediction-driven*, in cui il passaggio dall'audio alla sua percezione non avviene più attraverso le tradizionali fasi (dette *bottom-up*, o *data-driven*) di segmentazione in eventi (note, rumori, fonemi, etc.), relativa caratterizzazione (*pitch*, *loudness*, timbro, ritmo, etc.) e integrazione in morfemi di più alto livello (melodie, armonie, parole, etc.), ma è il prodotto di un sistema reazionato [17] il cui passaggio fondamentale è la riconciliazione tra lo stato attuale del modello predittivo adottato e l'informazione proveniente dall'analisi psicoacustica del segnale in ingresso. In altre parole, la percezione uditiva, e più in generale la percezione umana, è considerata come l'incontro tra le informazioni derivanti dai sensi e dalle previsioni. Per queste ragioni, il problema della trascrizione automatica non viene più visto come passaggio obbligato per la comprensione dell'audio, ma come un separato, ancorché stimolante, problema ingegneristico [64].

Per quanto riguarda più strettamente il riconoscimento delle timbriche degli strumenti musicali, l'opera di Martin [62], a cui il presente lavoro sotto

certi aspetti si ispira, poggia su numerosi risultati psicologici e psicoacustici, che lo hanno portato a formulare nuove proprietà caratteristiche dei timbri (elencati e commentati nel paragrafo 2.5), e ad adottare un modello gerarchico per il classificatore realizzato.

Per ulteriori approfondimenti sul tema, si consigliano le tesi di Scheirer [86, 88] e i modelli proposti in [51, 54, 97].

2.3 Studi psicoacustici sulla percezione del timbro

Come accennato, già dai primi studi di von Helmholtz del XIX secolo, fortemente motivati dai risultati di Fourier, appariva evidente come le timbriche degli strumenti musicali tradizionali fossero in buona parte influenzate dallo spettro del segmento di sostegno[4] delle note, in cui il suono si avvicina maggiormente alla definizione di segnale periodico. In effetti, si sosteneva che il timbro dipendesse unicamente dal numero di parziali[5] presenti, e dal rapporto delle loro ampiezze. Solo nel secolo successivo fu ripresa in considerazione la differenza di fase, che fino ad allora si riteneva essere ininfluente ai fini della discriminazione dei suoni. Gli esperimenti citati in [33, 81] dimostrarono che effettivamente l'orecchio umano non è completamente sordo al mutamento della fase delle diverse armoniche, ma che l'effetto è così poco percepibile che la sua rilevanza scompare in una camera riverberante, in cui le relazioni di fase vengono sconvolte.

Si è ottimisticamente ritenuto a lungo che le note musicali (*tones*) fossero periodiche, perlomeno per buona parte della loro durata, e così negli studi di inizio secolo si effettuava una semplice media dello spettro. I limiti di un simile approccio apparirono evidenti con i primi tentativi di sintesi sonora, in quanto si scoprì che i segnali puramente periodici suonano "piatti" e innaturali. Si consideri, ancora, uno strumento che suona in una camera riverberante, dove la risposta in frequenza, e non solo la fase, varia enormemente da punto a punto. Il fatto che un essere umano riconosca lo strumento da qualsiasi punto della sala, induce a pensare che lo spettro non sia l'unica caratteristica discriminante. Altre proprietà del timbro vanno ricercate nella

[4] L'inviluppo della forma d'onda di una nota si suddivide tipicamente in quattro fasi, chiamate rispettivamente attacco, decadimento, sostegno e rilascio.

[5] Una parziale è una componente dominante dello spettro di un suono, la cui frequenza può essere o non essere multiplo intero della frequenza fondamentale. In realtà, parlando di note suonate da strumenti musicali, si dovrebbe per la precisione parlare di "parziali," più che di "armoniche."

ricchezza di micro-cambiamenti di una moltitudine di parametri, durante l'esecuzione della nota, che difficilmente sono ottenibili a partire da una analisi tradizionale.

Mentre appare ormai chiaro che l'informazione timbrica non può essere catturata semplicemente effettuando un'analisi spettrale di un segmento della fase di sostegno, non sembra esserci un accordo tra i vari studiosi rispetto alla rilevanza relativa dei diversi segmenti, in particolare quelli di attacco e sostegno. In alcuni studi si sostiene che la fase di attacco è necessaria e sufficiente per l'identificazione, mentre in altri, anche più recenti, la stessa cosa è detta della fase di sostegno [39, 48, 52]. Il vero problema, a parere dell'autore, è che si fa sentire sempre più pressante la necessità di un modello in grado di rappresentare adeguatamente l'evoluzione temporale delle caratteristiche spettrali. Mentre in altri campi, quali il *pitch-tracking*, sono state applicate con successo tecniche di *bayesian modeling* [102], sembra che il numero di fattori in gioco in campo timbrico vanifichi gli sforzi in questa direzione, tanto che i principali sistemi di classificazione timbrica, analizzati nel paragrafo 2.6, si basano tutti su un canonico approccio di *pattern-matching*, pur utilizzando le tecniche più varie.

Negli anni 70 gli sforzi delle ricerche si intensificarono, principalmente in due direzioni: l'analisi più accurata di note isolate, e studi percettivi sulla capacità di discriminare timbriche differenti. Questi aspetti, investigati tipicamente negli strumenti musicali della tradizione musicale occidentale, ma estendibili facilmente a qualsiasi tipo di sorgente sonora quasi-periodica, vengono approfonditi nei due paragrafi seguenti.

2.3.1 Studio spettrografico di note isolate

Sicuramente anche grazie alla spinta della disponibilità di metodi ed apparecchiature digitali, una serie di articoli mirati allo studio degli spettrogrammi dei diversi strumenti musicali (perlopiù di note isolate) popolarono i periodici specializzati [55, 72, 73, 79, 80]. Ne scaturì un *corpus* non indifferente di osservazioni e scoperte, spesso eterogenee e di difficile generalizzazione, relative ai singoli strumenti, di cui viene proposto di seguito un piccolo campionario.

Pianoforte Le parziali di ordine inferiore presentano un'attenuazione esponenziale, mentre quelle di ordine superiore appaiono di ampiezza costante dopo i primi 50 millisecondi circa.

Alcune parziali sembrano smorzarsi completamente, ma riappaiono successivamente nella nota.

Le corde in oscillazione libera, in generale, danno luogo a suoni inarmonici, ovvero suoni in cui le parziali di ordine superiore occupano frequenze che si scostano dai rispettivi multipli interi della fondamentale, a causa della rigidezza meccanica. Le note basse del pianoforte godono di un grado di inarmonicità particolarmente elevato, che conferisce loro un timbro particolarmente "caldo."

Violino Durante la fase di attacco della nota, le frequenze delle varie parziali sono erratiche, probabilmante a causa di alcune discontinuità di fase.

Gli archi pizzicati presentano un suono impulsivo che si attenua esponenzialmente con una costante di tempo inversamente proporzionale all'accoppiamento tra la corda e il corpo risonante. Se eccitati con l'archetto, invece, la fase di attacco risulta molto lenta. Il contenuto in frequenza dipende dalla pressione esercitata con l'archetto, e della sua posizione rispetto al ponticello.

Ottoni Nell'ultima parte della fase di attacco si osserva una rapida ed irregolare oscillazione nell'ampiezza delle parziali, dovuta al lasco accoppiamento tra le labbra dell'esecutore e lo strumento. Il fenomeno è più accentuato se lo strumento è suonato con maggiore intensità, ed è presente in particolare nelle parziali di ordine superiore. Nel corno francese è presente più di una oscillazione.

Nei "pianissimo" il suono è pressoché sinusoidale, e ad intensità crescenti corrisponde una forma d'onda più vicina a quella impulsiva.

Tromba Le parziali di ordine superiore compaiono più tardi nello spettrogramma.

Xilofono Si registra la presenza di una banda di rumore in bassa frequenza che sembra accompagnare la fase di attacco

Flauto Le ampie fluttuazioni in ampiezza (*tremolo*) sono accompagnate da fluttuazioni in frequenza (*vibrato*).

Clarinetto Le parziali di ordine dispari sono molto più accentuati rispetto a quelli di ordine pari. Più in dettaglio, questo vale al di sotto di una frequenza di *cut-off*, attorno ai 3 kHz per un clarinetto in si ♭, e la differenza è più accentuata nel registro superiore. Per note al di sopra di questa frequenza, viceversa, l'energia delle parziali pari e dispari è comparabile.

Considerazioni come l'ultima riportata sono note da tempo, e in realtà queste proprietà acustiche sono alla base dell'antica arte della produzione di strumenti tradizionali. Una di esse ha una particolare importanza per l'identificazione degli strumenti musicali da parte dell'orecchio umano: la collocazione delle *formanti*, ovvero di picchi di risonanza, o aree di accentuazione spettrale, poste a frequenze fisse, indipendentemente dalla particolare nota suonata. Ad esempio, l'oboe è caratterizzato da due formanti principali, a circa 1 kHz e 3 kHz, rispettivamente, separate da una regione di "anti-risonanza" intorno ai 2100 Hz. Sfortunatamente, codificare questo tipo di informazione in un classificatore automatico, basandosi sulla registrazione di un insieme di note, è molto difficile.

2.3.2 Distanze soggettive tra strumenti musicali

Gli studi di Grey, Gordon e Wessel verso la fine degli anni 70 [35, 36, 103] mirarono all'isolamento dei fattori discriminanti tra i vari strumenti musicali, cercando di generalizzare i risultati sopra riportati in caratteristiche ad un più elevato livello di astrazione. Le tecnica maggiormente impiegata, e che ottenne i risultati più convincenti, fu la *taratura multidimensionale* (*multidimensional scaling*).

Il *multidimensional scaling* è un metodo della statistica multivariata messo a punto negli anni 60 da Kruskal, che ben si adatta allo studio di fenomeni complessi, come quello della percezione del timbro. Il ricercatore, a partire da un insieme di registrazioni di note suonate da diversi strumenti, propone ad una serie di ascoltatori tutte le possibili coppie di suoni, e raccoglie le distanze soggettive percepite[6] in una matrice. A questo punto i singoli stimoli possono essere proiettati in uno spazio avente un numero di dimensioni arbitrario, in modo che la separazione tra gli stimoli sia massima. Grey adottò uno spazio tridimensionale, e riuscì ad associare ai tre assi le seguenti proprietà.

Asse I Distribuzione dell'energia spettrale. Strumenti come il corno francese e il sassofono soprano presentano un'ampiezza di banda molto più stretta rispetto, ad esempio, al trombone. Questa proprietà in letteratura è stata più tardi reinterpretata con il nome di *centroide spettrale*.

[6] Queste distanze possono essere calcolate in base alla capacità degli ascoltatori di distinguere gli stimoli, o di associare ad essi dei nomi.

Asse II Sincronia dell'attacco e decadimento delle diverse parziali. Collegato a questo aspetto è la quantità complessiva di fluttuazioni spettrali durante l'intera esecuzione della nota. La famiglia dei legni sembra avere elevati valori di allineamento.

Asse III Lungo questo asse sono raccolte le differenze in termini di evoluzione temporale dell'energia spettrale, come la presenza di perturbazioni inarmoniche nella fase di attacco.

Successivamente, ci furono diversi tentativi di miglioramento di questa tecnica attraverso l'*analisi-resintesi* [37, 38, 104], che in buona parte confermarono i risultati ottenuti. Questi studi effettuavano una analisi di Fourier dei campioni e ne modificavano i parametri prima di rigenerare gli stimoli attraverso sintesi additiva. L'obiettivo era quello di isolare le modifiche che portano alla percezione di due timbriche diverse. Per studi recenti di questo tipo, si vedano [47, 100].

Modelli multidimensionali come questi sono intrinsecamente limitati dal fatto che, come è stato dimostrato da numerosi studi psicologici, lo spazio percettivo non è metrico, e quindi alcuni autori [95] criticano come forzature le interpretazioni fisiche di queste dimensioni.

Questi esperimenti ebbero anche il merito di mettere in luce caratteristiche macroscopiche, ma spesso sottovalutate, che favoriscono la discriminazione. Ad esempio, l'orecchio umano sembra fortemente influenzato dal grado di percussività dei suoni, e dalle modulazioni parallele (periodiche o aperiodiche) dei formanti, sia in ampiezza che in frequenza. Conseguentemente, misurare l'ampiezza e la frequenza del *vibrato* (oscillazioni in frequenza) e del *tremolo* (oscillazioni in ampiezza) può rivelarsi un buon inizio per un sistema di classificazione. Lo stesso può dirsi per il *jitter* (modulazione casuale in frequenza), presente ad esempio nella tromba e negli archi. La conoscenza dell'estensione dello strumento musicale è un'indicazione preziosa per il nostro sistema uditivo, basti pensare a quanto inusuale potrebbe sembrare un do 2 suonato da un violino! Questo dovrebbe indurre il ricercatore ad assicurarsi che il soggetto (sia esso un essere umano, o un sistema informatico) abbia avuto precedente esperienza di un campione rappresentativo di note, per evitare grossolani errori di classificazione.

Molti autori hanno sottolineato inoltre l'importanza del *contesto* in cui le note vengono suonate: le capacità degli ascoltatori di riconoscere uno strumento aumentano sensibilmente se i campioni diventano, da singole note, intervalli di note legate o semplici melodie [52]. La fase di *transizione*, o *articolazione*, quindi, è depositaria di una cospicua quantità di informazione, che, per buona parte, rimane tuttora inesplorata.

2.4 Scelta del *front-end*

Il primo passo che un qualsiasi sistema di comprensione dell'audio deve effettuare è trasformare i dati in ingresso, si supponga audio PCM, in una rappresentazione (o *front-end*) più adatta allo scopo.

Tradizionalmente, è stato utilizzata la trasformata di Fourier a tempo discreto[7] (*Discrete-time Fourier Transform*, DFT). Negli ultimi tempi, questa rappresentazione ha dimostrato dei limiti intrinseci, ed è stata criticata come base di queste applicazioni [1]. Ad esempio, è noto che la risoluzione in frequenza e quella temporale sono legate da un rapporto di inversa proporzionalità (*incertezza di Fourier*), ed è quindi necessario un *trade-off* tra le due proprietà. Inoltre, quando si ha a che fare con sorgenti complesse, le cose si complicano immediatamente: già con un bicordo di un intervallo giusto, le parziali tendono a sovrapporsi, e la separazione delle due note diventa un compito non banale, anche perché lo spettro risultante dipende dalle fasi delle singole note, a cui l'orecchio umano è praticamente insensibile. La rappresentazione di Fourier si rivela inadeguata per sorgenti immerse in scene sonore rumorose e in generale acusticamente "ostili," come in presenza di strumenti a percussione ad ampio spettro (si pensi ad esempio ai piatti), dal momento che l'informazione dovuta al rumore viene spalmata su tutto lo spettro, rendendo difficile l'individuazione di questi eventi. Anche per compiti relativamente più semplici, quali la risoluzione della frequenza fondamentale per segnali che ne sono privi (ad esempio la voce umana trasmessa attraverso una linea telefonica [78], o le note di alcuni strumenti, come il fagotto), la DFT non sembra il sistema di rappresentazione più indicato.

Alcuni autori [4, 7, 99] hanno affrontato il problema della risoluzione in frequenza, cercando di fornire un *front-end* che avesse una risoluzione *logaritmica*, tenendo conto della "geometricità" della sensibilità cocleare all'altezza dei suoni. Brown [7] sostiene che la trasformata a Q costante (ovvero avente un costante rapporto tra frequenza centrale e intervallo di risoluzione) faciliti i compiti di *pitch-detection*, di riconoscimento degli strumenti, e in generale di separazione dei segnali, ma riconosce che anche in questo caso si è di fronte a compromessi di risoluzione tempo-frequenza.

Col tempo sono state proposte altre trasformate, con alterni successi. Si veda ad esempio [6] per un'applicazione di spettri del second'ordine e di

[7] La stessa trasformata, prende in letteratura nomi diversi, anche in base alla particolare implementazione adottata: *Short-Time Fourier Transform*, *Fast Fourier Transform* e *Phase Vocoder*.

ordine superiore a segnali musicali. Le trasformate ondine (*wavelet*) risolvono in buona parte i limiti della DFT, ma vengono spesso ignorate per la mancanza di implementazioni efficienti o di pacchetti specifici che le supportino.

Riassumendo, si elencano le qualità auspicabili da una rappresentazione ideale di livello intermedio [61]:

1. Separabilità delle sorgenti sonore. Man mano che la rappresentazione aumenta in livello di astrazione, ci si aspetta che ai suoi elementi corrispondano sempre più gli eventi sonori.

2. Invertibilità. È necessario che sia possibile effettuare a ritroso il percorso di analisi, ovvero sintetizzare un segnale che sia percettivamente simile a quello originale.

3. Plausibilità psicologica. È desiderabile che la rappresentazione sia il più possibile simile a quella adottata dagli esseri umani. Questo per due motivi: per ottenere risultati più vicini alla nostra percezione, e per accostarsi ai diversi problemi con un sistema che sembra risolverli con "naturalezza."

4. Robustezza rispetto al rumore di fondo o al numero di sorgenti sonore.

Con questi obiettivi in mente, agli inizi degli anni 90 ci fu un fiorire di articoli [67, 68, 107], in cui ogni studioso proponeva il proprio modello dell'apparato uditivo umano, perlopiù derivante dai precedenti studi di Licklider (1951). In particolare, la pubblicazione di Meddis e Hewitt è quella che in questi anni ha riscosso maggiore successo ed è stata onorata da più di una implementazione all'elaboratore, e verrà di seguito descritta per sommi capi. Il modello è composto di una serie di fasi di elaborazione disposte in cascata, attravverso le quali passa il segnale.

Filtro dell'orecchio esterno e medio	Filtro passabanda che simula il guadagno delle frequenze centrali alle quali il nostro orecchio è più sensibile.
Filtro timpanico	Una serie di filtri (*gammatone filterbank*) simulano l'effetto del timpano sul segnale, separandolo in canali caratterizzati da una propria frequenza centrale e disposti con risoluzione quasi-logaritmica sull'asse della frequenza. Sollecitato dalle onde di pressione, il timpano comunica il movimento agli ossicini dell'orecchio interno, che a loro volta muovono il liquido contenuto nella coclea.
Trasduzione dell'energia	Nella coclea risiedono le cellule ciliate, responsabili della trasduzione dell'energia meccanica in energia elettrica

trasmessa al sistema nervoso umano. Il modello, basato su un sistema non-deterministico a tempo discreto reazionato, emette un quanto di energia (in inglese, *fires*) in un determinato canale con una probabilità che cresce al crescere dell'energia meccanica dello stesso canale, e decresce in prossimità di recenti eventi di innesco.

Calcolo cumulativo Gli intervalli che separano le emissioni delle cellule ciliate vengono cumulate in un istogramma per ogni canale, attraverso il calcolo di funzioni di autocorrelazione. La rappresentazione che ne deriva, chiamata *correlogramma*, tiene conto, nell'originale implementazione di Meddis e Hewitt, degli ultimi 7,5 ms di segnale in ingresso. Questo passaggio è il più delicato, oltre che il più oneroso computazionalmente, in quanto realizza il passaggio più oscuro e più intimamente collegato alla percezione del *pitch*.

È stato dimostrato [67] che un sistema di questo tipo fornisce interpretazioni più attendibili e immediate per alcuni fenomeni che mettono in crisi la DFT, come il problema della fondamentale mancante, di parziali irrisolte causate dall'alto grado di inarmonicità, il *pitch-tracking* di note glissate o con vibrato accentuato e il riconoscimento di accordi. Esso infatti codifica automaticamente informazioni che tradizionalmente si trovano nei livelli più alti della rappresentazione, come la struttura formantica e le "proprietà di gruppo" delle parziali.

I correlogrammi sono stati adottati come *front-end* per numerosi sistemi di CASA, fra i quali [18, 62]. Il sistema di Ellis copre frequenze da 100 Hz a 10 kHz, con una dinamica di 16 bit e una frequenza di campionamento del segnale originario pari a 22,05 kHz. Martin ha sensibilmante migliorato questi limiti. Slaney [94] ha reso disponibile un *toolbox* per MATLAB che realizza i principali modelli computazionali dell'orecchio umano. Purtroppo questi sistemi sono ancora prototipali, e non raggiungono prestazioni soddisfacenti per applicazioni in tempo reale. I calcoli sono infatti molto onerosi, anche se si adattano facilmente ad essere implementati su architetture parallele [61].

Ad Ellis va anche il merito di avere introdotto il concetto di *weft* [19], una rappresentazione strettamente legata al correlogramma, che realizza il livello successivo di astrazione, ovvero quello di "oggetto sonoro" quasi-periodico intonato. Può essere legato ad una singola nota, come a un accordo: a qualsiasi suono caratterizzato da una propria "coerenza interna," percepito dall'orecchio umano come un'unica sorgente sonora.

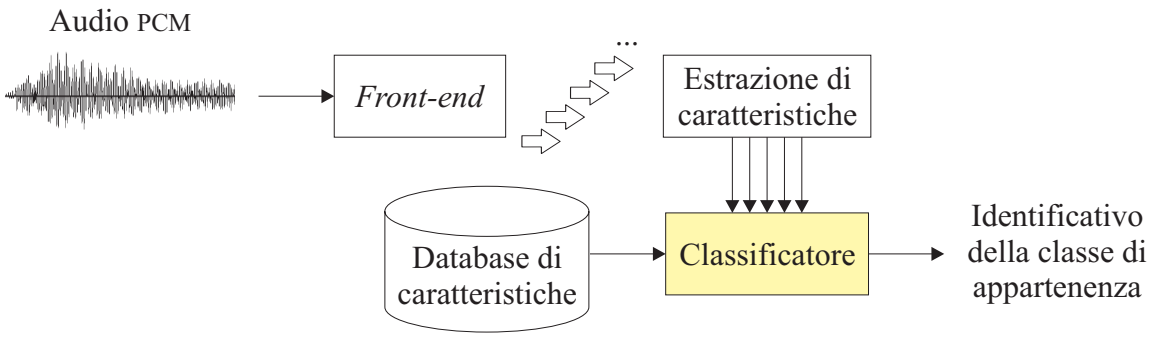


Figura 2.1. Schema a blocchi dell'approccio *pattern recognition* al problema della classificazione automatica.

2.5 Estrazione di caratteristiche significative

Tutti i sistemi di riconoscimento timbrico che saranno analizzati nella sezione 2.6 adottano un approccio di *pattern-recognition* più o meno tradizionale. Tranne alcune eccezioni, in cui si sono utilizzate reti neurali a mappa autoorganizzante, è stata realizzata una classificazione di tipo *assistito*, ovvero al classificatore è stato fornito l'identificativo dello strumento musicale associato alle note o ai brani con i quali il sistema è stato addestrato. Affinchè questi sistemi non vengano sommersi dall'enorme quantità di dati in uscita dal *front-end*, è necessario calcolare, a partire da questi, un numero più limitato di caratteristiche (*features*) per ogni nota, o brano, che si intende classificare, o con cui si intende addestrare il sistema (figura 2.1). Questi passaggi sono necessari anche in considerazione del fatto che i dati grezzi (*raw*) che rappresentano l'ampiezza istantanea della forma d'onda non sono in alcun modo collegabili direttamente alla sorgente, variando enormemente persino tra due registrazioni successive nelle stesse condizioni.

È importante che i valori delle caratteristiche considerate per un determinato strumento siano quanto più invarianti possibile rispetto al particolare esecutore, ambiente di registrazione, impianto microfonico, e realizzazione dello strumento stesso: idealmente, uno Stradivari suonato da Uto Ughi in sala di registrazione dovrebbe essere riconoscibile come violino quanto quello di un *clochard* in una metropolitana. Per raggiungere un obiettivo così ambizioso, appare chiaro da subito che conviene attenersi a caratteristiche che siano strettamente legate alle proprietà fisiche dello strumento e alla percezione umana di queste proprietà. In questo senso, utilizzare un *front-end* appositamente studiato per questi scopi rappresenta un notevole vantaggio.

È tuttavia altrettanto importante che questi valori consentano una sufficiente separazione tra le varie classi. Con questo non si vuol dire che ogni caratteristica debba essere rilevante per ogni timbro. Infatti, come sarà più chiaro dopo la lettura del capitolo 3, la salienza delle caratteristiche dipende dal particolare contesto.

Di seguito sono esposte le caratteristiche considerate a questo scopo in letteratura.

2.5.1 Caratteristiche del *pitch*

La percezione dell'altezza del suono e della sua evoluzione temporale fornisce numerosi spunti per il processo di riconoscimento.

Pitch — L'attributo del *pitch* può essere diversamente interpretato, a seconda che le sorgenti da identificare abbiano un'estensione ridotta (ad esempio miagolii, tonfi, etc.) oppure, come nel caso degli strumenti musicali, le frequenze fondamentali possono variare di una o più ottave. Nel primo caso, il *pitch* può essere utilizzato come variabile ordinaria, altrimenti si delinea il concetto di *pitch range*, che va trattato diversamente. Infatti, la frequenza delle note suonate da un generico strumento è distribuita uniformemente all'interno di questo intervallo, e quindi non soddisfa l'ipotesi di base di distribuzione normale delle caratteristiche, alla base dei principali metodi di classificazione.

Vibrato — I parametri delle oscillazioni periodiche del *pitch* provocate volontariamente in alcuni strumenti musicali e nella voce umana sono poco caratterizzanti, in quanto la loro ampiezza (o *profondità* del vibrato) e la loro frequenza variano molto di più in funzione dell'esecutore o della particolare esecuzione, più che in base alla classe di appartenenza dello strumento. Questa caratteristica, però, viene solitamente considerata congiuntamente ad altre *feature* ad essa strettamente correlate, come i parametri di oscillazione del centroide o del tremolo.

Jitter — L'ampiezza delle modulazioni in frequenza casuali, o *jitter*, è indice della qualità dell'accoppiamento meccanico tra la sorgente dell'eccitazione e il corpo vibrante.

Portamento Citando da *La nuova enciclopedia della musica* Garzanti [31]:

> Il portamento è una tecnica di esecuzione che consiste nel passare rapidamente da un suono a un altro sfiorando rapidamente tutti, o per buona parte, i suoni intermedi. Ha largo impiego nel canto e negli strumenti ad arco, ma viene praticato anche negli strumento a pizzico e a fiato.

2.5.2 Caratteristiche dell'inviluppo

L'inviluppo dell'ampiezza istantanea si compone tradizionalmente di almeno quattro fasi: attacco, decadimento, sostegno e rilascio. Esse sono state introdotte con le prime tecniche di sintesi, in modo da introdurre una semplice modulazione non periodica, che simulasse quella degli strumenti musicali tradizionali. Trattandosi di un modello, identificarne i parametri, cioè le durate, è, nel migliore dei casi, difficile, se non una forzatura. I tempi di attacco variano enormemente da strumento a strumento, da esecutore a esecutore, e persino da nota a nota, quindi l'estrazione di un frammento di durata fissa appare un tentativo alquanto goffo per l'isolamento delle varie fasi. Valide alternative possono essere una attenta rilevazione dei picchi in ampiezza, o il monitoraggio del tasso di *zero crossing*, ovvero il passaggio dell'ampiezza istantanea per il valore nullo.

Alcune classi di caratteristiche (ad esempio quelle relative al *pitch* o allo spettro) andrebbero valutate solo per la fase di sostegno, ma le problematiche relative alla segmentazione automatica dell'inviluppo suggeriscono di evitarla, e valutare queste caratteristiche su tutta la durata della nota o del brano, in quanto si assume che la durata della fase di sostegno sia dominante sulle altre fasi e sulle eventuali pause, e quindi queste ultime non influenzano drammaticamente i risultati.

Durata delle diverse fasi Ammesso di riuscire ad ottenere valori accurati di queste quantità (generalmente ci si accontenta dei valori relativi all'attacco e al sostegno), esse sono indice della "forma" dell'inviluppo, e possono essere variamente combinate fra loro dando luogo a nuove *feature*.

Pendenza delle diverse fasi Assieme ai valori del punto precedente, specie quello relativo all'attacco, possono dare una misura del grado di percussività del suono.

Oscillazioni nella fase di attacco Sono, come detto, caratteristiche di alcune classi di strumenti musicali.

Tremolo L'ampiezza e la frequenza delle oscillazioni in ampiezza, al contrario di quelle del vibrato, possono essere discriminanti di per sé. Il flauto, per esempio, è lo strumento che presenta in assoluto le ampiezze maggiori.

2.5.3 Caratteristiche dello spettro

Molte delle caratteristiche appartenenti a questa classe si basano sull'ampiezza delle parziali. Il problema della risoluzione delle parziali per suoni quasi-periodici consiste nel determinare le frequenze della fondamentale e delle parziali di ordine superiore, che possono non coincidere con le frequenze multiple intere della fondamentale, e le relative ampiezze. Esistono diverse soluzioni, di cui è ricca la letteratura relativa al *pitch-tracking*, che rivelano quanto delicata sia la questione. Per evitare che i valori dipendano dall'intensità del suono o dalla qualità della registrazione, si usa generalmente normalizzarli, dividendoli per il valore massimo, che si ricorda può non coincidere con quello della fondamentale.

Centroide spettrale Somma pesata delle frequenze spettrali. È di gran lunga la caratteristica più utilizzata, ed è storicamente interpretata come indice della "brillantezza" del suono. Spesso questa quantità è posta in rapporto con il *pitch* (centroide relativo). Purtroppo, il calcolo del centroide secondo questa definizione è facilmente inquinabile in caso di audio polifonico o rumoroso.

Ampiezza delle parziali L'importanza delle singole parziali viene analizzata fino a che la risoluzione in frequenza lo consente—per le rappresentazioni che adottano una risoluzione logaritmica sull'asse delle frequenze, si tratta di un limite intrinseco. Generalmente, è sufficiente risolvere fino alla sesta o alla settima parziale. Per le parziali successive, ha più senso utilizzare delle proprietà di gruppo, coerentemente con gli studi riportati in [10].

Inarmonicità Un elevato scostamento delle parziali dalle frequenze armoniche (multiple intere della fondamentale) è caratteristico della vibrazione delle corde libere.

Ampiezza di banda È calcolata come la media pesata delle differenze tra le componenti spettrali e il centroide.

Rapporto tra parziali pari e dispari Come insegna l'esperienza relativa al clarinetto, può essere utile confrontare gli apporti delle parziali pari rispetto a quelle dispari.

Irregolarità spettrale Sono state impiegati diversi indici per calcolare questa quantità. Una scelta possibile consiste nella differenza media tra l'ampiezza di una parziale e quella delle due parziali successive [62].

2.5.4 Altre caratteristiche

Oscillazioni del centroide L'ampiezza delle oscillazioni periodiche del centroide sono generalmente indotte dal fenomeno del vibrato. Oltre all'ampiezza e alla frequenza assolute, quindi, è di interesse il rapporto con le rispettive grandezze legate al vibrato, nonché la fase che separa le due oscillazioni.

Relazioni fra tremolo e vibrato Per lo stesso motivo, in alcuni sistemi si prendono in considerazione i rapporti tra le frequenze e le ampiezze delle modulazioni in ampiezza e in frequenza, e la fase che le separa. Le stesse grandezze possono essere calcolate separatamente per le diverse parziali.

Spettro formantico L'identificazione delle proprietà di risonanza degli strumenti musicali, o più in generale delle sorgenti sonore, è un compito difficile, e di solito ci si accontenta di determinare i valori del tasso di *roll-off* e della frequenza di *cut-off*.

Coefficienti di *speech processing* I coefficienti di predizione lineare (*Linear Prediction Coefficients*, LPC), del *cepstrum* e del *mel-frequency cepstrum* (MFCC) sono *feature* largamente utilizzate con successo nell'ambito dello *speech processing*. Purtroppo, la loro applicazione in questa area di ricerca non ha replicato i risultati sperati [109]. Al contrario, il tasso di attraversamento dello zero[8] (*zero-crossing rate*, ZCR) è proficuamente utilizzato in diversi sistemi come semplice misura del contenuto in frequenza.

2.5.5 Alcune considerazioni

Molte caratteristiche riportate sono solitamente calcolate "istante per istante," ovvero, a partire dai dati forniti dal *front-end* adottato, che sono a loro

[8] Si parla di un attraversamento dello zero quando l'ampiezza istantanea assume segni diversi in due istanti successivi.

volta calcolati su finestre temporali[9] la cui durata è dell'ordine dei millisecondi. Considerando che un evento sonoro difficilmente dura meno di mezzo secondo, si pone il problema di condensare tutti i valori ottenuti in un unico valore reale, sbarazzandosi così della dimensione temporale. Come è stato anticipato nella sezione 2.3, non si è ancora trovato un metodo che effettui questa operazione in modo soddisfacente, e rispettoso della percezione uditiva. Ci si accontenta così degli ordinari momenti centrali statistici, quali l'integrale (momento di ordine zero), la media (momento di ordine uno), la varianza (momento di ordine due), e più raramente, l'asimmetria, o *skewness* (momento di ordine tre) [28], e l'autocorrelazione [105]. In altri casi, si opera sulle differenze dei valori adiacenti, sommandone ad esempio i valori assoluti, o calcolandone media e varianza. Il sistema SoundFisher calcola queste statistiche pesando i dati per il valore dell'ampiezza relativa ad ogni finestra[10]. In questo modo si enfatizzano, in prima approssimazione, le regioni percettivamente rilevanti del suono.

Combinando questi metodi con le caratteristiche succitate, si ottiene un insieme sterminato di potenziali variabili. Se, per un verso, questo può sembrare un vantaggio, dall'altro pone dei seri limiti di realizzazione di sistemi di riconoscimento in tempo reale su architetture non parallele. Per di più, secondo un noto risultato della teoria della classificazione [14], volendo mantenere il tasso di errore costante, un classificatore necessita di una sequenza di *training* che cresce esponenzialmente con il numero delle variabili in gioco, a causa del rumore che viene inevitabilmente introdotto. Questo paradosso è noto in letteratura come la "sciagura delle dimensioni" (*curse of dimensionality*). Nel paragrafo 3.2.11 verranno introdotte alcune tecniche statistiche che consentono di ridurre il numero di dimensioni conservando al meglio il contenuto informativo.

La particolare tecnica di classificazione adottata tiene conto automaticamente, in alcuni casi, della correlazione tra le diverse caratteristiche. Ad esempio, nel caso si adotti un modello a misture gaussiane, il sistema considera, per costruzione, tutte le combinazioni lineari delle variabili esaminate, che è quindi inutile, se non dannoso, valutare separatamente (correlazione lineare).

---

[9] Le finestre possono sovrapporsi (*overlapped windows*) e/o essere caratterizzate da transizioni più dolci della semplice finestra rettangolare (ad esempio finestre di Bartlett, Hanning, Hamming, etc.), al fine di evitare l'introduzione di componenti spurie in alta frequenza.

[10] Si tratta della radice del valore quadratico medio dell'ampiezza istantanea (*root mean square*, RMS).

Le caratteristiche esposte nei precedenti paragrafi sono quasi tutte distribuite normalmente per la maggior parte degli strumenti. Le uniche eccezioni sono il *pitch*, di cui si è già discusso a pagina 21, e le *feature* relative allle oscillazioni in ampiezza e in frequenza. In questi casi, più che di una distribuzione continua, si tratta di una mistura finita di due distribuzioni: la prima, avente una media e una varianza assai ridotte, è relativa agli strumenti, come il pianoforte, che non danno all'esecutore la possibilità di applicare alcuna modulazione periodica; la seconda contempla l'insieme complementare. Si noti, tuttavia, che anche questo modello più accurato è messo in discussione dal fatto che considerare un consistente numero di note eseguite senza modulazioni può distorcere le stime relative alla seconda distribuzione.

2.6 Sistemi di classificazione esistenti

I sistemi di riconoscimento automatico dei timbri noti dalla letteratura non si distinguono solo per i *feature set* adottati, ma anche e soprattutto per le tecniche di classificazione utilizzate. Alla base di questi lavori, c'è spesso un'ipotesi di base, ovvero il modello di mistura gaussiana (*Gaussian Mixture Model*), con cui si assume che le distribuzioni condizionali (cioè quelle relative ai singoli strumenti, o classi) siano delle multinormali. La definizione di distribuzione multinormale è fornita nel paragrafo 3.2.2. Allo stesso modo, nel capitolo 3 sono illustrate nel dettaglio le principali tecniche di classificazione, che in questo paragrafo verranno solo menzionate.

G. De Poli, P. Prandoni, P. Tonella [13]
: Utilizzando una serie di reti neurali a mappa auto-organizzante (*Self-Organizing Map*, SOM) sono stati classificati 40 strumenti musicali, basandosi su un campione di una sola nota per strumento, alla medesima altezza. Non sono stati effettuati test con dati indipendenti.

I. Kaminskyj, A. Materka [50]
: Sono state messe a confronto una tecnica di *k-nearest neighbor* e una tecnica neurale *feed-forward* per la classificazione di note isolate di chitarra, pianoforte, marimba e fisarmonica. I risultati sono incoraggianti (il tasso di errore è pari al 1,9%), ma sono stati ottenuti con test che utilizzavano le stesse registrazioni con cui i sistemi erano stati addestrati.

SoundFisher [105, 106]
: Messo a punto da Wold e Blum della MuscleFish utilizza la tecnica dell'analisi discriminante quadratica (QDA, paragrafo 3.2.4). Trattandosi di un sistema commerciale, non sono stati pubblicati i dettagli implementativi. Le prestazioni sono

più che soddisfacenti, e possono essere messe alla prova attraverso la *demo* del sito http://www.musclefish.com. L'interfaccia consente l'immissione di interrogazioni "*fuzzy*," almeno nelle intenzioni degli autori, assegnando dei limiti di tolleranza ai diversi parametri.

A. Fraser, I. Fujinaga [28]

Viene adottata una tecnica di *k-nearest neighbor* (paragrafo 3.2.9), in cui si utilizza una metrica euclidea pesata, ovvero nella misura della distanza ogni caratteristica è moltiplicata per un coefficiente di rilevanza. Questi coefficienti sono determinati per mezzo di un algoritmo genetico in modo da minimizzare il tasso di errore. Le prestazioni, come è lecito aspettarsi, peggiorano all'aumentare del numero di strumenti analizzati: per tre strumenti si hanno tassi di errore variabili tra il 2,7% e il 6,1%, mentre per 39 strumenti variano tra il 36,7% e il 45,6%. I dati di addestramento e quelli di test provengono dallo stesso CD.

J. Marques, P. J. Moreno [59]

Le tecniche della QDA (paragrafo 3.2.4) e delle *Support Vector Machines* (SVM, paragrafo 3.2.9) vengono messe a confronto per un problema di classificazione di otto strumenti, utilizzando dati provenienti da diverse registrazioni sia per i dati di *training* che per i dati di test. Le SVM si sono comportate meglio, fornendo risultati erronei solo nel 30% dei casi, contro il 37% della QDA. Utilizzando dati provenienti dalla stessa registrazione i risultati migliorano di un ordine di grandezza. Sono stati validati diversi insiemi di caratteristiche, provenienti dalle ricerche di *speech recognition*. I risultati migliori sono stati ottenuti con i coefficienti MFCC.

K. D. Martin, Y. E. Kim [63]

Vengono confrontate le tecniche di analisi discriminante canonica (paragrafo 3.2.6) e di *k-nearest neighbor*, concludendo debolmente in favore della prima. L'aspetto più interessante di questo lavoro è l'introduzione di tecniche di classificazione gerarchica per questo problema, che saranno meglio sfruttate nel successivo lavoro di Martin, a cui è dedicato il paragrafo 2.6.1. I test sono stati effettuati con una convalida incrociata (*cross-validation*) usando gli stessi dati dell'addestramento.

J. C. Brown [8]

Sfruttando il *front-end* suggerito in [7] e la tecnica QDA, il sistema riesce a riconoscere brani di registrazioni musicali com-

merciali di solo sassofono e oboe sbagliando nel 7,5% dei casi.

A. Eronen [20] Basandosi sui precedenti lavori di Martin [63, 62], vengono introdotti nuovi tipi di caratteristiche, modificata la gerarchia e, per la prima volta, selezionate, seppure manualmente, le migliori caratteristiche in ogni nodo decisionale. I test su note isolate hanno dato risultati molto soddisfacenti: mediamente 20% di errori di classificazione sui singoli strumenti.

2.6.1 Martin e la classificazione gerarchica

Considerato il gran numero di punti di contatto con il presente lavoro si è voluto dedicare un paragrafo alla tesi di Ph. D. di Martin [62] del Massachussets Institute of Technology (MIT).

Facendo seguito alle esperienze di classificazione con Kim [63], viene ampliato il discorso sulla classificazione gerarchica. Gli strumenti in esame vengono raggruppati in "superclassi," ottenendo una gerarchia a tre livelli (figura 2.2), non molto dissimile dalla classificazione tradizionale degli strumenti musicali[11]. L'enfasi posta sulla struttura gerarchica si basa sugli studi psicologico-percettivi di Rosch [84] e Minsky [70], che spingono a ritenere che il processo di riconoscimento timbrico sia intrinsecamente stratificato, come avviene in altri domini applicativi: prima di capire che si tratta di un alano, un essere vivente viene riconosciuto come quadrupede, e successivamente come cane. I vantaggi di un classificatore gerarchico sono esposti nel paragrafo 4.2.3.

La classificazione vera e propria si basa su una stima di media e varianza delle distribuzioni delle singole caratteristiche, in funzione del *pitch*. Assumendo che ad ogni frequenza della nota suonata corrisponda una distribuzione normale di ogni caratteristica, e che le caratteristiche siano tra loro indipendenti, si perviene ad una semplice formula moltiplicativa[12], che

[11] Gli strumenti della tradizione occidentale vengono suddivisi in base alle caratteristiche costruttive e di eccitazione. Al primo livello della gerarchia si distingue tra idiofoni (il suono scaturisce dalla vibrazione dell'intero strumento, generalmente per percussione), membranofoni (i vari generi di tamburo), cordofoni, areofoni ed elettrofoni. I cordofoni si suddividono ulteriormente in strumenti a corde pizzicate (chitarra, sitar, banjo, etc.) e a corde strofinate (principalmente gli archi orchestrali). Tra gli aerofoni si distinguono i flauti, gli strumenti ad ancia (singola e doppia) e a bocchino.

[12] Si tratta di una semplificazione della (3.6), analizzata nella sezione 3.1.

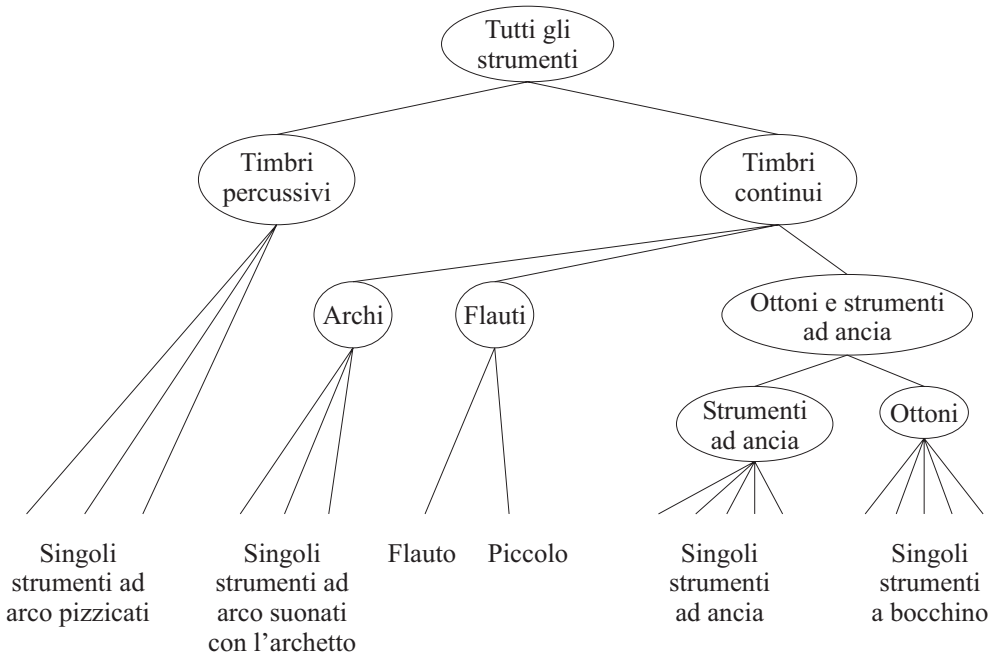


Figura 2.2. La gerarchia adottata nel lavoro di Martin. Al primo livello, si distingue tra timbri impulsivi (nell'insieme considerato si tratta di soli archi pizzicati) e "continui" (*sustained*, ovvero caratterizzati da una dominante fase di sostegno). Questi ultimi, a loro volta, sono suddivisi in flauti, archi e strumenti ad ancia e a bocchino.

consente di stabilire a quale famiglia o a quale strumento appartiene con la massima probabilità. Un altro aspetto che distingue il lavoro da quello precedente in collaborazione con Kim, è l'aver effettuato l'addestramento e i test con frasi musicali, e non più con note isolate.

La struttura gerarchica adottata da Martin è fissa e si basa in parte sulla classificazione tradizionale, e in parte sulla letteratura relativa alla rilevanza delle caratteristiche timbriche, trattata nella sezione 2.3. Questo, a parere dell'autore, è il motivo per cui, come già risultava dai test effettuati in [63], la classificazione "piatta," ovvero effettuata direttamente al livello inferiore della gerarchia, risulta spesso più efficace rispetto alla classificazione gerarchica. Sono infatti riportati tassi di errore del 32,5% per la prima e del 38,7% per la seconda, relativamente a un problema di classificazione tra 14 strumenti. Una imposizione della gerarchia da parte del ricercatore, in realtà, può essere limitante, e rischia di mettere in difficoltà il sistema di classificazione a causa di raggruppamenti non basati sui dati di *training*. Non stupisce, perciò, che

lo stesso Martin elenchi tra i possibili miglioramenti del proprio sistema la costruzione automatica dell'organizzazione gerarchica.

Un altro aspetto incerto della ricerca è l'assunzione "ingenua," per ammissione dello stesso Martin, di indipendenza tra le varie caratteristiche. In effetti, la maggior parte dei metodi di classificazione conosciuti sfrutta la correlazione tra le variabili, informazione spesso catturata attraverso la stima della matrice di covarianza. L'assunzione consente tuttavia di utilizzare modelli più generali di quello delle misture gaussiane, e questo vantaggio è stato sfruttato effettuando separatamente una stima non parametrica per la distribuzione del *pitch*, che nel sistema è utilizzato come una comune *feature*. Inoltre, in una comunicazione personale, Martin afferma che questa ipotesi gli ha consentito di addestrare il classificatore con un numero più limitato di campioni, aggirando così la "sciagura delle dimensioni."

capitolo 3

Tecniche di classificazione

Il problema della classificazione automatica, o *pattern recognition*, trova i suoi prodromi nei lavori pionieristici di *statistica multivariata* di Pearson [77] di inizio secolo, sviluppati da Fisher [21], Mahalanobis [56] e Hotelling [44] negli anni 30. In seguito, parallelamente all'approccio statistico classico (*analisi discriminante*, o *discriminant analysis*—una branca della statistica multivariata), nascono e si sviluppano le tecniche di classificazione non-parametrica.

L'obiettivo è quello di identificare la natura dell'oggetto in esame all'interno di una rosa di classi possibili, siano esse note a priori o meno. Perché il processo sia automatizzabile, è necessario effettuare una serie di misurazioni di caratteristiche significative sui campioni (*feature extraction*), in modo da avere a disposizione dei vettori numerici, visti come realizzazione di un vettore casuale, facilmente manipolabili dal calcolatore, che fornisce in uscita la classe di appartenenza stimata (figura 2.1). La classificazione è detta assistita (*supervised*) se vengono forniti al calcolatore esempi di osservazioni di cui sia nota la classe. L'analisi dei *cluster* (*cluster analysis*, *analisi dei grappoli*) e le reti neurali a mappa auto-organittante (*Self-Organizing Map*, SOM) sono tipici esempi di tecniche di classificazione non-assistita.

La classificazione automatica è strettamente correlata, e per certi versi sovrapposta, all'apprendimento automatico [57]. Esso presenta infatti una simile tassonomia delle tecniche, che possono essere parametriche o non parametriche, mentre l'apprendimento può essere assistito e non-assistito.

Si espongono in questo capitolo i fondamenti della teoria della classificazione, e vengono illustrati in dettaglio l'approccio statistico al problema (sezione 3.2) e le principali tecniche di analisi dei *cluster* (sezione 3.3).

Si assume che il lettore conosca i fondamenti della statistica e dell'algebra lineare (si vedano, ad esempio, [71] e [12]). [24] fornisce una gradevole introduzione alla statistica multivariata e alle sue applicazioni, con un encomiabile ricchezza di esempi, e una serie di istruttivi algoritmi disponibili via `ftp`. Un'ottima monografia sull'analisi discriminante, ricca di risultati e spunti stimolanti (incluse le regole di *k-nearest neighbor* e le *kernel rules*), contenente una imponente bibliografia di oltre 1200 titoli, è [66]. Per una panoramica molto tecnica e articolata sulle principali tecniche parametriche e non-parametriche esistenti e sulla classificazione automatica in generale (seppure limitata a due classi), si rimanda a [14].

L'analisi dei *cluster* è coperta da una vecchia, ma autorevole monografia [41], e da due testi di impostazione applicativa, completi di algoritmi [83, 98].

3.1 Formalizzazione del problema

In questo paragrafo si fornisce una definizione rigorosa del problema della classificazione. La notazione è mutuata principalmente da [24].

Una *osservazione* (o *oggetto*, o *caso*, o *entità*) è determinata da un vettore p-dimensionale di variabili casuali, che rappresenta le misurazioni effettuate sull'oggetto da classificare. La natura (nota o incognita) di un'osservazione è detta *classe*. Nella presente trattazione si assume che le variabili siano continue, o comunque misurate su una metrica razionale. Per una efficace esposizione dei quattro tipi di variabili esistenti (nominali, ordinali, intervallari e razionali), si veda [98].

Le osservazioni già classificate, se esistono, sono raccolte in k matrici di dati $\mathbf{D}_j$, che hanno tante righe (N_j) quante osservazioni appartenenti a quella classe, e tante colonne (p) quante variabili. Si definisce $N \stackrel{\text{def}}{=} \sum_{j=1}^{k} N_j$ il numero totale di osservazioni.

Data una generica osservazione $\mathbf{y}$ di natura incognita ed un numero k di classi si definisce *classificatore* (o *regola di classificazione*) una funzione $g(\mathbf{y}) : \mathbb{R}^p \to \{1, \ldots, k\}$. Se $\mathbf{y}$ appartiene alla classe j e $g(\mathbf{y}) \neq j$ si dice che il classificatore g commette un *errore* nella classificazione di $\mathbf{y}$.

Sia X una variabile casuale a valori discreti $\{1, \ldots, k\}$ che indichi l'appartenenza ad una classe. Si definiscono le *probabilità a priori* di appartenenza alla j-esima classe le quantità

$$\pi_j \stackrel{\text{def}}{=} \Pr[X = j] \qquad 1 \leq j \leq k. \tag{3.1}$$

Naturalmente $\sum_{j=1}^{k} \pi_j = 1$. Sia $\mathbf{Y}$ un vettore casuale continuo di dimensione

p, e sia

$$X_g \stackrel{\text{def}}{=} g(\mathbf{Y}) \tag{3.2}$$

la variabile casuale che rappresenta la classificazione di g per $\mathbf{Y}$. Si definisce inoltre la *probabilità* o *tasso di errore* di un classificatore

$$\gamma_g \stackrel{\text{def}}{=} \Pr[X_g \neq X]. \tag{3.3}$$

Si definisce *classificatore ottimo*, *classificatore a minimo tasso di errore*, o *classificatore di Bayes* una funzione

$$g^*(\cdot) \stackrel{\text{def}}{=} \underset{g:\mathbb{R}^p \to \{1,\ldots,k\}}{\arg\min}\ \gamma_g, \tag{3.4}$$

ovvero un classificatore che renda minima la probabilità di errore per la generica osservazione $\mathbf{Y}$.

Ogni classificatore $g(\cdot)$ ripartisce lo spazio campionario $\mathbb{R}^p$ in *regioni di classificazione* $C_1, \ldots, C_k$, tali per cui

$$g(\mathbf{y}) = j \Leftrightarrow \mathbf{y} \in C_j \qquad 1 \leq j \leq k.$$

Un classificatore può essere espresso da un insieme di k funzioni

$$g_i : \mathbb{R}^p \to \{1, \ldots, p\} \qquad 1 \leq i \leq k$$

scelte in modo tale che

$$g(\mathbf{y}) = j \quad \Rightarrow \quad g_j(\mathbf{y}) > g_i(\mathbf{y}) \qquad \mathbf{y} \in \mathbb{R}^p, \quad j \neq i \in \{1, \ldots, k\}. \tag{3.5}$$

Questo consente di effettuare la decisione basandosi sul calcolo delle k funzioni ed attribuendo all'osservazione in esame la classe per cui la $g_i(\cdot)$ relativa assume in quel punto il valore massimo rispetto alle altre. Le $g_i(\cdot)$ vengono dette *funzioni discriminanti*. Si osservi che non sono univocamente definite per un determinato classificatore, dal momento che, dato un insieme $\{g_1(\cdot), \ldots, g_k(\cdot)\}$ ed una funzione monotona crescente $f(\cdot)$, la (3.5) continua a valere anche per l'insieme $\{f(g_1(\cdot)), \ldots, f(g_k(\cdot))\}$. Pur dando luogo alle stesse regioni di classificazione, le funzioni trasformate possono essere più facili da calcolare. Questa proprietà viene sfruttata ad esempio nella dimostrazione del risultato della sezione 3.2.4, in cui si fa uso dell'addizione di costanti positive, la moltiplicazione per costanti strettamente positive e la funzione logaritmo.

Sia $f_j(\cdot)$ la funzione densità di probabilità (*probability density function*, PDF) del vettore $\mathbf{Y}$ data la sua appartenenza alla classe j, o, per brevità, PDF della classe j. Facendo uso della formula di Bayes si ottiene che la probabilità

che un'osservazione $\mathbf{y}$ faccia parte della classe j (*probabilità a posteriori*) è data da

$$\pi_{j\mathbf{y}} \stackrel{\text{def}}{=} \Pr[X = j|\mathbf{Y} = \mathbf{y}] = \frac{\pi_j f_j(\mathbf{y})}{\sum_{h=1}^{k} \pi_h f_h(\mathbf{y})} \qquad 1 \leq j \leq k. \tag{3.6}$$

La probabilità di classificare come appartenente alla classe i un'osservazione di natura j è data da

$$q_{ij} = \Pr[\mathbf{Y} \in C_i|X = j] = \int_{C_i} f_j(\mathbf{y})\, d\mathbf{y}, \tag{3.7}$$

e si verifica facilmente che la probabilità di errore di un classificatore è pari a

$$\gamma_g = \sum_{j=1}^{k} \pi_j(1 - q_{jj}) = 1 - \sum_{j=1}^{k} \pi_j q_{jj}. \tag{3.8}$$

Si dimostra che ogni classificatore $\check{g}(\cdot)$ avente come funzioni discriminanti

$$\check{g}_j(\mathbf{y}) = \pi_j f_j(\mathbf{y}) \qquad 1 \leq j \leq k, \tag{3.9}$$

è un classificatore ottimo[1].

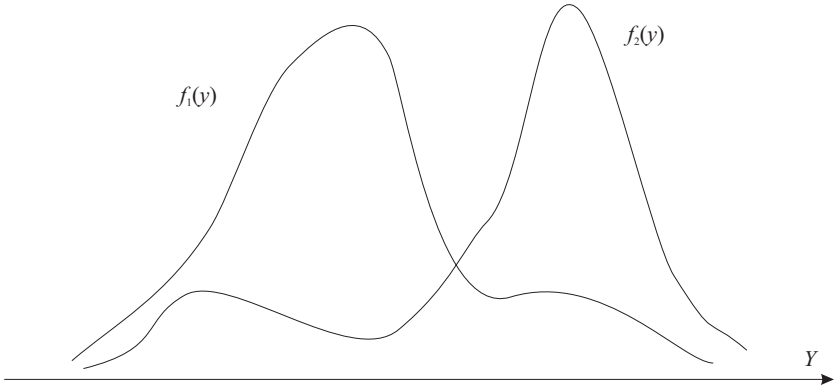
La figura 3.1 esemplifica quanto detto finora in un semplice caso ($p = 1$, $k = 2$).

3.1.1 Fattori di costo e teoria delle decisioni

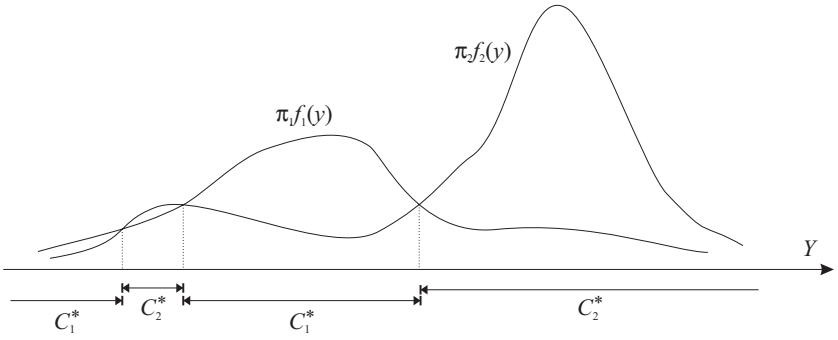
È possibile generalizzare [16, 66] la teoria suesposta tenendo conto di condizioni particolari in cui alcuni errori di classificazione danno luogo a costi diversi da altri. Si pensi al classico esempio delle diagnosi mediche, in cui classificare un paziente sano come malato è molto meno grave di classificarne uno malato come sano.

Si introduce la matrice dei costi $\mathbf{C}$, il cui generico elemento c_{ij} rappresenta il *fattore di costo* della classificazione di un'osservazione come elemento della classe i ($X_g = i$) quando è di natura j ($X = j$). Si noti che non è necessario imporre la condizione $c_{ii} = 0$. Nel paragrafo precedente si era implicitamente

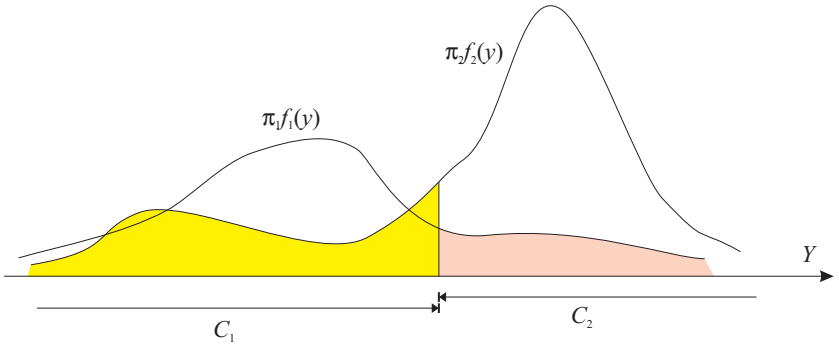
[1] Dal punto di vista strettamente formale, in realtà, questo è scorretto, perché nei punti in cui due o più $\check{g}_j(\mathbf{y})$ assumono lo stesso valore, alcune disuguaglianze della (3.5) vengono violate. Tuttavia si preferisce non appesantire ulteriormente la notazione, e assumere che, in questi casi, la decisione è arbitraria.



(a) Funzioni di distribuzione di due variabili casuali.



(b) Funzioni discriminanti e regioni di classificazione del classificatore ottimo assumendo $\pi_1 = \frac{1}{3}$ e $\pi_2 = \frac{2}{3}$.



(c) Regioni di classificazione di un classificatore *non* ottimo e relativo tasso di errore (somma delle aree colorate). Si noti che anche per il classificatore ottimo il tasso di errore non è nullo.

Figura 3.1. Esempio di classificazione ottima e di errore di classificazione per $k = 2$ variabili casuali ($p = 1$).

assunto $\mathbf{C} = 1 - \mathbf{I}$. In altre parole, si attribuiva costo 1 ad un qualsiasi errore di classificazione, e costo zero in caso di classificazione corretta.

Si definisce *rischio condizionato* $R_i(\mathbf{y})$ il valore atteso del costo di una classificazione di $\mathbf{y}$ nella classe i. Si deriva

$$R_i(\mathbf{y}) = \sum_{j=1}^{k} c_{ij} \Pr[X = j | \mathbf{Y} = \mathbf{y}] = \sum_{j=1}^{k} c_{ij} \pi_{j\mathbf{y}}. \tag{3.10}$$

Il *rischio totale* del classificatore è definito come

$$R_g \stackrel{\text{def}}{=} \int_{\mathbb{R}^p} R_{X_g}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}, \tag{3.11}$$

in cui $f_{\mathbf{Y}}(\cdot)$ è la PDF marginale di $\mathbf{Y}$, data da

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{y}) \qquad \mathbf{y} \in \mathbb{R}^p \tag{3.12}$$

Si dimostra che ogni classificatore $\check{g}(\cdot)$ avente come funzioni discriminanti

$$\check{g}_j(\mathbf{y}) = \sum_{j=1}^{k} c_{ij} \pi_j f_j(\mathbf{y}) \qquad 1 \leq j \leq k, \tag{3.13}$$

rende minimo il rischio totale $R_{\check{g}}$.

Nel seguito, se non verrà esplicitato il contrario, si assumerà $\mathbf{C} = 1 - \mathbf{I}$, ricercando il classificatore ottimo introdotto nel paragrafo 3.1.

3.2 L'approccio statistico

La (3.9) del paragrafo precedente sembrerebbe fornire, insieme, il problema e la soluzione. Purtroppo, però, le quantità di quella equazione sono ignote, e ne è richiesta una stima a partire dai dati disponibili, forniti al calcolatore nella *fase di addestramento* (o *training*).

Per quanto riguarda le probabilità a priori π_j si profilano tre alternative:

1. $\hat{\pi}_j = \frac{1}{k}$;
2. una semplice stima basata sulla sequenza di *training*, data da $\hat{\pi}_j = \frac{N_j}{N}$;
3. una stima fornita da esperti (per esempio basandosi su popolazioni più significative della sequenza di *training*, o sulla mera esperienza).

Il problema della stima delle PDF di un vettore casuale è un problema vastissimo ed ancora oggetto di ricerca nella sua formulazione generale. In questa trattazione ci si occuperà del caso più studiato in letteratura in cui le popolazioni seguono una distribuzione multinormale. Nel seguito, perciò, a volte si assumerà tacitamente che le distribuzioni trattate obbediscono a questa legge. Per una trattazione comprendente risultati ed esempi riguardanti anche altre distribuzioni, si rimanda a [24].

L'ipotesi di multinormalità dei dati può essere suffragata da opportuni test statistici, illustrati nella sezione 3.2.10.

3.2.1 Distanza standard

La maggior parte delle tecniche di classificazione e raggruppamento analizzate in questo capitolo coinvolge un processo di misurazione. La distanza standard riveste un ruolo centrale in questo ambito, e viene quindi introdotta insieme ad altri concetti di base, come le distribuzioni di probabilità multinormali e la stima dei parametri, in questi primi paragrafi.

Un generico vettore casuale $\mathbf{Y}$ di dimensione p avente matrice di covarianza $\boldsymbol{\Sigma}$ induce sullo spazio campionario $\mathbb{R}^p$ la metrica

$$\|\mathbf{y}\|_{\boldsymbol{\Sigma}^{-1}} = \sqrt{\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}}, \tag{3.14}$$

da cui si deriva la *distanza standard*, o *distanza di Mahalanobis*[2], definita come

$$\Delta_{\mathbf{Y}}(\mathbf{y}_1, \mathbf{y}_2) \stackrel{\text{def}}{=} \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_1 - \mathbf{y}_2)}. \tag{3.15}$$

La distanza standard può essere interpretata come una generalizzazione della familiare distanza euclidea, in cui vale $\boldsymbol{\Sigma} = \mathbf{I}$. Per altri tipi di generalizzazione della metrica euclidea, si veda [98].

Di particolare interesse per l'analisi discriminante si rivela la distanza standard di un punto dalla media $\boldsymbol{\mu}$ della distribuzione in questione. La quantità $\Delta_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\mu})$ fornisce infatti una comoda misura dello scostamento di un'osservazione dagli "standard" di una determinata classe, permettendo di individuare immediatamente le osservazioni atipiche (*outlier*). La visualizzazione della distanza standard è un passo fondamentale per la comprensione

[2] Alcuni autori definiscono distanza di Mahalanobis il quadrato della quantità indicata nella (3.15). In questo caso, tuttavia, essa non soddisfa più gli assiomi che caratterizzano una metrica. Si spera che il numero sfortunato capitato a questa nota a pié di pagina non generi ulteriore confusione.

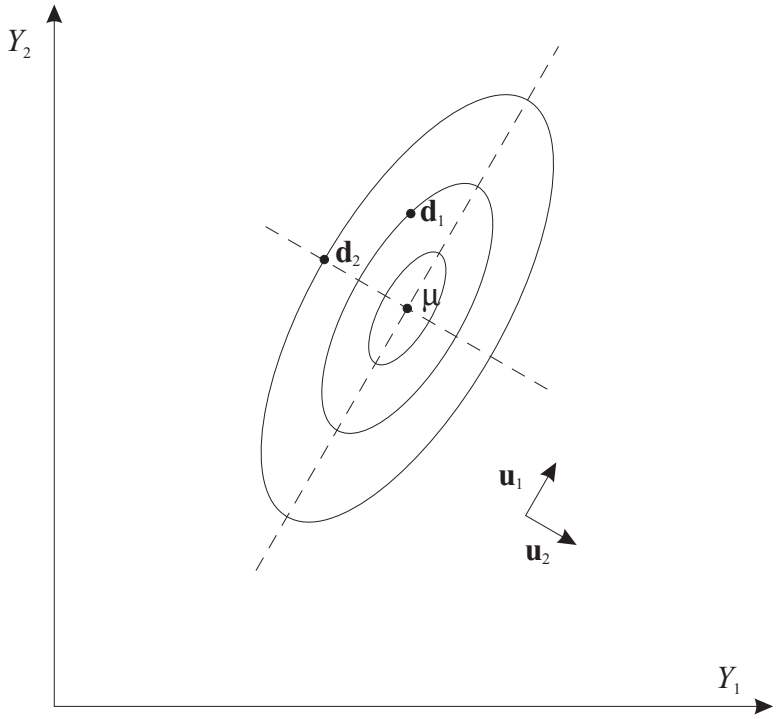


Figura 3.2. Luoghi di distanza standard costante dalla media $\boldsymbol{\mu}$ per una distribuzione bivariata avente matrice di covarianza $\boldsymbol{\Sigma} = \begin{bmatrix} 5{,}5 & 2{,}5981 \\ 2{,}5981 & 8{,}5 \end{bmatrix}$. $\mathbf{d}_1$ e $\mathbf{d}_2$ rappresentano due osservazioni aventi la stessa distanza euclidea dalla media. $\mathbf{u}_1$ e $\mathbf{u}_2$ sono gli autovettori della matrice di covarianza.

dei concetti che seguiranno. Si dimostra che, per $p = 2$, i luoghi dei punti aventi distanza standard costante dalla media $\boldsymbol{\mu}$ sono delle ellissi[3] centrate attorno a $\boldsymbol{\mu}$. La figura 3.2 mostra i luoghi dei punti a distanza standard costante dalla media per una distribuzione bivariata. Le ellissi più esterne corrispondono a costanti più grandi. Si noti che i punti $\mathbf{d}_1$ e $\mathbf{d}_2$ hanno la medesima distanza euclidea dalla media, ma, in termini di distanza standard, $\mathbf{d}_1$ è più vicino a $\boldsymbol{\mu}$ di $\mathbf{d}_2$.

Siano λ_i gli autovalori della matrice di covarianza ordinati in modo che

$$|\lambda_i| \geq |\lambda_j| \qquad 1 \leq i < j \leq p.$$

Essi sono reali, in quanto $\boldsymbol{\Sigma}$ è reale e simmetrica. Siano $\mathbf{u}_i$ gli autovettori

[3] Per $p = 3$ sono ellissoidi, per $p > 3$ iperellissoidi.

normalizzati associati ai λ_i. In seguito si assumerà che gli autovalori sono distinti, ipotesi ragionevole nel caso in cui si debba stimare la matrice di covarianza da dati reali. Si dimostra [3, 12] che in queste condizioni gli autovettori formano un sistema ortonormale. Il valore assoluto degli autovalori quantifica l'estensione della distribuzione nella direzione dell'autovettore corrispondente. Ad esempio, per la distribuzione trattata nella figura 3.2, si ottengono i seguenti risultati, facilmente interpretabili alla luce di quanto detto:

$$\lambda_1 = 10 \qquad \mathbf{u}_1 = \begin{bmatrix} 1/2 \\ \sqrt{3}/2 \end{bmatrix}, \qquad \lambda_2 = 4 \qquad \mathbf{u}_2 = \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix}.$$

Un altro modo di leggere questo risultato è il seguente: l'autovettore associato all'autovalore dominante fornisce i coefficienti della combinazione lineare delle variabili che presenta la massima variabilità (a meno di una costante moltiplicativa, che è un grado di libertà nella scelta dei coefficienti). Le combinazioni lineari associate agli altri autovettori di conseguenza ordinati possiedono una variabilità via via decrescente, ma massima, sotto il vincolo di indipendenza dalle combinazioni lineari precedenti. Si potrebbe dire che l'informazione relativa alla distribuzione del vettore casuale è maggiormente concentrata nelle combinazioni lineari delle variabili studiate associate (attraverso i relativi autovettori) agli autovalori dominanti. Considerazioni di questo tipo sono alla base dell'Analisi delle Componenti Principali (*Principal Component Analysis*, PCA), che si preoccupa di sfruttare la correlazione delle variabili al fine di condensare una buona parte dell'informazione in un numero inferiore di variabili ortogonali.

Un'altra interessante proprietà della distanza standard è la sua invarianza alle trasformazioni lineari. Data cioè una matrice non singolare $\mathbf{T}$, si ha

$$\Delta_{\mathbf{Y}}(\mathbf{y}_1, \mathbf{y}_2) = \Delta_{\mathbf{Y}}(\mathbf{T}\mathbf{y}_1, \mathbf{T}\mathbf{y}_2). \tag{3.16}$$

Questo è un ulteriore vantaggio sulla distanza euclidea, invariante solo rispetto alle trasformazioni rigide (rototraslazioni e riflessioni).

3.2.2 Distribuzioni multinormali

In questo paragrafo si richiamano le definizioni di distribuzione normale e multinormale, e le relative proprietà.

Si dice che una variabile casuale continua Y ha una una distribuzione *normale*, o *gaussiana* (univariata) se possiede una PDF del tipo

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right), \tag{3.17}$$

e si verifica che essa ha media μ e varianza σ^2. Per brevità, si scriverà $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Generalizzando, un vettore casuale p-variato $\mathbf{Y}$ ha una una distribuzione *multinormale* (o normale multivariata) se possiede una PDF del tipo

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-p/2}(\det \boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{3.18}$$

con $\boldsymbol{\Sigma}$ matrice definita positiva, e si verifica che esso ha vetore media $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$. Per brevità, si scriverà $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Si ricorda che la *matrice di covarianza* di un vettore casuale $\mathbf{Y}$ avente media $\boldsymbol{\mu}$ è definita come

$$\mathrm{Cov}[\mathbf{Y}] \stackrel{\text{def}}{=} \mathrm{E}\left[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\right].$$

Sulla diagonale di questa matrice si leggono le varianze delle rispettive variabili (con un pasticcio notazionale, $\sigma_{ii} = \sigma_i^2 \stackrel{\text{def}}{=} \mathrm{Var}[Y_i]$), e sul generico elemento σ_{ij} la covarianza tra le variabili Y_i e Y_j.

È possibile dimostrare che, se $\mathbf{Y}$ è un vettore multinormale di dimensione p, per ogni matrice $\mathbf{A}$ di dimensioni $k \times p$ ed ogni vettore $\mathbf{b}$ di dimensione k, il vettore casuale $\mathbf{Y} = \mathbf{A}\mathbf{Y} + \mathbf{b}$ è un vettore multinormale di dimensione k.

Sia $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Si dimostra che la variabile casuale

$$\Delta^2_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\mu}), \tag{3.19}$$

ovvero il quadrato della distanza standard di $\mathbf{Y}$ dalla media, segue una distribuzione chi-quadrato con p gradi di libertà (χ^2_p).

Il vettore media e la matrice di covarianza costituiscono statistiche sufficienti per una distribuzione multinormale. Da un semplice confronto tra la (3.15) e la (3.18) segue che nel caso di distribuzioni multinormali i luoghi dei punti per cui la PDF ha valore costante presentano la stessa forma ellittica di cui si è parlato nel paragrafo precedente. In altri termini, qualora di una distribuzione si conoscano solo media e matrice di covarianza, la distanza standard fornisce informazioni tanto più aderenti alla realtà quanto più la distribuzione si avvicina ad una multinormale.

Potrebbe trasparire fin d'ora un semplice procedimento per classificare osservazioni all'interno di k distribuzioni multinormali: una volta stimati i parametri relativi a queste classi, e tenendo in debito conto le probabilità a priori, si attribuisce l'osservazione alla classe la cui distanza standard tra media e osservazione è minima. Sebbene questo procedimento porti al classificatore che abbiamo definito ottimo, sarà chiaro alla fine del paragrafo 3.2.4 che, paradossalmente, i risultati ottenuti in questo modo sono talvolta drogati, in quanto dipendono fortemente dalla sequenza di *training*.

3.2.3 Stima dei parametri

Si introducono in questa sezione i principali metodi di stima dei due parametri che caratterizzano una distribuzione multinormale: il vettore media e la matrice di covarianza.

Si definiscono formalmente le matrici delle osservazioni della j-esima classe introdotte a pagina 32, ciascuna contenente N_j vettori osservazione. Sia

$$\mathbf{D}_j \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{d}'_{j1} \\ \vdots \\ \mathbf{d}'_{jN_j} \end{bmatrix} \qquad 1 \leq j \leq k. \tag{3.20}$$

Si introduce inoltre la matrice

$$\mathbf{D} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{d}'_1 \\ \vdots \\ \mathbf{d}'_N \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{d}'_{11} \\ \vdots \\ \mathbf{d}'_{1N_1} \\ \mathbf{d}'_{21} \\ \vdots \\ \mathbf{d}'_{kN_k} \end{bmatrix}, \tag{3.21}$$

dove $N \stackrel{\text{def}}{=} \sum_{j=1}^{k} N_j$, ottenuta incolonnando le matrici $\mathbf{D}_j$. Si noti che i $\mathbf{d}_i$ sono vettori colonna.

Si supponga inizialmente $k = 1$, ovvero $\mathbf{D}_1 = \mathbf{D}$. Si assume che le $\mathbf{d}_i$ siano realizzazioni indipendenti del medesimo vettore casuale $\mathbf{Y}$.

La media è stimata dalla *media campionaria*, o *centroide*

$$\mathbf{m} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mathbf{d}_i. \tag{3.22}$$

È raro che questa statistica necessiti di un'alternativa, godendo delle principali proprietà auspicabili per uno stimatore e coincidendo peraltro con lo stimatore di massima verosimiglianza[4].

Per la matrice di covarianza, esistono due alternative principali: lo stimatore *plug-in*

$$\mathbf{S}_{\text{P}} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{d}_i - \mathbf{m})(\mathbf{d}_i - \mathbf{m})' = \frac{1}{N} \sum_{i=1}^{N} \mathbf{d}_i \mathbf{d}'_i - \mathbf{m}\mathbf{m}', \tag{3.23}$$

[4]Per una trattazione esauriente sulla teoria degli stimatori di massima verosimiglianza si rimanda a [24, sezione 4.3]

e la *matrice di covarianza campionaria*

$$\mathbf{S} \stackrel{\text{def}}{=} \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{d}_i - \mathbf{m})(\mathbf{d}_i - \mathbf{m})' = \frac{N}{N-1}\mathbf{S}_{\text{P}}. \tag{3.24}$$

Il vantaggio della $\mathbf{S}_{\text{P}}$ è che coincide con lo stimatore di massima verosimiglianza, ma è uno stimatore distorto. Al contrario, si dimostra che $\mathbf{S}$ è non-distorto. Anche se, evidentemente, la differenza tra i due è minima per N sufficientemente grande, verrà utilizzata l'una o l'altra formula a seconda delle necessità. Alcuni risultati dipendono infatti dalle proprietà della particolare statistica adottata.

Condizione necessaria per la definita positività delle matrici $\mathbf{S}$ ed $\mathbf{S}_{\text{P}}$ è $N \geq p+1$. Se le osservazioni $\mathbf{d}_i$ sono effettivamente indipendenti, comunque, la condizione è anche sufficiente con probabilità 1.

Data una coppia di vettori casuali p-variati $\mathbf{X}$ e $\mathbf{Y}$ aventi medie $\boldsymbol{\mu}_{\mathbf{X}} \neq \boldsymbol{\mu}_{\mathbf{Y}}$, e comune matrice di covarianza $\boldsymbol{\Sigma}$, si dimostra che lo stimatore

$$\mathbf{S}_{\text{pooled}} \stackrel{\text{def}}{=} \frac{(N_{\mathbf{X}}-1)\mathbf{S}_{\mathbf{X}} + (N_{\mathbf{Y}}-1)\mathbf{S}_{\mathbf{Y}}}{(N_{\mathbf{X}}+N_{\mathbf{Y}}-2)}, \tag{3.25}$$

denominato *matrice di covarianza campionaria comune* (*pooled sample covariance matrix*), è non-distorto per $\boldsymbol{\Sigma}$. Si tratta di una specie di media pesata delle due matrici, e viene talvolta utilizzata anche nel caso in cui l'ipotesi di uguale matrice di covarianza non è verificata. Essa è infatti la versione non distorta dello stimatore di massima verosimiglianza della matrice di covarianza comune, data dalla media pesata delle stime *plug-in* delle singole distribuzioni. La (3.25) è infine facilmente generalizzabile per $k > 2$ vettori casuali $\mathbf{X}_i$, definendo

$$\mathbf{S}_{\text{pooled}} \stackrel{\text{def}}{=} \frac{\sum_{j=1}^{k}(N_i-1)\mathbf{S}_{\mathbf{X}_i}}{(N-k)}. \tag{3.26}$$

Sia ora $k \geq 1$. Il seguente notevole risultato, noto in letteratura sotto il nome di equazioni MANOVA (*Multivariate ANalysis Of VAriance*), esprime media campionaria e varianza *plug-in* ($\mathbf{m}_{\text{Tot}}$ e $\mathbf{S}_{\text{Tot P}}$) di un insieme di k popolazioni, in funzione delle loro medie campionarie e varianze *plug-in* ($\mathbf{m}_j$ e $\mathbf{S}_{\text{P}j}$). Posto

$$\hat{\pi}_j = \frac{N_j}{N} \qquad 1 \leq j \leq k, \tag{3.27}$$

si ha

$$\mathbf{m}_{\text{Tot}} = \sum_{j=1}^{k} \hat{\pi}_j \mathbf{m}_j \tag{3.28}$$

$$\begin{aligned}\mathbf{S}_{\text{P Tot}} &= \sum_{j=1}^{k} \hat{\pi}_j \mathbf{S}_{\text{P}j} + \sum_{j=1}^{k} \hat{\pi}_j (\mathbf{m}_j - \mathbf{m}_{\text{Tot}})(\mathbf{m}_j - \mathbf{m}_{\text{Tot}})' \qquad (3.29)\\ &= \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{N_j} (\mathbf{d}_{ji} - \mathbf{m}_j)(\mathbf{d}_{ji} - \mathbf{m}_j)' \\ &\quad + \frac{1}{N} \sum_{j=1}^{k} N_j (\mathbf{m}_j - \mathbf{m}_{\text{Tot}})(\mathbf{m}_j - \mathbf{m}_{\text{Tot}})' \\ &\stackrel{\text{def}}{=} \mathbf{S}_{\text{W}} + \mathbf{S}_{\text{B}}.\end{aligned}$$

La matrice di covarianza totale consta quindi di due termini: il primo, $\mathbf{S}_{\text{W}}$, rende conto della variabilità all'interno (*within*) di ciascuna classe; il secondo, $\mathbf{S}_{\text{B}}$, è una misura della dispersione tra (*between*) le classi[5]. Questo rappresenta un notevole vantaggio computazionale rispetto alle definizioni (3.22) e (3.23), in quanto non è necessario riconsiderare le k matrici dei dati, le cui dimensioni non sono limitate superiormente. Un'immediata applicazione di queste equazioni è la seguente. Si supponga di possedere una matrice di dati per una classe, di cui sono disponibili le stime $\mathbf{m}$ e $\mathbf{S}_{\text{P}}$. Se in un secondo momento si rendono disponibili atri dati per quella classe, è possibile aggiornare le stime senza doverle ricalcolare.

In [24] sono analizzati in dettaglio metodi di stima più generali, come gli stimatori di massima verosimiglianza, e avanzati, come gli stimatori *bootstrap*. Alcuni studiosi (si veda [66, sezione 5.7] per approfondimenti) propongono metodi di stima robusta rispetto alle aberrazioni delle distribuzioni multinormali che saranno esposte nella sezione 3.2.8. In sostanza esse danno automaticamente un peso inferiore alle osservazioni atipiche.

3.2.4 Classificatore ottimo per classi multinormali

Si torna ora a ricercare il classificatore ottimo del paragrafo 3.1 per popolazioni aventi densità multinormali, partendo dal risultato (3.9) di pagina 34,

[5] Le stesse matrici moltiplicate per N prendono il nome di *matrici di dispersione*.

che permette di classificare un'osservazione in modo da rendere massima la probabilità a posteriori $\Pr[X = j|\mathbf{Y} = \mathbf{y}]$.

L'ipotesi di multinormalità consente di calcolare in forma chiusa, a partire dalle sole medie $\boldsymbol{\mu}_j$ e matrici di covarianza $\boldsymbol{\Sigma}_j$, l'espressione $\check{g}_j(\mathbf{y}) = \pi_j f_j(\mathbf{y})$ per ciascuna classe. Ottenuti questi valori, è sufficiente trovarne il massimo e decidere per la classe relativa, con la certezza di avere effettuato la scelta con la minima probabilità di errore. Sfruttando inoltre la proprietà delle funzioni discriminanti illustrata a pagina 33, è possibile effettuare calcoli più efficienti. Si considerano infatti le

$$g_j^{\text{QDA}}(\mathbf{y}) \stackrel{\text{def}}{=} \log(\check{g}_j(\mathbf{y})) + C \stackrel{\text{def}}{=} \log(\pi_j f_j(\mathbf{y})) + \frac{p}{2}\log(2\pi), \tag{3.30}$$

dove la costante serve ad eliminare il fattore $(2\pi)^{-p/2}$ comune a tutte le distribuzioni (si veda la (3.18)). Si verifica che

$$g_j^{\text{QDA}}(\mathbf{y}) = \mathbf{y}'\mathbf{A}_j\mathbf{y} + \mathbf{b}_j'\mathbf{y} + c_j \qquad 1 \leq j \leq k, \tag{3.31}$$

dove

$$\mathbf{A}_j = -\frac{1}{2}\boldsymbol{\Sigma}_j^{-1} \tag{3.32}$$

$$\mathbf{b}_j = \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j, \tag{3.33}$$

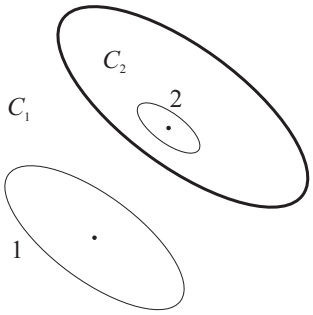
$$c_j = \log \pi_j - \frac{1}{2}\log(\det \boldsymbol{\Sigma}_j) - \frac{1}{2}\boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j. \tag{3.34}$$

La (3.31) è una forma quadratica, e per questo motivo si parla di *analisi discriminante quadratica* (*quadratic discriminant analysis*, QDA). Per $k = 2$ e $p = 2$ i bordi delle regioni di classificazione assumono la forma, relativamente semplice, di sezioni coniche. Nella figura 3.3, tratta da [16], ne vengono illustrati alcuni casi, assumendo $\pi_1 = \pi_2 = \frac{1}{2}$. Si noti che, in alcuni casi, la sezione conica degenera in linee rette. In particolare, se $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, la curva è una retta diretta come l'autovettore dominante e posta esattamente tra $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$.

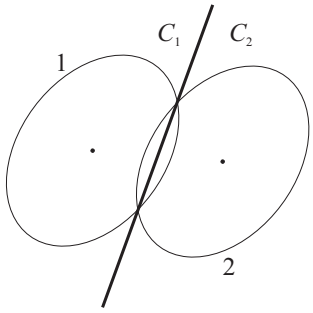
È possibile ricavare le probabilità a posteriori introdotte nella (3.6),

$$\pi_{j\mathbf{y}} = \frac{\exp\left(g_j^{\text{QDA}}(\mathbf{y})\right)}{\sum_{h=1}^{k}\exp\left(g_h^{\text{QDA}}(\mathbf{y})\right)} \qquad 1 \leq j \leq k. \tag{3.35}$$

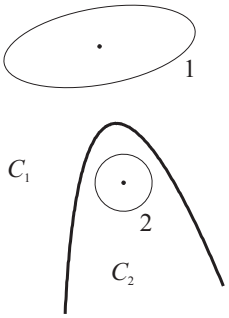
Sostituendo nelle (3.32) i parametri $\boldsymbol{\mu}_j$ e $\boldsymbol{\Sigma}_j$ con le relative stime campionarie, si ottiene, a partire dalle matrici di *training* $\mathbf{D}_j$, un classificatore automatico ottimo. Nonostante il nome promettente, però, un classificatore di questo tipo rischia di aderire troppo ai dati, che sono generalmente



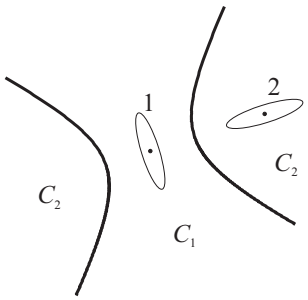
(a) Un'ellisse.



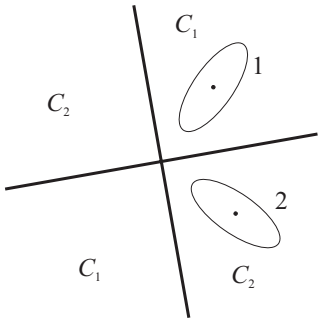
(b) Una linea.



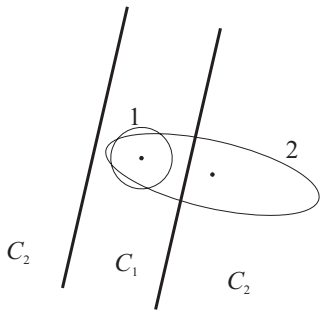
(c) Una parabola.



(d) Un'iperbole.



(e) Due linee.



(f) Due linee parallele.

Figura 3.3. Alcune regioni di classificazione quadratiche per $k = 2$ e $p = 2$.

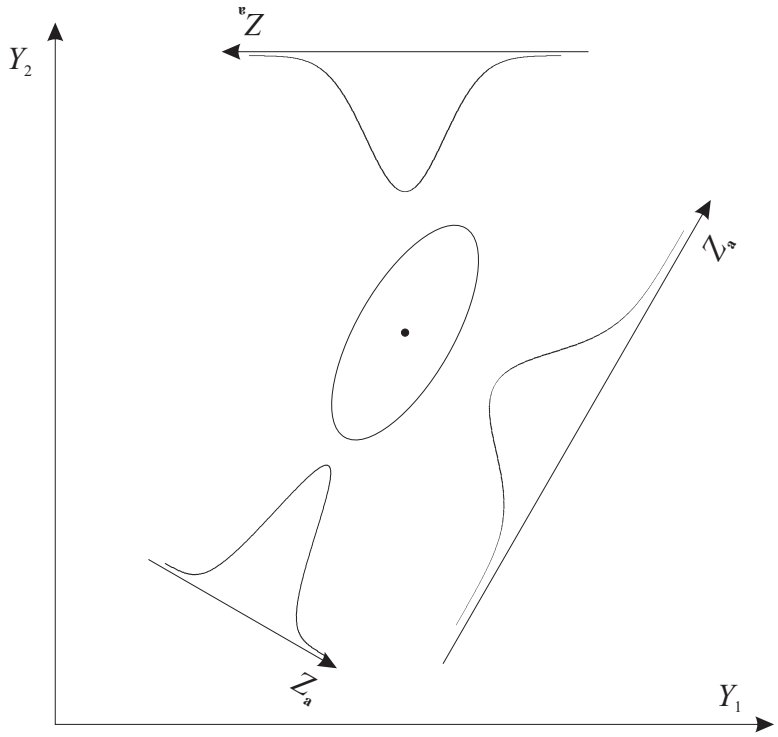


Figura 3.4. Visualizzazione di combinazioni lineari di una distribuzione normale bivariata.

affetti da rumore e che possono non provenire da distribuzioni esattamente multinormali. Si rischia così di tracciare confini artificiosi tra le classi, dando origine ad errori di *overfitting*, o eccessiva aderenza ai dati. Per ovviare a questi fenomeni, si può ricorrere ad una tecnica altrettanto semplice, ma meno sensibile alle sequenze di addestramento, illustrata nel seguente paragrafo.

3.2.5 Classificazione lineare

Innanzitutto, qualche definizione e risultato riguardante la combinazione lineare di variabili casuali. Sia $\mathbf{a}$ un vettore reale di dimensione p. Il prodotto scalare $Z_{\mathbf{a}} = \mathbf{a}'\mathbf{Y}$ proietta il vettore casuale $\mathbf{Y}$ in una variabile (monodimensionale). In altri termini, $Z_{\mathbf{a}}$ è una combinazione lineare delle componenti di $\mathbf{Y}$. Come illustrato dalla figura 3.4, se $\mathbf{Y}$ è un vettore multinormale, $Z_{\mathbf{a}}$ è, per quanto esposto nel paragrafo 3.2.2, una variabile casuale normale. Si deduce inoltre dal paragrafo 3.2.1 che, se si considerano solamente vettori $\mathbf{a}$ normalizzati ($\mathbf{a}'\mathbf{a} = 1$), $Z_{\mathbf{a}}$ ha varianza massima per $\mathbf{a} = \mathbf{u_1}$.

Dalla linearità dell'operatore valore atteso $\mathrm{E}[\cdot]$ segue che

$$\mu_{Z_\mathbf{a}} \stackrel{\text{def}}{=} \mathrm{E}[Z_\mathbf{a}] = \mathbf{a}'\mathrm{E}[\mathbf{Y}] \stackrel{\text{def}}{=} \mathbf{a}'\boldsymbol{\mu}_\mathbf{Y}, \tag{3.36}$$

mentre per la varianza vale

$$\sigma^2_{Z_\mathbf{a}} = \mathbf{a}'\boldsymbol{\Sigma}_\mathbf{Y}\mathbf{a}. \tag{3.37}$$

Più in generale, considerando un insieme di l combinazioni lineari i cui coefficienti stanno in una matrice $\mathbf{A}$ di dimensioni $l \times p$, per il vettore $\mathbf{Z} = \mathbf{A}'\mathbf{Y}$ vale

$$\boldsymbol{\mu}_\mathbf{Z} = \mathbf{A}'\boldsymbol{\mu}_\mathbf{Y} \tag{3.38}$$

$$\boldsymbol{\Sigma}_\mathbf{Z} = \mathbf{A}'\boldsymbol{\Sigma}_\mathbf{Y}\mathbf{A}. \tag{3.39}$$

Ad esempio, se $\mathbf{A}$ è la matrice $\mathbf{U}$ degli autovettori di $\boldsymbol{\Sigma}_\mathbf{Y}$, $\boldsymbol{\Sigma}_\mathbf{Z} = \mathbf{U}'\boldsymbol{\Sigma}_\mathbf{Y}\mathbf{U}$, e si dimostra che è diagonale (ovvero le variabili sono incorrelate), ed i suoi elementi sono gli autovalori.

Naturalmente i risultati fin qui esposti valgono anche per le relative statistiche campionarie $\mathbf{m}$ e $\mathbf{S}$.

Si consideri ora un problema di discriminazione tra $k = 2$ classi, di cui si dispongono i campioni $\mathbf{D}_1$ e $\mathbf{D}_2$. Per semplicità si assume inizialmente che le due classi abbiano uguale probabilità a priori, $\pi_1 = \pi_2 = \frac{1}{2}$. L'*analisi discriminante lineare* (*linear discriminant analysis*, LDA) si propone di trovare il vettore $\mathbf{a}$ per cui la cifra di merito

$$D(\mathbf{a}) = \frac{|\mathbf{a}'\mathbf{m}_1 - \mathbf{a}'\mathbf{m}_2|}{(\mathbf{a}'\mathbf{S}_{\text{pooled}}\mathbf{a})^{1/2}} \tag{3.40}$$

è massima, dove $\mathbf{S}_{\text{pooled}}$ è la matrice di covarianza comune delle due popolazioni, definita nella (3.25). Aiutandosi con la figura 3.5 non è difficile convincersi che la (3.40) cresce al crescere della distanza euclidea tra le medie delle proiezioni monodimensionali delle due popolazioni, e decresce al crescere della loro variabilità. Si dimostra che il vettore cercato è un qualsiasi vettore proporzionale a

$$\mathbf{a}_0 = \mathbf{S}^{-1}_{\text{pooled}}(\mathbf{m}_1 - \mathbf{m}_2), \tag{3.41}$$

in corrispondenza del quale

$$D(\mathbf{a}_0) = ((\mathbf{m}_1 - \mathbf{m}_2)\mathbf{S}^{-1}_{\text{pooled}}(\mathbf{m}_1 - \mathbf{m}_2))^{1/2}, \tag{3.42}$$

che equivale alla distanza standard[8] tra $\mathbf{m}_1$ ed $\mathbf{m}_2$ per una distribuzione

[7] Sebbene nel testo ci si riferisca a due popolazioni (campioni di vettori casuali), in questa figura si visualizzano per semplicità le relative distribuzioni.

[8] Alcuni autori la definiscono proprio come valore massimo di $D(\mathbf{a})$.

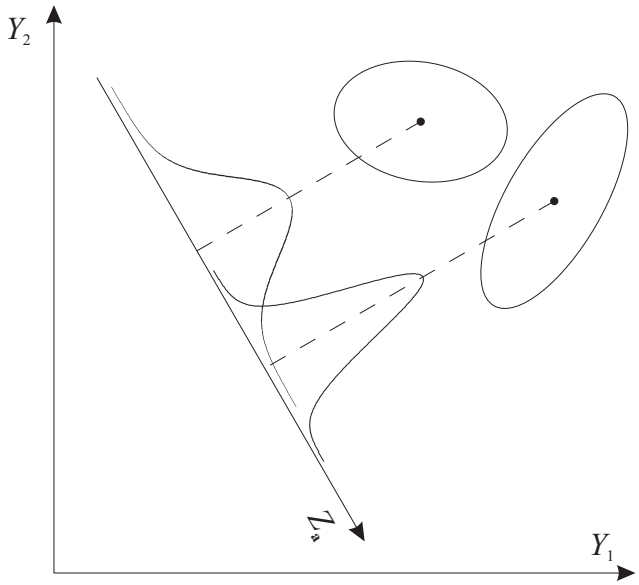


Figura 3.5. Una proiezione di due distribuzioni normali bivariate[7].

avente covarianza $\mathbf{S}_{\text{pooled}}$.

Si introduce, con parziale sovrapposizione rispetto alla definizione di pagina 33, la *funzione discriminante lineare*

$$z(\mathbf{y}; a, b) = a \cdot \mathbf{a}_0' \mathbf{y} + b, \tag{3.43}$$

definita a meno dei due gradi di libertà a e b, dove $\mathbf{y}$ rappresenta l'osservazione da classificare.

L'intenzione è quella di trasformare il vettore di osservazione secondo la (3.43), e classificarla nell'uno o nell'altro gruppo a seconda che lo scalare risultante sia maggiore o minore di un certo valore di soglia f, da determinare. Dal punto di vista geometrico, questo corrisponde a fissare la posizione dell'iperpiano[9] normale alla direzione definita da $\mathbf{a}_0$ in modo che separi il meglio possibile le due classi. Esso è infatti chiamato *iperpiano separatore*. In altri termini, il classificatore cercato ha la forma

$$g^{\text{LDA}}(\mathbf{y}) = \begin{cases} 1 & \text{se } a \cdot \mathbf{a}_0' \mathbf{y} + b > f, \\ 2 & \text{altrimenti.} \end{cases} \tag{3.44}$$

[9]Si tratta di un piano se $p = 3$, di una retta se $p = 2$, di una costante se $p = 1$.

Contrariamente alla scelta di a, quella di b (e congiuntamente di f) è determinante per la regola di decisione, e dipende dalle distribuzioni delle proiezioni delle due classi

$$Z_i = z(\mathbf{Y_i}; a, b) \qquad i = 1, 2. \tag{3.45}$$

Esse sono una combinazione lineare di variabili casuali, e quindi, per il teorema limite centrale, ha senso assumere che siano normali, anche se la distribuzione multivariata delle due popolazioni si discosta da quella multinormale. Un metodo per determinare la soglia potrebbe essere quello di determinare analiticamente il punto di intersezione "più significativo" delle due gaussiane[10] a partire dalle stime dei loro parametri ricavate grazie alla (3.36) e alla (3.37). Nell'analisi discriminante lineare, tuttavia, si assume che le due normali abbiano uguale varianza e si pone così la soglia a metà delle medie

$$f = \frac{m_{Z_1} + m_{Z_2}}{2}. \tag{3.46}$$

Una possibile scelta di a e b è

$$a = 1, \tag{3.47}$$

$$b = -\frac{1}{2}\mathbf{a}_0'(\mathbf{m}_1 + \mathbf{m}_2), \tag{3.48}$$

per cui il punto medio tra m_{Z_1} e m_{Z_2} è pari a zero, e viene scelto come punto di soglia f. Per inciso, dividendo i coefficienti così trovati per $D(\mathbf{a}_0)$, la varianza di Z è unitaria.

Se le probabilità a priori π_j sono disuguali, ovvero se è più probabile osservare esemplari di una delle due classi, la regola della LDA viene generalizzata ponendo la soglia $f = \log(\pi_2/\pi_1)$. Riassumendo

$$g^{\text{LDA}}(\mathbf{y}) = \begin{cases} 1 & \text{se } \mathbf{a}_0'\mathbf{y} - \frac{1}{2}\mathbf{a}_0'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) > \log\left(\frac{\pi_2}{\pi_1}\right), \\ 2 & \text{altrimenti.} \end{cases} \tag{3.49}$$

Equivalentemente, è possibile porre

$$a = 1, \tag{3.50}$$

$$b = -\frac{1}{2}\mathbf{a}_0'(\mathbf{m}_1 + \mathbf{m}_2) + \log\left(\frac{\pi_1}{\pi_2}\right), \tag{3.51}$$

[10] Si ricorda infatti che, a meno che abbiano uguale varianza, esse si intersecano in due punti.

lasciando la soglia $f = 0$. Con quest'ultima scelta dei coefficienti, si ricavano le probabilità a posteriori (3.6) stimate da questo metodo

$$\hat{\pi}_{1\mathbf{y}}^{\text{LDA}} = \frac{\mathrm{e}^{z(\mathbf{y})}}{1 + \mathrm{e}^{z(\mathbf{y})}} \tag{3.52}$$

$$\hat{\pi}_{2\mathbf{y}}^{\text{LDA}} = 1 - \hat{\pi}_{1\mathbf{y}}. \tag{3.53}$$

Si dimostra, come ci si poteva aspettare dalla figura 3.3, che l'analisi discriminante lineare corrisponde al classificatore ottimo solo nel caso in cui le due distribuzioni multinormali abbiano uguale matrice di covarianza (*omoschedasticità*).

L'analisi discriminante lineare, come evidenziato da Fisher, è strettamente imparentata con la regressione lineare. È infatti possibile mostrare che i due problemi sono riconducibili uno all'altro, e quindi equivalenti. Non stupisce perciò che si possano adattare risultati di un campo già disponibili in letteratura per applicarli nell'altro.

3.2.6 Analisi discriminante canonica

È possibile estendere la teoria della LDA ad un numero $k \geq 2$ di classi. Si parla in questo caso di *analisi discriminante canonica* (*canonical discriminant analysis*, CDA), o *analisi discriminante lineare multipla* (*multiple linear discriminant analysis*), e la quantità che si cerca di rendere massima, con riferimento alle quantità introdotte nella (3.29), è

$$D(\mathbf{a}) = \frac{\mathbf{a}'\mathbf{S}_{\text{P Tot}}\mathbf{a}}{\mathbf{a}'\mathbf{S}_{\text{W}}\mathbf{a}}. \tag{3.54}$$

Altrimenti detto, si cerca la combinazione lineare (avente come coefficienti gli elementi di $\mathbf{a}$) che massimizzi il rapporto tra la varianza della proiezione della popolazione totale e la varianza delle proiezioni delle singole classi.

La soluzione è piuttosto complessa, e fa uso di concetti e risultati avanzati di algebra lineare, e ci si limita ad esporre per completezza l'algoritmo per il calcolo di $\mathbf{a}$, riportandolo impudentemente da [24].

La *diagonalizzazione simultanea* (o *scomposizione simultanea*) di una matrice simmetrica e definita positiva $\mathbf{W}$ ed una matrice simmetrica $\mathbf{A}$ (aventi uguali dimensioni) è definita come una coppia di matrici $(\mathbf{H}, \mathbf{\Lambda})$, aventi uguali dimensioni e con $\mathbf{\Lambda}$ diagonale, tale che

$$\mathbf{W} \stackrel{\text{def}}{=} \mathbf{H}\mathbf{H}', \tag{3.55}$$

$$\mathbf{A} \stackrel{\text{def}}{=} \mathbf{H}\mathbf{\Lambda}\mathbf{H}'. \tag{3.56}$$

Siano $\widetilde{\mathbf{\Lambda}}$ e $\widetilde{\mathbf{U}}$ le matrici di scomposizione spettrale di

$$\mathbf{W} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{U}} = \sum_{j=1}^{p} \tilde{\lambda}_i \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i', \tag{3.57}$$

cioè, rispettivamente, la matrice degli autovalori (diagonale) e degli autovettori di $\mathbf{W}$. Un metodo efficiente per il calcolo della matrice radice quadrata di una matrice simmetrica definita positiva è

$$\mathbf{W}^{1/2} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}^{1/2}\widetilde{\mathbf{U}} = \sum_{j=1}^{p} \tilde{\lambda}_i^{1/2} \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i'. \tag{3.58}$$

Siano $\widehat{\mathbf{\Lambda}}$ e $\widehat{\mathbf{U}}$ le matrici di scomposizione spettrale della matrice simmetrica $\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}$. Si dimostra che le matrici cercate sono date da

$$\mathbf{H} = \mathbf{W}^{1/2}\widehat{\mathbf{U}}, \tag{3.59}$$

$$\mathbf{\Lambda} = \widehat{\mathbf{\Lambda}}. \tag{3.60}$$

Siano ora $\mathbf{\Lambda}$ e $\mathbf{H}$ le matrici di scomposizione simultanea delle matrici $\mathbf{S}_{\mathrm{W}}$ (simmetrica e definita positiva) e $\mathbf{S}_{\mathrm{B}}$, e si riordinino le loro colonne secondo l'ordine decrescente degli autovalori presenti sulla diagonale di $\mathbf{\Lambda}$. Sia

$$\mathbf{\Gamma} \stackrel{\text{def}}{=} \left[\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_p\right] \stackrel{\text{def}}{=} \left(\mathbf{H}'\right)^{-1}. \tag{3.61}$$

Definendo

$$m \stackrel{\text{def}}{=} \min(p, k-1), \tag{3.62}$$

si dice *j-esima variata canonica* ognuna delle m combinazioni lineari delle osservazioni date dalle

$$z_j(\mathbf{y}) \stackrel{\text{def}}{=} \boldsymbol{\gamma}_j \mathbf{y} \qquad 1 \leq j \leq m, \tag{3.63}$$

ed è unica a meno di una costante proporzionale. Secondo la notazione matriciale

$$\check{\mathbf{\Gamma}} \stackrel{\text{def}}{=} \left[\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_m\right], \tag{3.64}$$

$$\mathbf{z}(\mathbf{y}) \stackrel{\text{def}}{=} \check{\mathbf{\Gamma}}'\mathbf{y}. \tag{3.65}$$

Si è perciò proiettato lo spazio campionario su uno spazio m-dimensionale, in modo che la separazione tra le classi sia massima. Si dimostra infatti che la

quantità (3.54) è massima per $\mathbf{a} = \boldsymbol{\gamma}_1$. Inoltre, sotto il vincolo di ortogonalità di $\mathbf{a}$ rispetto alle $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{j-1}$, la (3.54) è massima per $\mathbf{a} = \boldsymbol{\gamma}_j$.

Siano

$$\mathbf{n}_j \stackrel{\text{def}}{=} \check{\boldsymbol{\Gamma}}' \mathbf{m}_j \qquad 1 \leq j \leq k \tag{3.66}$$

le proiezioni delle medie campionarie delle classi in questo spazio. La CDA, tenendo conto anche delle probabilità a priori π_j, adotta le seguenti funzioni discriminanti

$$g_j^{\text{CDA}}(\mathbf{y}) \stackrel{\text{def}}{=} \mathbf{n}_j' \left(\mathbf{z}(\mathbf{y}) - \frac{1}{2}\mathbf{n}_j \right) + \log \pi_j \qquad 1 \leq j \leq k, \tag{3.67}$$

cioè classifica l'osservazione $\mathbf{y}$ nel gruppo j-esimo per cui $g_j^{\text{CDA}}(\mathbf{y})$ è maggiore rispetto alle altre valutazioni $g_i^{\text{CDA}}(\mathbf{y})$, per $i \neq j$. Al solito, le stime delle probabilità a posteriori sono date da

$$\hat{\pi}_{j\mathbf{y}}^{\text{CDA}} = \frac{\exp\left(g_j^{\text{CDA}}(\mathbf{y})\right)}{\sum_{h=1}^{k} \exp\left(g_h^{\text{CDA}}(\mathbf{y})\right)} \qquad 1 \leq j \leq k. \tag{3.68}$$

Anche in questo caso più generale si dimostra che l'approccio lineare è ottimo solo se le matrici di covarianza delle classi sono identiche. Tuttavia, come si analizzerà più in dettaglio nel paragrafo 3.2.8, il metodo della CDA si comporta dignitosamente anche in condizioni di eteroschedasticità.

3.2.7 Stime del tasso di errore

Una volta ottenuto un classificatore, è prassi comune verificarne la validità stimando la probabilità di errore introdotta nella

$$\gamma_g \stackrel{\text{def}}{=} \Pr[X_g \neq X]. \tag{3.3}$$

Si riportano due semplici stime basate sui dati di *training*. Ricordando le definizioni (3.20) e (3.21), si introduce il vettore

$$\mathbf{x} \stackrel{\text{def}}{=} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \tag{3.69}$$

che memorizza nella sua i-esima componente il numero associato alla classe di appartenenza dell'i-esima osservazione.

Si definisce infine la quantità e_i, che vale 0 se il classificatore identifica la i-esima osservazione correttamente, 1 altrimenti

$$e_i \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } x_i = g(\mathbf{d}_i) \\ 1 & \text{se } x_i \neq g(\mathbf{d}_i). \end{cases} \tag{3.70}$$

La stima più semplice del tasso di errore γ_g, chiamata stima *plug-in*, o *tasso di errore apparente*, è data da

$$\hat{\gamma}_{\text{plug-in}} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} e_i. \tag{3.71}$$

Essa si rivela spesso ottimistica, poiché, di fatto, confonde i dati di *training* con quelli di convalida (o *assessment*), aderendo troppo ai dati osservati. In questi casi, invece, si preferisce generalmente utilizzare una parte dei dati disponibili per la fase di addestramento, e la rimanente parte per la convalida (una ripartizione molto popolare è 70% e 30%, rispettivamente). Si parla in questo caso di convalida incrociata, o *cross-validation*. Estremizzando questo argomento, si perviene alla definizione della stima che in letteratura prende il nome di *leave-one-out*, letteralmente "lasciane fuori uno." Senza definirlo formalmente, si introduce il classificatore $g_{-i}(\cdot)$, ottenuto escludendo dalla sequenza di addestramento $\mathbf{D}$ la sola osservazione i-esima. Il passo successivo è la definizione di

$$e_{i,-i} \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } x_i = g_{-i}(\mathbf{d}_i) \\ 1 & \text{se } x_i \neq g_{-i}(\mathbf{d}_i). \end{cases} . \tag{3.72}$$

È ora possibile battezzare

$$\hat{\gamma}_{\text{leave-one-out}} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} e_{i,-i}, \tag{3.73}$$

ovvero la somma degli errori di classificazione di ciascun elemento sul classificatore addestrato senza di esso. Evidentemente, il computo di questa stima è enormemente più oneroso di quello della stima *plug-in*, in quanto prevede N addestramenti diversi del classificatore in esame. Generalmente, specie se serve una convalida in tempi brevi, si preferisce adottare soluzioni intermedie come quella accennata poc'anzi.

3.2.8 Classificazione lineare e quadratica a confronto

Come era stato anticipato alla fine del paragrafo 3.2.4, anche per classi aventi distribuzioni multinormali, le prestazioni della LDA (o della CDA) possono essere superiori rispetto al classificatore "ottimo" della QDA. Il paradosso è generato dal fatto che le stime dei parametri delle distribuzioni vengono effettuate su un numero *finito* di campioni, e sono quindi affette da errore. Quanto detto può essere suffragato da evidenze sperimentali, calcolando $\hat{\gamma}_{\text{leave-one-out}}$ per entrambi i classificatori (si veda ad esempio [24, esempio 7.2.4]).

La scelta tra i due metodi si basa su diversi fattori [66]. Se ne elencano alcuni:

1. Dal punto di vista della complessità computazionale della sola fase di identificazione (la più critica per applicazioni in tempo reale) per la QDA è $\Theta\left(k(p^2+2p)\right)$, cioè $\Theta(kp^2)$, mentre per la CDA è $\Theta\left(k(p+mp+m)\right)$, cioè $\Theta(kpm)$, con m definita nella (3.62). Come si vede, sotto questo aspetto i metodi sono pressoché equivalenti.

2. La CDA si comporta meglio della QDA se i dati disponibili per la fase di addestramento sono pochi, in quanto aderisce meno al rumore inevitabilmente presente.

3. La CDA è ottima in caso di omoschedasticità. Se le matrici di covarianza sono "molto diverse" (si veda [23] per una trattazione monografica sulla comparazione tra matrici di covarianza) conviene quindi spostarsi verso la QDA.

4. L'approssimazione introdotta dalla CDA è più marcata se le classi sono poco separate. Viceversa, in queste situazioni la QDA è preferibile, in quanto disegna regioni di classificazione più accurate.

5. La robustezza rispetto alla non-normalità delle distribuzioni è un elemento di decisione fondamentale in presenza di aberrazioni ("la mappa non è il territorio").

   Code corte Se le classi sono più compatte rispetto ad una multinormale, le prestazioni non peggiorano in ogni caso.

   Code lunghe Questo tipo di deformazione compromette i risultati di entrambe le tecniche. Tuttavia, se sono disponibili parecchi dati di addestramento e non ci sono evidenti asimmetrie, conviene scegliere la QDA.

   Curtosi Se le popolazioni hanno forme concave ("a fagiolo") molto pronunciate, la QDA fornisce risultati migliori.

   Asimmetrie (*skewness*) Entrambe le tecniche si comportano bene, ma la CDA è da preferirsi, specie se il difetto non è accentuato.

Esistono in letteratura numerose proposte di compromesso tra i due metodi esposti. Ad esempio Friedman [30] propone l'*analisi discriminante regolarizzata*, operando una sorta di combinazione lineare convessa tra QDA e CDA. Infine, una alternativa piuttosto popolare nello stesso ambito è rappresentata dalla *regressione logistica* [15].

3.2.9 Cenni ad altre tecniche di classificazione

Si accennano in questo paragrafo ad alcune altre tecniche di classificazione, rimandando alla letteratura per eventuali approfondimenti. Per semplicità, si assume di voler discriminare tra $k = 2$ classi.

k-nearest neighbor (k-NN)

Sia k un numero dispari[11]. Data l'osservazione di natura incognita $\mathbf{y}$, siano

$$\mathbf{d}_j^{\mathrm{NN}} \qquad 1 \leq j \leq k \tag{3.74}$$

le osservazioni appartenenti alla sequenza di *training* che sono più vicine ad $\mathbf{y}$. La regola di decisione attribuisce $\mathbf{y}$ alla classe che contiene il maggior numero di queste osservazioni. L'implementazione banale di questa tecnica, ovvero quella che calcola la distanza tra $\mathbf{y}$ e tutti dati della sequenza di addestramento, non brilla per efficienza, e per questo sono stati messi a punto algoritmi migliori. La scelta della particolare metrica adottata dipende dal problema in esame e dalla morfologia delle classi.

Kernel rules

È la tecnica duale della *k-nearest neighbor*. Si consideri una regione dello spazio p-dimensionale centrata in $\mathbf{y}$ e di forma arbitraria[12], e siano

$$\mathbf{d}_j^{\mathrm{KR}} \qquad 1 \leq j \leq t \tag{3.75}$$

le t osservazioni appartenenti alla sequenza di *training* che cadono in questa regione. La regola di decisione attribuisce $\mathbf{y}$ alla classe che contiene il maggior numero di queste osservazioni. Valgono le stesse considerazioni di efficienza e flessibilità esposte per la tecnica di *k-nearest neighbor*.

Support vector machines, SVM

Similmente alla LDA, esposta in dettaglio nel paragrafo 3.2.5, questa tecnica separa le classi attraverso degli iperpiani (uno solo, nel semplice caso di due classi). Essa, però, non richiede alcuna assunzione riguardo alle loro distribuzioni. L'iperpiano viene fissato in modo da rendere massima la sua distanza con l'osservazione più vicina. Se le due classi non sono separabili da un iperpiano, questo semplice criterio non porta ad alcuna soluzione, e quindi lo si

[11] La collisione con il nome della variabile utilizzata nel resto del capitolo per indicare il numero di classi non è stata risolta, a causa della diffusione del nome di questa tecnica.

[12] Generalmente di utilizzano ipersfere, iperellissoidi, o parallelepipedi in p dimensioni.

modifica introducendo delle penalizzazioni in caso di classificazione erronea. L'estensione del metodo a $k > 2$ classi non è univoca. Per un *tutorial* su questa tecnica si veda [9].

3.2.10 Test per le proprietà dei campioni

Un elemento importante nella realizzazione di un classificatore è rappresentato dai test statistici e da quantificazioni di determinate proprietà dei campioni. Ad esempio, ci si potrebbe chiedere quanto "normali" sono i dati, o da che grado di curtosi sono affetti, per adottare un modello adeguato di classificazione. Si riportano di seguito una serie di procedure di test e statistiche per campioni multinormali tratte da [24, 66].

Test 1 $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ [24, sezione 6.2]

Si vuole testare l'ipotesi che la media del vettore casuale di cui il campione in esame è una realizzazione abbia media $\boldsymbol{\mu}_0$. Si supponga che il campione sia composto di N osservazioni, abbia media $\mathbf{m}$ e matrice di covarianza $\mathbf{S}$. Hotelling [45] propone la seguente statistica, generalizzazione della t di Student[13] del caso monovariato

$$T^2 \stackrel{\text{def}}{=} N \cdot (\boldsymbol{\mu}_0 - \mathbf{m})' \, \mathbf{S}^{-1} \, (\boldsymbol{\mu}_0 - \mathbf{m}) = N \cdot D^2(\boldsymbol{\mu}_0, \mathbf{m}), \tag{3.76}$$

dove

$$D(\mathbf{y}_1, \mathbf{y}_2) \stackrel{\text{def}}{=} \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)' \, \mathbf{S}^{-1} \, (\mathbf{y}_1 - \mathbf{y}_2)} \tag{3.77}$$

rappresenta la versione campionaria della distanza standard.

Sia f il $(1 - \alpha)$-quantile della distribuzione F con p ed $(N - p)$ gradi di libertà. Si accetti l'ipotesi H_0 se

$$T^2 \leq \frac{p(N-1)}{N-p} f \tag{3.78}$$

o, equivalentemente, se

$$D(\boldsymbol{\mu}_0, \mathbf{m}) \leq \sqrt{\frac{p(N-1)}{N(N-p)} f} \, . \tag{3.79}$$

Si dimostra che, se le classi sono multinormali e possiedono la stessa matrice di covarianza, la probabilità condizionale di rifiutare l'ipotesi nulla nel caso in cui sia vera è α (*test di livello* α).

[13] Pseudonimo che W. S. Gossett fu costretto ad usare nel 1908 per poter pubblicare i propri risultati.

Test 2 $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ [24, sezione 6.4]

Si desidera testare l'ipotesi che due campioni hanno media coincidente. Si accetti l'ipotesi se

$$T^2 \leq \frac{p(N_1 + N_2 - 2)}{N_1 + N_2 - p - 1} f \tag{3.80}$$

o, equivalentemente, se

$$D(\boldsymbol{\mu}_0, \mathbf{m}) \leq \sqrt{\frac{p(N_1 + N_2)(N_1 + N_2 - 2)}{N_1 N_2 (N_1 + N_2 - p - 1)} f}\,, \tag{3.81}$$

con N_1 ed N_2 pari al numero di osservazioni delle due popolazioni, ed f pari all'$(1-\alpha)$-quantile della distribuzione F con p ed $(N_1 + N_2 - p - 1)$ gradi di libertà. Anche in questo caso si dimostra che, se le classi sono multinormali e possiedono la stessa matrice di covarianza, la probabilità condizionale di rifiutare l'ipotesi nulla nel caso in cui sia vera è α.

Questo test ha diverse applicazioni. Ad esempio, può servire per verificare che due classi distinte non differiscano solo per "rumore statistico" (test di significatività globale), oppure per assicurarsi che due sessioni di campionamento distinte della stessa classe presentino medie campionarie "sufficientemente vicine."

Test 3 $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k$ [24, sezione 7.4]

Si desidera testare l'ipotesi che k classi aventi distribuzioni multinormali ed omoschedastiche hanno media coincidente, generalizzando il test 2. Con riferimento alle quantità introdotte nella (3.29), si dimostra che la statistica dei rapporti di massima verosimiglianza logaritmica (*log-likelihood ratio statistic*) è data da

$$\mathrm{LLRS_m} = N \log \det(\mathbf{S}_{\mathrm{W}}^{-1} \mathbf{S}_{\mathrm{P\,Tot}}). \tag{3.82}$$

Conseguentemente, dalla teoria degli stimatori di massima verosimiglianza (si veda, ad esempio, [24, sezione 4.3]) si ha che, asintoticamente, $\mathrm{LLRS_m} \sim \chi^2_d$, dove $d = p(k-1)$. Operativamente, si accetti l'ipotesi di ridondanza se e solo se $\mathrm{LLRS_m} \leq c$, dove c è il $(1-\alpha)$-quantile della distribuzione χ^2_d: la decisione è corretta con probabilità approssimativamente pari ad α.

Si dimostra che la (3.82) può essere anche espressa come

$$\mathrm{LLRS_m} = N \sum_{j=1}^{m} \log(1 + \hat{\lambda}_j), \tag{3.83}$$

dove le $\hat{\lambda}_j$ sono gli autovalori della matrice $\mathbf{S}_{\mathrm{W}}^{-1} \mathbf{S}_{\mathrm{B}}$, e m è definito nella (3.62).

Test 4 $H_0 : \mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Si desidera testare la normalità della variabile casuale p-variata $\mathbf{Y}$ dato un suo campione. Si calcolino le quantità

$$c_j \stackrel{\text{def}}{=} \frac{(N-p-1)ND(\mathbf{d}_j, \mathbf{m})}{p\left[(N-1)^2 - ND(\mathbf{d}_j, \mathbf{m})\right]}. \tag{3.84}$$

Si dimostra che le c_j seguono una distribuzione F con p e $(N-p-1)$ gradi di libertà. Se a_j denota l'area alla destra di c_j sottesa dalla $F_{p,N-p-1}$, sotto l'ipotesi nulla si ha che

$$a_1, \ldots, a_N \overset{iid}{\sim} \mathcal{U}(0,1), \tag{3.85}$$

ovvero le a_j sono indipendenti ed uniformemente distribuite in $(0,1)$.

Test 5 $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_k$ [24, sezione 6.6 esercizio 13]

Una misura di omoschedasticità tra k distribuzioni normali p-variate, tra le tante disponibili, è data da

$$\text{LLRS}_\text{h} \stackrel{\text{def}}{=} N \log(\det \mathbf{S}_\text{W}) - \sum_{j=1}^{k} N_j \log(\det \mathbf{S}_{\text{P}j}), \tag{3.86}$$

con $\mathbf{S}_\text{P}$ e $\mathbf{S}_\text{W}$ definiti nella (3.23) e nella (3.29), rispettivamente. Si dimostra che la (3.86) corrisponde alla statistica dei rapporti di massima verosimiglianza logaritmica. Dalla teoria della stima di massima verosimiglianza si deriva quindi che, asintoticamente, LLRS $\sim \chi^2_d$, con $d = (k-1)p(p+1)/2$.

Test 6 $H_0 : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ [24, sezione 6.6 esercizio 14]

Si desidera testare l'uguaglianza di media e matrice di covarianza di due distribuzioni normali p-variate. Si dimostra che

$$\text{LLRS}_\text{eq} \stackrel{\text{def}}{=} N \log(\det \mathbf{S}_{\text{P Tot}}) - \sum_{j=1}^{2} N_j \log(\det \mathbf{S}_{\text{P}j}), \tag{3.87}$$

dove $\mathbf{S}_{\text{P Tot}}$ è la matrice di covarianza complessiva introdotta nella (3.29), corrisponde alla statistica dei rapporti di massima verosimiglianza logaritmica. Si deriva quindi che, asintoticamente, LLRS $\sim \chi^2_d$, con $d = p(p+3)/2$.

Test 7 $H_0 : \mathbf{Y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ [66, sezione 6.2.2]

Si desidera testare la normalità e, congiuntamente, l'omoschedasticità delle variabili casuali p-variate $\mathbf{Y}_i$ dati i loro campioni. Si tratta di una generalizzazione del test 4 [42].

Sia

$$\nu \stackrel{\text{def}}{=} N - p - k. \tag{3.88}$$

Si calcolino le quantità

$$c_{ij} \stackrel{\text{def}}{=} \frac{\nu N_i D(\mathbf{d}_{ij}, \mathbf{m}_i)}{p\left[(\nu + p)(N_i - 1) - N_i D(\mathbf{d}_{ij}, \mathbf{m}_i)\right]} \quad 1 \leq i \leq k, \tag{3.89}$$

con $D(\cdot, \cdot)$ distanza standard campionaria basata sulla matrice di covarianza comune $\mathbf{S}_{\text{pooled}}$, definita nella (3.25). Si dimostra che le c_{ij} seguono una distribuzione F con p e ν gradi di libertà. Se a_{ij} denota l'area alla destra di c_{ij} sottesa dalla $F_{p,\nu}$, sotto l'ipotesi nulla si ha che

$$a_{i1}, \ldots, a_{iN_i} \stackrel{iid}{\sim} \mathcal{U}(0, 1) \quad 1 \leq i \leq k. \tag{3.90}$$

Test 8 Misura di asimmetria (*skewness*) [66, sezione 6.2.3]

Si calcoli la quantità [58]

$$S \stackrel{\text{def}}{=} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[(\mathbf{d}_i - \mathbf{m})' \mathbf{S}^{-1} (\mathbf{d}_j - \mathbf{m})\right]^3. \tag{3.91}$$

Sotto le ipotesi di multinormalità si ha che, asintoticamente,

$$\frac{N}{6} S \sim \chi^2_d, \tag{3.92}$$

con $d = (p/6)(p + 1)(p + 2)$. Si osservi che questo indice di asimmetria, al contrario del corrispondente indice scalare ($p = 1$), non fornicse alcuna informazione riguardo alla direzione della eventuale asimmetria.

Test 9 Misura di curtosi [66, sezione 6.2.3]

Si calcoli la quantità [58]

$$K \stackrel{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} \left[(\mathbf{d}_j - \mathbf{m})' \mathbf{S}^{-1} (\mathbf{d}_j - \mathbf{m})\right]^2. \tag{3.93}$$

Sotto le ipotesi di multinormalità si ha che, asintoticamente,

$$\frac{K - p(p + 2)}{\sqrt{p(p + 2)(8/N)}} \sim \mathcal{N}_p(0, 1). \tag{3.94}$$

I test 8 e 9 possono vedersi come test alternativi per la normalità.

3.2.11 Test per la ridondanza di variabili

Nell'ambito di una analisi discriminante, o di una procedura di classificazione automatica in generale, può aver senso chiedersi se alcune variabili, si supponga in un numero q, non siano ridondanti, cioè se non aggiungano alcuna informazione utile alla classificazione rispetto all'insieme delle rimanenti $p-q$. La risposta a questa domanda si rivela di notevole interesse se alcune variabili sono particolarmente costose o difficili da ottenere. La restrizione del numero di variabili si rivela infine necessaria se è disponibile una casistica limitata per la fase di addestramento. Si dimostra infatti [14] che, a parità di tasso di errore, un classificatore necessita di una sequenza di *training* che cresce esponenzialmente con il numero delle variabili p.

Un possibile modo di procedere è quello di valutare, in presenza e in assenza delle variabili in esame, una cifra di merito, per esempio una stima del tasso di errore, o una misura della separazione tra le classi, come la (3.42) per la LDA, o la prima variata canonica per la CDA.

Paradossalmente, l'eliminazione di variabili che singolarmente presentano un basso indice di separazione può rivelarsi una pessima idea. Viceversa, è possibile che una variabile con un alto contenuto informativo ai fini della classificazione sia superflua se utilizzata con altre variabili. Per un'illuminante illustrazione di questo fenomeno si veda [24, esempio 5.3.3].

Nel semplice caso in cui si abbiano $k = 2$ classi multinormali e con uguale matrice di covarianza, sono disponibili due test di ridondanza di variabili.

Test 10 H_0 : la variabile Y_j è ridondante [24, teorema 6.5.1]

Siano, al solito, N_1 ed N_2 il numero di osservazioni per ogni classe; sia D la distanza standard tra le due medie definita dalla (3.42), e sia D_{-j} la distanza standard calcolata senza fare uso della variabile Y_j. Si calcoli

$$|t_j| = \sqrt{(N_1 + N_2 - p - 1) \cdot \frac{D^2 - D_{-j}^2}{m + D_{-j}^2}}\,, \tag{3.95}$$

dove

$$m \stackrel{\text{def}}{=} \frac{(N_1 + N_2)(N_1 + N_2 - 2)}{N_1 N_2}. \tag{3.96}$$

Si dimostra che, per classi multinormali e omoschedastiche, la quantità t_j segue una distribuzione t con $(N_1 + N_2 - p - 1)$ gradi di libertà. Operativamente, si accetti l'ipotesi di ridondanza se e solo se $|t_j| \leq c$, dove c è il $(1-\alpha/2)$-quantile della distribuzione t con (N_1+N_2-p-1) gradi di libertà: la decisione è corretta con probabilità α.

È possibile calcolare D_{-j} senza dover affrontare una analisi discriminante da zero. Sia infatti a_{0j} la j-esima componente del vettore $\mathbf{a}_0$ calcolato nella (3.41), e sia $\check{s}_{jj}$ il j-esimo elemento diagonale della matrice $\mathbf{S}_{\text{pooled}}^{-1}$. Si verifica che

$$D_{-j}^2 = D^2 - \frac{a_{0j}^2}{\check{s}_{jj}}. \tag{3.97}$$

Test 11 H_0 : le ultime $(p-q)$ variabili Y_j $(q+1 \leq j \leq p)$ sono ridondanti [24, teorema 6.5.2]

Sia D_p la distanza standard tra le due medie definita dalla (3.42), e sia D_q la distanza standard ottenuta utilizzando solo le prime q variabili. Si calcoli

$$R_q = \frac{N_1 + N_2 - p - 1}{p - q} \cdot \frac{D_p^2 - D_q^2}{m + D_q^2}, \tag{3.98}$$

con m definita come nella (3.96). Si dimostra che, per classi multinormali e omoschedastiche, la quantità R_q segue una distribuzione F con $(p-q)$ e $(N_1 + N_2 - p - 1)$ gradi di libertà.

Quest'ultimo test riguarda sottoinsiemi di variabili, ed è una generalizzazione di entrambi i test 2 e 10, in cui si consideravano rispettivamente singoletti e l'insieme vuoto. Si osserva che una ricerca esaustiva su tutti i sottoinsiemi ha complessità $\Theta(2^p)$, e può diventare facilmente intrattabile al crescere di p. Volendo trovare il sottoinsieme di cardinalità $q < p$ dell'insieme delle variabili disponibili, tale che il suo tasso di errore sia minimo, si dimostra [14, teorema 32.1] che è necessario, nel caso generale, cercarlo esaustivamente tra tutti i $\binom{q}{p}$ sottoinsiemi di cardinalità q. L'unica cosa certa è che, aumentando il numero di variabili, il tasso di errore del classificatore *ottimo* non decresce (ma, come già detto, il tasso di errore può aumentare per una determinata regola). Si rendono perciò necessarie delle procedure euristiche di selezione delle variabili, riportate alla fine del seguente test.

Test 12 H_0 : la variabile Y_j è ridondante (per la classificazione tra $k > 2$ classi) [66]

Si intende generalizzare il test 10 per un numero arbitrario di classi, sempre in condizioni di omoschedasticità. Sia $j = p$, ovvero, senza perdita di generalità, si riordinino le variabili in modo che quella in esame sia l'ultima. Siano

$$\mathbf{W} \stackrel{\text{def}}{=} (N-k)\mathbf{S}_{\text{pooled}}, \tag{3.99}$$

$$\mathbf{B} \stackrel{\text{def}}{=} (k-1)\mathbf{S}_{\text{B}}, \tag{3.100}$$

dove $\mathbf{S}_{\text{pooled}}$ è la matrive di covarianza comune generalizzata definita nella (3.26). Si suddividano inoltre queste matrici

$$\mathbf{W} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}, \tag{3.101}$$

$$\mathbf{B} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{3.102}$$

isolando le componenti relative alla p-esima variabile (gli scalari $\mathbf{W}_{22}$ e $\mathbf{B}_{22}$). Si definiscono le quantità scalari[14]

$$\mathbf{W}_{2.1} \stackrel{\text{def}}{=} \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}, \tag{3.103}$$

$$\mathbf{B}_{2.1} \stackrel{\text{def}}{=} (\mathbf{B}_{22} + \mathbf{W}_{22}) - (\mathbf{B}_{21} + \mathbf{W}_{21})(\mathbf{B}_{11} + \mathbf{W}_{11})^{-1}(\mathbf{B}_{12} + \mathbf{W}_{12}) - \mathbf{W}_{2.1}, \tag{3.104}$$

$$\Lambda_{-p} \stackrel{\text{def}}{=} \frac{|\mathbf{W}_{2.1}|}{|\mathbf{W}_{2.1} + \mathbf{B}_{2.1}|}. \tag{3.105}$$

Si dimostra che, sotto l'ipotesi di ridondanza, la statistica

$$\Lambda \stackrel{\text{def}}{=} \frac{(N - k - p + 1)(1 - \Lambda_{-p})}{(k - 1)\Lambda_{-p}} \tag{3.106}$$

segue una distribuzione F con $(k - 1)$ e $(N - k - p - 1)$ gradi di libertà.

A partire da questo test, è possibile costruire delle procedure euristiche passo-passo per la determinazione di un sottoinsieme subottimo di variabili [66, sezione 12.3.3]. Le due tecniche fondamentali sono la selezione in avanti (*forward selection*) e l'eliminazione all'indietro (*backward elimination*). La prima, partendo dall'insieme vuoto, aggiunge all'insieme corrente la variabile che presenta il valore massimo per la (3.106), a patto che superi un valore di soglia, ad esempio il $(1 - \alpha)$-quantile di $F_{k-1,N-k-p+1}$. Viceversa, nell'eliminazione all'indietro, partendo dall'insieme di tutte le variabili vengono eliminate una ad una quelle che presentano il valore minimo per la (3.106), a patto che sia sotto il valore di soglia.

[14] Si preferisce mantenere la notazione matriciale per uniformarsi a quella di [66], che tratta il caso più generale di ridondanza di $q \geq 1$ variabili.

3.3 Analisi dei cluster

L'*analisi dei cluster* (*cluster analysis*, a volte tradotta come *analisi dei grappoli*) è una tencica nata e diffusasi negli anni 60 e 70, mirata all'individuazione di agglomerati di dati all'interno di una popolazione nota. Gli obiettivi finali possono essere i più disparati, ad esempio l'individuazione o la convalida di un'ipotesi di ricerca a partire dai dati, l'isolamento di *pattern* caratteristici in determinate sotto-popolazioni, o la classificazione dei dati. In questa sezione verranno esposti i principali strumenti dell'analisi dei *cluster* con quest'ultimo scopo in mente.

L'analisi dei *cluster* si basa su procedure semplici e facilmente automatizzabili, fa largo uso di euristiche e poggia su una matematica piuttosto elementare. Per questi motivi, essa è spesso snobbata dagli statistici, che la vedono come il "fratello povero" [24, pagina 123] dell'*analisi delle misture finite*, strettamente connessa all'analisi discriminante, dalle basi teoriche certamente più solide e rigorose. D'altra parte, proprio la sua semplicità ne ha favorito la diffusione tra i ricercatori delle scienze naturali, e la leggibilità dei suoi risultati, l'alto potenziale euristico (appunto) e la disponibilità di numerosi strumenti di analisi automatica ne fanno uno strumento valido e meritevole di considerazione.

Nella tradizione di questa disciplina, ma la definizione potrebbe essere estesa alla classificazione in generale, si distingue tra tecniche di tipo Q e tecniche di tipo R. Nel primo caso, come si assume in questo intero capitolo se non specificato diversamente, vengono analizzate e classificate le *osservazioni*, mentre nelle tecniche di tipo R vengono esaminate le variabili, ad esempio, come si è fatto nella sezione 3.2.11, per eliminare variabili superflue o dallo scarso contenuto informativo.

Le tecniche di analisi dei *cluster* possono suddividersi in due ampie categorie: i metodi di ripartizione e i metodi gerarchici. Prima della loro trattazione si introducono le principali procedure di trasformazione delle variabili.

3.3.1 Trasformazione delle variabili e normalizzazione

Sebbene la normalizzazione, e in generale la trasformazione delle variabili, possa essere utilizzata anche nei metodi statistici, viene introdotta in questo contesto, in quanto nei metodi presentati nel paragrafo 3.2 non era indispensabile. Viceversa, nell'analisi dei *cluster* il suo utilizzo è fortemente consigliato, poiché rende il risultato indipendente dalle unità di misura adottate per le variabili. Inoltre, la normalizzazione fa sì che tutte le variabili contribuiscano in ugual misura alla classificazione.

Per *trasformazione* di una variabile, o attributo, si intende la derivazione

di nuove variabili attraverso l'applicazione di funzioni a quelle originarie. In formula

$$Y_i' \stackrel{\text{def}}{=} f_i(Y_i) \qquad 1 \leq i \leq p. \tag{3.107}$$

In alcuni casi può essere utile applicare ad alcune variabili delle trasformazioni non lineari, al fine di correggerne la distorsione. Le più usate sono $\log(\cdot)$, $\log(\cdot + 1)$, $\sqrt{\cdot}$, $\arctan(\cdot)$ e $\cosh(\cdot)$.

Tra le trasformazioni lineari, la più usata è senz'altro la

$$Y_i' = \frac{Y_i - \mathrm{E}[Y_i]}{\sqrt{\mathrm{Var}[Y_i]}} \qquad 1 \leq i \leq p, \tag{3.108}$$

spesso denominata *normalizzazione*. Naturalmente, nella pratica si utilizzano le stime campionarie di queste quantità. Si verifica facilmente che, le variabili così trasformate, hanno media (campionaria) nulla e varianza (campionaria) unitaria. Si noti infine che le nuove variabili sono adimensionali.

Esistono forme alternative di normalizzazione. Ad esempio, nel caso in cui i valori delle variabili siano non negative, si può far uso della

$$Y_i' = \frac{Y_i}{\max\limits_{1 \leq j \leq N} \mathbf{e}_i' \mathbf{d}_j} \qquad 1 \leq i \leq p, \tag{3.109}$$

dove si è indicato con $\mathbf{e}_i$ l'i-esimo versore, e quindi il prodotto scalare $\mathbf{e}_i' \mathbf{d}_j$ rappresenta la componente i-esima dell'osservazione j-esima (si ricordi che $\mathbf{d}_j$ è stato definito come un vettore colonna). In sostanza, si dividono i dati rilevati di ciascuna variabile per il valore massimo, in modo che i tutti i valori delle nuove variabili siano comprese nell'intervallo unitario. Affinché quest'ultimo sia il *più piccolo* intervallo contenente tutti i nuovi valori, si può ricorrere alla

$$Y_i' = \frac{\left(Y_i - \min\limits_{1 \leq j \leq N} \mathbf{e}_i' \mathbf{d}_j \right)}{\left(\max\limits_{1 \leq j \leq N} \mathbf{e}_i' \mathbf{d}_j - \min\limits_{1 \leq j \leq N} \mathbf{e}_i' \mathbf{d}_j \right)} \qquad 1 \leq i \leq p. \tag{3.110}$$

Se si vuole applicare una tecnica di tipo R ai dati, le tecniche di normalizzazione vanno applicate alle osservazioni, anziché alle variabili. In pratica, vengono utilizzate le stesse formule, previa trasposizione della matrice dei dati $\mathbf{D}$.

Se il metodo adottato prevede sia una analisi Q che una analisi R, è opportuno decidere se normalizzare rispetto alle variabili o alle osservazioni, in quanto effettuare entrambe le operazioni in cascata fornisce risultati poco

interpretabili, dipendenti peraltro dall'ordine in cui le due normalizzazioni vengono applicate.

Nei paragrafi a seguire si assumerà di lavorare con una matrice delle osservazioni con variabili normalizzate.

3.3.2 Metodi di ripartizione

L'obiettivo di questa classe di algoritmi è la ripartizione dei dati disponibili in n sottoinsiemi (*cluster*) $C_1, \ldots, C_n$, quindi tali per cui

$$C_1 \cup \ldots \cup C_n = \{\mathbf{d}_i | 1 \leq i \leq N\} \tag{3.111}$$

$$C_j \cap C_k = \emptyset \qquad j \neq k, \tag{3.112}$$

in modo che gli elementi di ogni sottoinsieme siano "il più compatti possibile." È l'interpretazione e la formalizzazione di questa proprietà alquanto sfumata che caratterizza i singoli algoritmi. Alcuni di essi procedono euristicamente, mentre altri cercano di ottimizzare una determinata funzione obiettivo.

In questa sezione verrà analizzato l'algoritmo kmeans, di gran lunga il più noto ed utilizzato. Esso utilizza come funzione obiettivo da minimizzare la somma dei quadrati delle distanze tra i punti e la media campionaria del *cluster* a cui appartengono. In formula

$$s_{\mathrm{W}} \stackrel{\mathrm{def}}{=} \sum_{j=1}^{n} \sum_{i=1}^{N_j} (\mathbf{d}_{ji} - \mathbf{m}_j)'(\mathbf{d}_{ji} - \mathbf{m}_j). \tag{3.113}$$

Come suggerisce il simbolo adottato per questa cifra di demerito, esiste un parallelo scalare, a meno del fattore $\frac{1}{N}$, dell'equazione MANOVA introdotta nella (3.29). È infatti

$$\begin{aligned} s_{\mathrm{Tot}} &\stackrel{\mathrm{def}}{=} \mathrm{tr}(\mathbf{S}_{\mathrm{Tot\,P}}) \\ &= \sum_{i=1}^{N} (\mathbf{d}_i - \mathbf{m}_{\mathrm{Tot}})'(\mathbf{d}_i - \mathbf{m}_{\mathrm{Tot}}) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{N_j} (\mathbf{d}_{ji} - \mathbf{m}_j)'(\mathbf{d}_{ji} - \mathbf{m}_j) \\ &\quad + \sum_{j=1}^{n} N_j (\mathbf{m}_j - \mathbf{m}_{\mathrm{Tot}})'(\mathbf{m}_j - \mathbf{m}_{\mathrm{Tot}}) \\ &= \mathrm{tr}(\mathbf{S}_{\mathrm{W}}) + \mathrm{tr}(\mathbf{S}_{\mathrm{B}}) \stackrel{\mathrm{def}}{=} s_{\mathrm{W}} + s_{\mathrm{B}}. \end{aligned} \tag{3.114}$$

In questo contesto, però, la s_{Tot} è una costante, mentre la suddivisione in classi è, per così dire, l'incognita. Questo giustifica la scelta della funzione obiettivo, che può essere vista come misura di variabilità intraspecifica, e mostra che è possibile scegliere equivalentemente di rendere massima la misura di separazione tra i gruppi s_{B}.

Prima di presentare l'algoritmo si fa notare che il numero di configurazioni possibili degli n *cluster* sugli N dati si dimostra [98] essere pari a

$$\frac{1}{n!}\sum_{j=1}^{n}(-1)^{n-j}\binom{n}{j}j^N, \tag{3.115}$$

che esplode facilmente per valori non banali dei due parametri[15]. Una ricerca esaustiva della configurazione ottima è quindi improponibile, e ci si accontenta di algoritmi subottimi.

Sia $\mathbf{x}$ il vettore di lunghezza N che conserva i codici associati ai *cluster* di appartenenza di ciascun dato. Il metodo kmeans, partendo da un assegnamento iniziale $\mathbf{x}_0$ e scandendo i dati uno ad uno, ad ogni passo calcola le medie e la funzione obiettivo, e assegna l'osservazione in esame al *cluster* per cui la nuova valutazione della funzione obiettivo è minima. Il procedimento si arresta allorquando $\mathbf{x}$ rimane invariato per N cicli consecutivi.

Questo algoritmo è ottimo ad ogni passo, ma non trova necessariamente la soluzione ottima cercata. È consigliabile pertanto ripetere la procedura con diverse configurazioni iniziali. Si tenga in considerazione, comunque, che la funzione obiettivo s_{W} soffre di alcune limitazioni, e fornisce risultati scadenti per *cluster* non sufficientemente compatti e separati, o aventi cardinalità molto diverse tra loro.

In [98] sono riportate numerose varianti di kmeans, con i listati in linguaggio FORTRAN a corredo. Ad esempio, hmeans ricalcola la funzione obiettivo solo dopo aver completato il ciclo su tutti i dati, e dà la possibilità di ridurre automaticamente il numero di *cluster* durante l'esecuzione. Per funzioni obiettivo alternative, sempre basate sulle matrici di dispersione, si veda [16, sezione 6.8.3].

3.3.3 Metodi gerarchici

In questa sezione verranno esaminati gli algoritmi gerarchici, che più di altri hanno riscosso successo all'interno delle comunità scientifiche di fisici, natura-

[15] Già per $N = 25$ ed $n = 3$ si hanno 141.197.991.025 configurazioni.

listi e sociologi, tanto che alcune pubblicazioni (ad esempio [83]) si riferiscono con il termine *cluster analysis* alla sola analisi gerarchica dei *cluster*.

L'obiettivo di questi algoritmi è l'organizzazione dei dati in una struttura gerarchica, che raggruppa osservazioni molto simili in piccoli *cluster* ai livelli più bassi, e osservazioni più lascamente collegate in *cluster* più grandi e generici ai livelli più alti, fino ad arrivare all'insieme di tutti i dati. Formalmente, si ottiene una sequenza di h partizioni di cardinalità strettamente crescente degli N dati. Sia n_i la cardinalità della i-esima partizione. Sarà allora

$$1 = n_1 < \ldots < n_h \leq N. \tag{3.116}$$

In altre parole, la prima partizione della sequenza è rappresentata da un solo insieme $C_1 = \{\mathbf{d}_i | 1 \leq i \leq N\}$, comprendente tutte le osservazioni; la seconda partizione prevede $n_2 \geq 2$ sottoinsiemi disgiunti e complementari di C_1, e così via, fino all'ultima partizione, che si noti non prevede necessariamente la frammentazione dei dati in N singoletti (*culster* degeneri).

I metodi di analisi gerarchica si distinguono in due categorie: le procedure *divisive*, che, al j-esimo passo, ripartiscono uno o più *cluster* del $(j-1)$-esimo livello in due o più *cluster* di dimensioni inferiori; le procedure *agglomerative*, decisamente le più utilizzate, accorpano al contrario i *cluster* più piccoli (solitamente partendo da quelli degeneri), fino ad arrivare all'insieme di tutti i dati.

3.3.3.1 Metodi agglomerativi binari

In questa sezione viene analizzata l'ampia classe delle procedure agglomerative, che raggruppano ad ogni passo i due cluster più vicini, partendo da quelli degeneri (le singole osservazioni). Con terminologia strettamente locale, si battezzano queste procedure agglomerative *binarie*. In [98] si trovano due esempi di procedure divisive, ed un algoritmo agglomerativo alternativo, qui non illustrato, basato sul concetto di *minimo albero di copertura*.

Vengono ora descritti i passi del generico algoritmo binario, e si discuteranno in seguito le scelte possibili per i suoi parametri.

Dopo aver provveduto alla normalizzazione della matrice dei dati, si costruisce la *matrice di somiglianza* $\mathbf{R}$ (*resemblance matrix*), il cui generico elemento r_{ij} rappresenta il valore di un *coefficiente di somiglianza* calcolato sulle osservazioni $\mathbf{d}_i$ e $\mathbf{d}_j$. La matrice è evidentemente simmetrica. Ci sono due tipi di coefficienti di somiglianza: i coefficienti di *similarità* (*similarity*), tanto più grandi quanto più gli oggetti esaminati sono simili, e i coefficienti di *diversità* (*dissimilarity*), caratterizzati dalla proprietà opposta. Si esaminano gli elementi della matrice, e la coppia di elementi maggiormente affini tra loro viene promossa a *cluster* (proprio). A questo punto la matrice di somiglianza

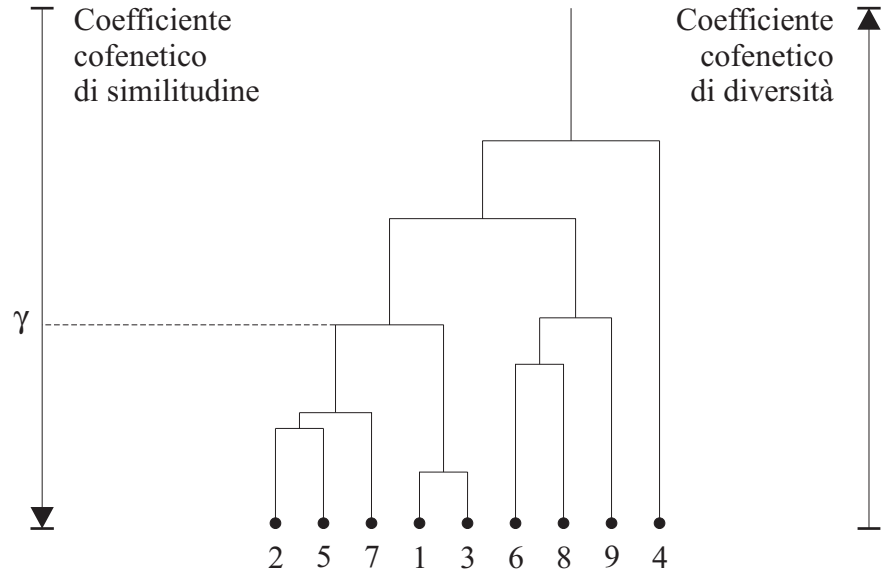


Figura 3.6. Dendrogramma relativo ad un'analisi di nove osservazioni.

deve essere aggiornata. In particolare, le due righe e le due colonne relative alle osservazioni scelte vanno eliminate, e deve essere calcolata una misura di somiglianza tra il *cluster* trovato e tutti gli altri oggetti. Questo implica l'esistenza di una generalizzazione del coefficiente di somiglianza, detto *coefficiente cofenetico*, che misura l'affinità tra coppie di *cluster* qualsiasi, propri o degeneri. Il tipo di generalizzazione adottato è un altro grado di libertà del ricercatore, prende spesso l'ambiguo nome di *metodo di raggruppamento* (*clustering method*) e verrà discusso più avanti. La procedura riprende considerando la nuova matrice fino a che essa si riduce ad uno scalare. Ci si riferirà nel seguito a questa matrice con il nome di *matrice di lavoro* $\mathbf{W}$.

I risultati degli algoritmi gerarchici sono facilmente rappresentabili graficamente attraverso alberi. Nel caso degli algoritmi binari si parla di *dendrogrammi*, o *fenogrammi*, di cui viene dato un esempio nella figura 3.6. Essi forniscono un colpo d'occhio attraente della soluzione trovata e ne favoriscono l'interpretazione. L'altezza a cui i due *cluster* si fondono rappresenta il valore del loro coefficiente cofenetico. Ad esempio, con riferimento alla figura, γ è il coefficiente cofenetico tra i *cluster* $\{\mathbf{d}_2, \mathbf{d}_5, \mathbf{d}_7\}$ e $\{\mathbf{d}_1, \mathbf{d}_3\}$. Si osservi che γ è in generale diverso dal coefficiente di somiglianza originario relativo a una qualsiasi coppia di osservazioni appartenenti rispettivamente al primo e al secondo *cluster*. Sempre riferendosi alla figura 3.6, i valori γ e r_{35}, ad esempio, possono essere differenti. Solo nel caso di *cluster* composti di due

sole osservazioni vale, per costruzione, $c_{ij} = r_{ij}$. Il valore del coefficiente cofenetico viene esteso a ciascuna di queste coppie, simboleggiandolo con c_{ij}. Nel caso in esame è $c_{12} = c_{15} = c_{17} = c_{23} = c_{35} = c_{37} = \gamma$. Si ottiene così, ad algoritmo terminato, la *matrice cofenetica* $\mathbf{C}$, dal punto di vista informativo equivalente al dendrogramma, che, per quanto detto, è generalmente diversa dalla matrice di somiglianza. Questo dà origine ad una distorsione, il cui grado dipende dal coefficiente di somiglianza e dal metodo di raggruppamento adottato. È possibile quantificare questa distorsione confrontando le due matrici, ad esempio attraverso il coefficiente di correlazione di Pearson, introdotto di seguito tra i coefficienti di somiglianza.

Scelta del coefficiente di somiglianza Il più semplice coefficiente di somiglianza è la *distanza euclidea*

$$r_{ij}^{\mathrm{e}} \stackrel{\mathrm{def}}{=} \|\mathbf{d}_i - \mathbf{d}_j\|_2 \stackrel{\mathrm{def}}{=} \sqrt{(\mathbf{d}_i - \mathbf{d}_j)'\,(\mathbf{d}_i - \mathbf{d}_j)}. \tag{3.117}$$

Una sua variante, la *distanza euclidea media*

$$r_{ij}^{\mathrm{d}} \stackrel{\mathrm{def}}{=} \frac{r_{ij}^{\mathrm{e}}}{\sqrt{p}}, \tag{3.118}$$

ha il vantaggio di poter essere usata anche in caso di valori mancanti. Il coefficiente di *differenza di forma* (*shape difference*) è definito da

$$r_{ij}^{\mathrm{z}} \stackrel{\mathrm{def}}{=} \sqrt{\frac{p}{p-1}} \left[d_{ij}^2 - \frac{1}{p^2} \left(\sum_{k=1}^{p} \mathbf{e}_k' \mathbf{d}_i - \sum_{k=1}^{p} \mathbf{e}_k' \mathbf{d}_j \right)^2 \right], \tag{3.119}$$

ed è nullo in caso di osservazioni coincidenti o le cui variabili differiscono tutte per la stessa quantità. Quelli introdotti fin qua sono tutti coefficienti di diversità, con valore minimo pari a zero e illimitati superiormente.

Tra i coefficienti di similitudine figura il *coefficiente del coseno*

$$r_{ij}^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \frac{\mathbf{d}_i' \mathbf{d}_j}{\|\mathbf{d}_i\|_2 \; \|\mathbf{d}_j\|_2}, \tag{3.120}$$

che misura appunto il coseno dell'angolo tra le due linee che collegano le osservazioni con l'origine, e il *coefficiente di correlazione*, introdotto da Pearson

$$r_{ij}^{\mathrm{p}} \stackrel{\mathrm{def}}{=} \frac{\mathbf{d}_i' \mathbf{d}_j - \frac{1}{p}(\mathbf{1}_p' \mathbf{d}_i)(\mathbf{1}_p' \mathbf{d}_j)}{\sqrt{\left[\mathbf{d}_i' \mathbf{d}_i - \frac{1}{p}\left(\mathbf{1}_p' \mathbf{d}_i\right)^2\right] \left[\mathbf{d}_j' \mathbf{d}_j - \frac{1}{p}\left(\mathbf{1}_p' \mathbf{d}_j\right)^2\right]}}, \tag{3.121}$$

dove con $\mathbf{1}_p$ si è indicato il vettore di lunghezza p avente tutti gli elementi unitari. Per questi due coefficienti di similitudine vale $-1 \leq r_{ij} \leq 1$.

In [83], da cui sono stati tratti i precedenti, si trova qualche altro coefficiente, mentre in [98] vengono elencate numerose misure di distanza tra vettori di variabili nominali e ordinali.

Scelta del metodo di raggruppamento Dati due *cluster* C_i e C_j, eventualmente degeneri, si pone il problema della misura della distanza tra essi, detta coefficiente cofenetico. Le prime tre possibilità elencate si basano direttamente sui valori del coefficiente di somiglianza tra le coppie di elementi appartenenti ai due diversi *cluster*. Il metodo detto di *legame singolo* (*single linkage*, o *nearest neighbor*) pone

$$c_{ij}^{\mathrm{s}} \stackrel{\mathrm{def}}{=} \begin{cases} \min\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \max\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di similitudine.} \end{cases} \tag{3.122}$$

Per il metodo di *legame completo* (*complete linkage*, o *farthest neighbor*) vale

$$c_{ij}^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \begin{cases} \max\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \min\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di similitudine.} \end{cases} \tag{3.123}$$

Un metodo tra questi due estremi calcola la media aritmetica

$$c_{ij}^{\mathrm{a}} \stackrel{\mathrm{def}}{=} \frac{1}{|C_i| + |C_j| - 1} \sum_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij}, \tag{3.124}$$

e fornisce alberi di compattezza intermedia rispetto a quelli forniti dai primi due.

È possibile esprimere queste distanze in un modo che rende l'algoritmo più efficiente, riferendosi direttamente alle quantità w_{ij} della matrice di lavoro. Si supponga di dover aggiornare questa matrice in seguito all'accorpamento di due *cluster* C_p e C_q in un unico *cluster* $C_i = C_p \cup C_q$. È necessario calcolare i coefficienti cofenetici c_{ij} tra C_i e tutti gli altri *cluster*. Si ha

$$c_{ij}^{\mathrm{s}} \stackrel{\mathrm{def}}{=} \begin{cases} \min\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \max\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di similitudine;} \end{cases} \tag{3.125}$$

$$c_{ij}^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \begin{cases} \max\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \min\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di similitudine;} \end{cases} \tag{3.126}$$

$$c_{ij}^{\mathrm{a}} \stackrel{\mathrm{def}}{=} \frac{|C_p|w_{jp} + |C_q|w_{jq}}{|C_i|}. \tag{3.127}$$

Un metodo ancora più semplice, proposto da Sokal e Sneath, è

$$c_{ij}^{\mathrm{SS}} \stackrel{\mathrm{def}}{=} \frac{1}{2}(w_{jp} + w_{jq}). \tag{3.128}$$

Il metodo di Ward utilizza il seguente coefficiente cofenetico

$$c_{ij}^{\mathrm{W}} \stackrel{\mathrm{def}}{=} \frac{|C_i||C_j|}{|C_i| + |C_j|}\|\mathbf{m}_i - \mathbf{m}_j\|_2^2, \tag{3.129}$$

ove $\mathbf{m}_i$ rappresenta il centroide di C_i. A partire dal coefficiente di somiglianza

$$r_{ij}^{\mathrm{W}} \stackrel{\mathrm{def}}{=} \frac{1}{2}\|\mathbf{d}_i - \mathbf{d}_j\|_2^2, \tag{3.130}$$

è possibile anche in questo caso aggiornare la matrice di lavoro ricorsivamente, ponendo

$$c_{ij}^{\mathrm{W}} \stackrel{\mathrm{def}}{=} \frac{(|C_p| + |C_j|)w_{jp} + (|C_q| + |C_j|)w_{jp} - |C_j|w_{pq}}{|C_i| + |C_j|}. \tag{3.131}$$

Si dimostra che questo metodo rende minima la (3.113) ad ogni passo. Nel caso si utilizzi come coefficiente cofenetico il quadrato della distanza euclidea tra i centroidi, la formula ricorsiva diventa

$$c_{ij}^{\mathrm{GB}} \stackrel{\mathrm{def}}{=} \frac{|C_p|}{|C_i|}w_{jp} + \frac{|C_q|}{|C_i|}w_{jq} - \frac{|C_p||C_q|}{|C_i|^2}w_{pq}. \tag{3.132}$$

Si può pensare di adottare queste ultime formule ricorsive anche con coefficienti di somiglianza diversi da quelli per cui sono stati studiati, con il rischio di ottenere dendrogrammi con *inversioni* (*reversal*), che presentano cioè fusioni a livelli inferiori rispetto a quelli dei *cluster* componenti. Questo fenomeno dà origine a dendrogrammi di difficile interpretazione.

Un altro effetto collaterale dell'analisi dei *cluster* è detto *concatenazione* (*chaining*), e si verifica quando un grosso *cluster* si forma accorpando un'osservazione alla volta, spostando progressivamente il proprio centro di massa lontano dagli elementi che lo hanno originato. Questo fenomeno è più o meno accentuato a seconda del particolare metodo di raggruppamento adottato.

capitolo 4

Architettura del classificatore realizzato

Il sistema di classificazione automatica è stato realizzato prima di avere a disposizione degli algoritmi di estrazione delle caratteristiche dai file audio. In prima battuta questo può sembrare un controsenso, ma si tratta in realtà di una precisa scelta di progetto. I due blocchi sono stati in questo modo disaccoppiati, consentendo che fossero sviluppati parallelamente, eventualmente da persone diverse. La realizzazione del classificatore in mancanza di dati reali ha imposto di astenersi dal formulare ipotesi sugli stessi che non fossero ampiamente giustificati dai risultati delle precedenti ricerche. Nel presente lavoro si assume perciò la sola aderenza dei dati al modello delle *misture normali* (o *misture gaussiane*, *Gaussian Mixture Model*, GMM), ovvero la multinormalità delle distribuzioni delle singole classi (*class-conditional distribution*). L'altra ipotesi, cioè la distribuzione uniforme del *pitch* all'interno dell'estensione in frequenza degli strumenti, è giustificata dal buon senso, e si tratta comunque di un'ipotesi conservativa.

Per questi motivi, l'applicazione deve includere degli strumenti di analisi oggettivi ed efficaci, che consentano di validare le ipotesi, di selezionare le *feature* più significative in determinati contesti, etc., guidando così la ricerca dell'estrazione delle caratteristiche verso quelle meglio discriminanti, e invarianti all'interno delle singole classi.

Prima di entrare nel dettaglio, è necessario precisare meglio, nei limiti delle difficoltà illustrate nel paragrafo 2.1, quale sia l'oggetto della classificazione. Il suono emesso da molti strumenti musicali cambia sensibilmente se li si eccita diversamente, o se cambiano le condizioni di risonanza. Classici

esempi sono gli archi, che possono essere suonati strofinandoli con l'archetto o pizzicandoli, oppure alcuni ottoni suonati con o senza sordina, per non parlare degli innumerevoli registri dell'organo liturgico, o della molteplicità delle timbriche riproducibili con i sintetizzatori elettronici. Di conseguenza, voler classificare gli *strumenti*, invece che le diverse *timbriche* ad essi associati, sembra una inutile complicazione, oltreché essere poco fedele al comportamento dell'orecchio umano. Per di più, un classificatore di questo tipo sarebbe difficilmente utilizzabile nelle applicazioni indicate nel capitolo 1. I banchi di suoni correntemente utilizzati nei sistemi di sintesi *wavetable*, infatti, distinguono tra le diverse timbriche associate allo stesso strumento musicale. Lo standard General MIDI, ad esempio, prevede due "strumenti" diversi per il basso elettrico suonato con le dita e quello suonato col plettro. A partire dal presente capitolo, perciò, i termini "classe," "strumento" e "timbro," se non specificato diversamente, saranno usati intercambiabilmente. Più difficile è stabilire se debbano essere assegnate classi diverse ai diversi registri di esecuzione, o a diverse regioni dell'estensione, considerata anche la particolarità della maggior parte delle implementazioni di sintesi *wavetable* di utilizzare uno stesso campione per l'esecuzione di svariate note. La questione rimane quindi aperta, anche perché sarebbe prematuro dare una risposta in questa prima fase del progetto.

4.1 Requisiti del sistema

In questo paragrafo verranno analizzate più da vicino i requisiti del sistema e l'interazione del classificatore con gli altri moduli.

4.1.1 I dati in ingresso

La sequenza di *training* consiste in una matrice di dimensioni $N_i \times p$ per ognuna delle k classi, dove N_i è il numero di eventi sonori analizzabili per l'i-esima classe, e p è il numero di caratteristiche[1] da essi estratte. Occorre una stima approssimativa di questi parametri, per poter dimensionare correttamente il sistema. Esistono in commercio numerose librerie di registrazioni di note singole suonate da diversi strumenti musicali, spesso con più di una dinamica e, dove applicabile, con diverse modalità di esecuzione. Una di esse,

[1] In questo capitolo, i termini "caratteristica," "variabile" e "*feature*," verranno considerati sinonimi.

in particolare, è lo standard *de facto*, essendo stata utilizzata nella maggior parte delle ricerche, sia psicoacustiche che di classificazione automatica, illustrate nel capitolo 2. Si tratta dei McGill University Master Samples, o MUMS [76], disponibili gratuitamente per ricerche di carattere accademico, in cui sono registrate le scale cromatiche dei principali strumenti della tradizione occidentale. Altre fonti possono essere le librerie disponibili in commercio, quali quelle della Korg. Si tratta, mediamente, di circa 30 campioni per ogni classe e per ogni libreria utilizzata. Supponendo di utilizzare tre o quattro diverse librerie, si stima che il numero di campioni per ogni classe sia pari a $N_i = 50 \div 100$. Il numero di timbriche tra cui discriminare, in un'applicazione realistica, è ugualmente dell'ordine delle centinaia[2] ($k = 100 \div 150$). Dalla sezione 2.5 si evince che il numero di variabili p che è possibile estrarre può arrivare anch'esso attorno al centinaio. Essendo però ottenute combinando caratteristiche "istantanee" con metodi che ricavano dalla loro evoluzione temporale un solo valore reale per ogni evento sonoro (vedi pagina 24), è possibile che alcune di esse siano poco rilevanti, perché prive di significato fisico, oppure abbiano significato solo per determinate classi di strumenti.

Da questa prima analisi, emerge che il numero di osservazioni è piuttosto esiguo. Come è stato detto nel paragrafo 3.2.11, all'aumentare delle caratteristiche considerate deve seguire un arricchimento esponenziale del *dataset* se si vuole mantenere costante il tasso di errore, o, altrimenti detto, fissata la cardinalità della sequenza di addestramento, le prestazioni decrescono esponenzialmente aumentando p. Ne deriva che è necessario adottare una strategia che permetta di risolvere il problema della dimensionalità, che per quanto detto non si può limitare alla selezione delle variabili migliori in assoluto, anche in considerazione dell'elevato numero di strumenti da classificare.

4.1.2 Integrazione nel dominio applicativo

Al di là delle applicazioni realizzabili nel lungo periodo, accennate nella sezione 1.1, il sistema va integrato a breve termine con altri moduli, sviluppati nell'ambito dello stesso progetto, che si occupino della segmentazione dei file audio in eventi sonori e della effettiva estrapolazione delle caratteristiche. Uno scenario verosimile è illustrato nella figura 4.1. Il segmentatore automatico isola all'interno del brano gli eventi sonori di interesse (singole note, eventualmente accordi). Sia t il numero di questi eventi. L'informazione relativa alla posizione e alla durata di ognuno di essi all'interno del file vie-

[2] Lo standard General MIDI prevede 128 strumenti.

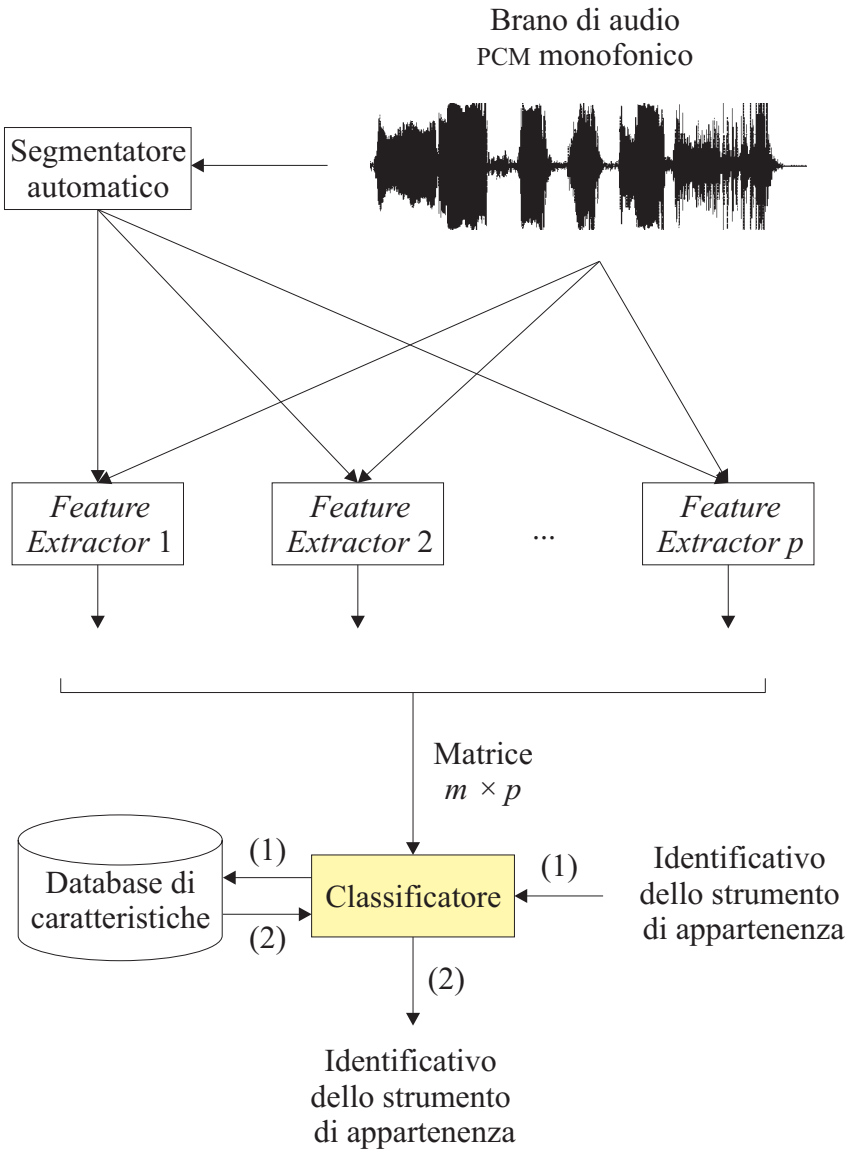


Figura 4.1. Schema a blocchi di una semplice applicazione, realizzabile nell'immediato. La variabile m rappresenta il numero di note individuate dal segmentatore automatico. I flussi contrassegnati da (1) sono relativi alla fase di *training*, quelli contrassegnati da (2) alla fase di classificazione vera e propria, o di test.

ne trasmessa agli algoritmi di estrazione, che, congiuntamente, forniscono in uscita la matrice dei dati relativa al brano. Questa matrice, insieme ad altre eventuali informazioni ausiliarie, passa al classificatore, dal quale può essere utilizzata in due modi distinti. Se il brano è un dato di *training*, dal quale apprendere qualcosa sullo strumento che lo ha suonato, al classificatore deve ovviamente essere trasmesso un identificativo della classe in esame (classificazione assistita). Se invece si tratta di un brano di test, il sistema confronta la matrice con i dati con cui è stato addestrato, e fornisce l'identificativo dello strumento per cui è massima la probabilità che l'abbia suonata (probabilità a posteriori, formula (3.6)).

Il software di segmentazione e analisi sarà probabilmente sviluppato in MATLAB, e si pongono quindi dei vincoli per l'interfaccia tra le due parti del sistema per lo scambio di informazioni.

Per il momento, il *pitch* è considerato alla stregua di una normale caratteristica timbrica. In realtà, questa informazione verrà utilizzata diversamente (paragrafo 4.3).

L'integrazione di un sistema siffatto all'interno di un *encoder* o comunque di un sistema di trascrizione automatica, pone degli stretti vincoli riguardo al tempo di classificazione. Come si è visto nella sezione 1.1, una prestazione in tempo reale può, per certe applicazioni di tipo "*batch*," non essere sufficiente. Sui tempi di addestramento, viceversa, non gravano vincoli particolari.

4.1.3 Interfaccia con l'utente

Il sistema, oltre a fornire un efficiente strumento di classificazione, deve poter essere utilizzato come *testbed* per analizzare l'efficacia delle diverse caratteristiche discriminanti e delle singole tecniche di classificazione per diversi gruppi (*cluster*) di strumenti musicali. Questo presuppone un'elevata flessibilità della sua architettura, la presenza di strumenti di analisi realizzati *ad hoc* e la possibilità di modificare facilmente i dati di *training*, le *feature* utilizzate, e così via. Ad esempio, può essere utile valutare l'invarianza di una determinata caratteristica rispetto alle diverse librerie, o la sua rilevanza all'interno di una determinata famiglia di strumenti.

Al fine di rendere maggiormente leggibili e interpretabili i risultati, si ritiene utile la possibilità di rappresentare graficamente le popolazioni studiate e le relative strutture gerarchiche di classificazione, oltre alla visualizzazione delle statistiche relative ai singoli strumenti. Nasce con questo l'esigenza di proiettare le popolazioni aventi una dimensionalità $p > 3$ in uno spazio tridimensionale.

4.2 Architettura generale del classificatore

Il presente lavoro prende le mosse da quello di Martin [62, 63], preservando e approfondendo la brillante idea della classificazione gerarchica, i cui vantaggi sono illustrati nel paragrafo 4.2.3. Più in dettaglio, viene realizzato quello che nella stessa tesi di dottorato [62, pagina 153] indica come possibile sviluppo futuro: l'automazione della costruzione della gerarchia.

Il classificatore realizzato assume la multinormalità delle distribuzioni delle singole classi, o modello a misture normali. Come detto, questa ipotesi euristica è stata ampiamente validata dalla letteratura.

4.2.1 Fase di *training*

La fase di apprendimento si svolge secondo la seguente modalità. Il *dataset* viene caricato in memoria ed eventualmente modificato. Vengono di seguito calcolate le statistiche necessarie, e le singole variabili vengono normalizzate. Alle classi viene applicato un algoritmo di raggruppamento gerarchico, in modo da pervenire ad un dendrogramma, del tipo mostrato in figura 3.6. L'idea è quella di percorrere l'albero ottenuto, in fase di riconoscimento, partendo dalla radice, effettuando ad ogni nodo una decisione, in questo caso binaria, tra i *cluster* disponibili, scendendo lungo il ramo più aderente ai dati in esame. I dendrogrammi, analizzati nella sezione 3.3.3.1, non si prestano ad essere utilizzati direttamente per una classificazione gerarchica, per due motivi. Innanzitutto, la rappresentazione dei dati in forma di dendrogramma introduce una distorsione, schiacciando di fatto p dimensioni in una sola, il coefficiente cofenetico. Inoltre, l'entità della distorsione e l'organizzazione stessa degli oggetti nell'albero binario dipende fortemente dal coefficiente di somiglianza e dal metodo di raggruppamento adottati. Questo significa che, in una struttura siffatta, una decisione sbagliata ad un livello troppo alto può compromettere irrimediabilmente il risultato finale. Il problema, illustrato nella figura 4.2, può essere affrontato utilizzando diverse tecniche euristiche, elencate di seguito.

Backtracking Si tratta della tecnica senz'altro più conosciuta per problematiche riguardanti alberi di decisione: si percorre la strada più "promettente," fino a che ci si imbatte in classi o *cluster* poco aderenti ai dati analizzati. A questo punto si torna sui propri passi (*backtrack*) e si considerano scelte alternative a quelle che hanno portato fuori strada.

Beam search Questo metodo, adottato nella classificazione gerarchica di Martin, è applicabile sono ad alberi di decisione aventi car-

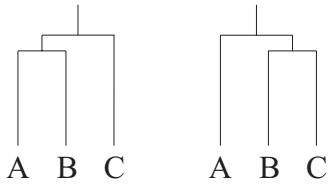


Figura 4.2. Si supponga che due diverse tecniche di raggruppamento originino i due dendrogrammi illustrati. Con le lettere maiuscole si intendono rappresentare dei generici oggetti, siano essi classi o agglomerati di classi (*cluster*). Essendo i coefficienti cofenetici relativi ai due nodi molto simili tra loro in entrambi i casi (le biforcazioni sono relativamente vicine), si può dire che i due dendrogrammi siano praticamente equivalenti. Eppure, dovendo classificare un'osservazione appartenete a B, può accadere che, al primo nodo, solo in uno dei due casi sia effettuata la scelta corretta, rispettivamente i *cluster* (AB) e (BC).

dinalità maggiore di due. Esso prevede infatti di discendere l'albero parallelamente, ad ogni passo, lungo i w rami più "promettenti," dove w è un parametro libero (ampiezza del raggio, *beam width*): generalmente si scelgono valori di w pari a due o tre, mentre per $w = 1$ si ha la classica euristica *greedy*.

Multilayer clustering

Questa tecnica è stata studiata *ad hoc* per questo lavoro, non avendo notizie di approcci simili in letteratura. Si tratta in sostanza di ricavare, a partire dal dendrogramma, una struttura decisionale gerarchica più ricca, in quanto prevede un numero arbitrario di rami uscenti da ogni nodo, sfruttando l'informazione relativa ai coefficienti cofenetici. In questo modo, due biforcazioni sufficientemente vicine come quelle della figura 4.2, vengono collassate in un unico nodo decisionale a tre vie. Questo modello presenta inoltre il vantaggio di avere un maggiore significato fisico. Nel paragrafo 4.5.6 verrà esposta nel dettaglio la realizzazione di questa tecnica.

Al momento della scelta della tecnica, scartata la *beam search*, in quanto non applicabile al dendrogramma, si è optato per un *multilayer clustering*, perché più semplice ed efficiente del *backtracking*.

Per un'applicazione esemplificativa della tecnica sopra esposta di *multilayer clustering* nel semplice caso bidimensionale, si vedano le figure 4.3–4.6.

Figura 4.3. Schematizzazione delle popolazioni di 18 classi multinormali bivariate.

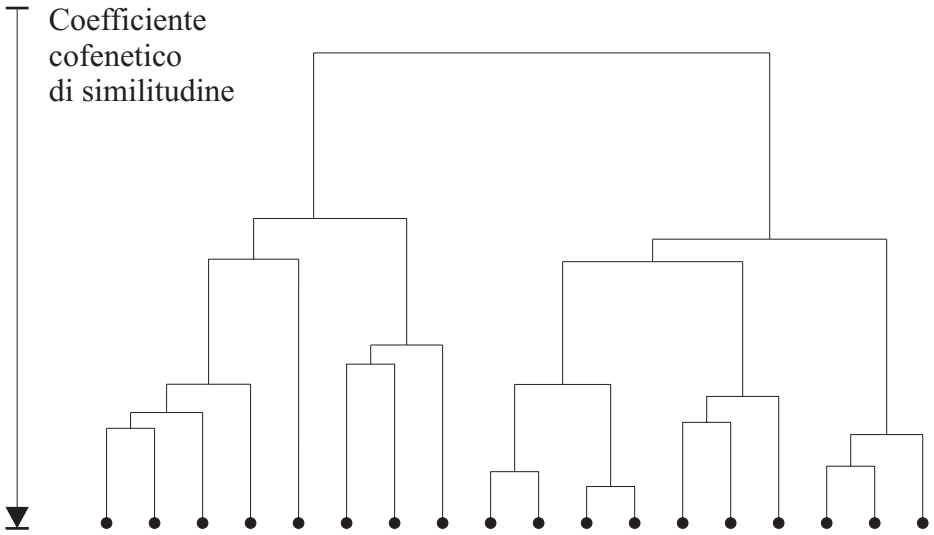


Figura 4.4. Un possibile dendrogramma relativo alle popolazioni della figura 4.3.

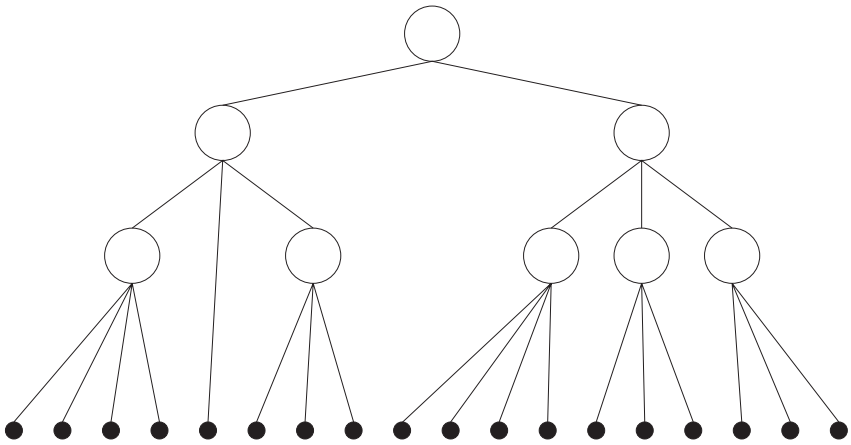


Figura 4.5. Albero decisionale ottenuto attraverso una tecnica di *multilayer clustering* a partire dal dendrogramma della figura 4.4.

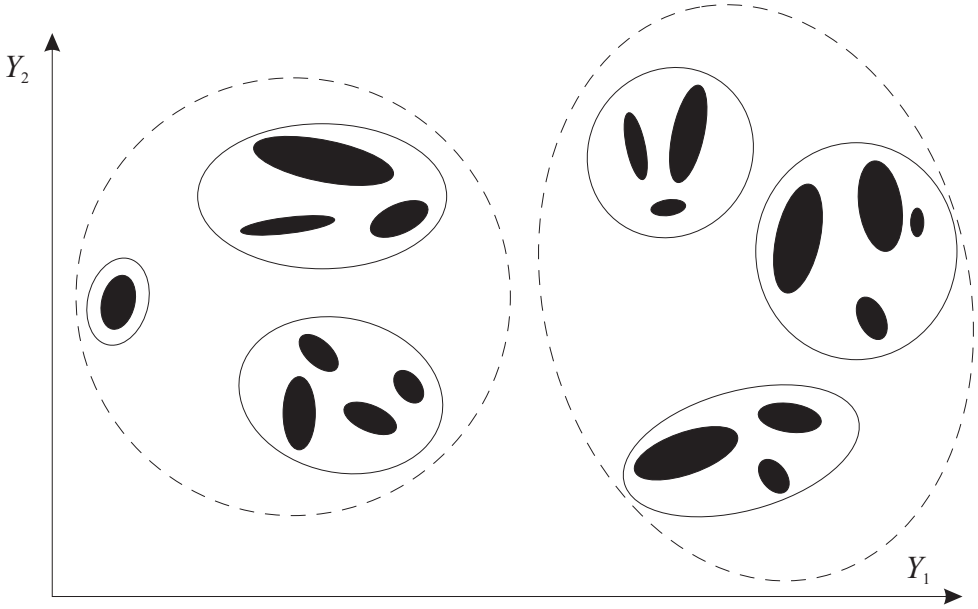


Figura 4.6. Rappresentazione sullo spazio statistico dell'albero decisionale della figura 4.5.

4.2.2 Fase di classificazione

Una volta ottenuta la struttura gerarchica esemplificata nella figura 4.5, si desidera classificare una matrice di dati relativa ad un brano eseguito da uno strumento musicale ignoto. Come è stato accennato nel paragrafo precedente, nel classificatore realizzato questo processo avviene discendendo l'albero a partire dalla radice, percorrendo di volta in volta il ramo relativo al *cluster* che con maggiore probabilità contiene lo strumento incognito. Le modalità di determinazione di queste probabilità possono cambiare da nodo a nodo, avvalendosi di una qualsiasi tecnica di classificazione. Il sistema, cioè, applica ad ogni nodo una tecnica di classificazione rispetto ai *cluster* in esso contenuti, considerando l'insieme delle classi presenti in ogni *cluster* come un'unica super-classe.

La scelta della tecnica adottata ad ogni nodo può essere manuale, oppure può essere effettuata automaticamente dal sistema in fase di addestramento, valutando per esempio quale tra le tecniche note fornisce il migliore tasso di errore, o in base al numero di campioni disponibili, il grado di omoschedasticità delle singole classi, e il loro grado di separazione.

Sebbene le reti neurali, in particolare quelle a mappa auto-organizzante, abbiano per certi aspetti rappresentato un'attraente alternativa nelle prime fasi di progettazione, l'approccio connessionistico è stato accantonato, in quanto non consente un'adeguata analisi introspettiva, ad esempio per quanto riguarda la rilevanza delle caratteristiche. Nondimeno, l'architettura aperta del classificatore descritto prevede l'utilizzo anche di queste tecniche, qualora per alcuni *cluster* forniscano risultati migliori.

4.2.3 Vantaggi di un classificatore gerarchico

Il modello gerarchico presentato gode di numerosi vantaggi, alcuni dei quali già individuati da Martin [62], altri propri delle innovazioni apportate in questo lavoro.

Efficienza Rispetto ad un classificatore tradizionale, le prestazioni aumentano notevolmente, dal momento che gli algoritmi di classificazione vengono invocati con popolazioni molto più snelle. In linea di principio, notoriamente, la complessità passa da $\Theta(f(k))$ a $\Theta(f(\log(k)))$. Ad esempio, per l'algoritmo della discriminante quadratica, era stata calcolata una complessità $\Theta(kp^2)$, che si ridurrebbe a $\Theta(p^2 \log(k))$. Per meglio comprendere l'importanza di questo punto, si ricorda che il numero k di strumenti considerati è dell'ordine delle centinaia.

Flessibilità e scalabilità

Come discusso nel paragrafo 3.2.8 per l'analisi quadratica e quella canonica, alcune tecniche di classificazione sono maggiormente efficaci in determinate condizioni, in dipendenza della morfologia delle popolazioni, o del loro grado di separazione. Per esempio, potrebbe verificarsi che per distinguere le timbriche associate ai diversi sassofoni (ammesso, come ci si aspetta, che cadano tutti nello stesso *cluster*) sia particolarmente efficace la tecnica di discriminazione canonica. Inoltre, il sistema cresce e migliora modularmente man mano che vengono implementate altre tecniche di classificazione.

Sensibilità

Un altro aspetto della flessibilità si ha nella possibilità di una scelta *dinamica* dell'insieme di caratteristiche. Si avrà perciò, similmente al punto precedente, la possibilità di utilizzare, in ogni nodo decisionale, un insieme ristretto di *feature*, quelle che meglio caratterizzano gli oggetti in esame. In questo modo si riesce ad aggirare il più volte citato problema dell'esplosione del numero di osservazioni richieste all'aumentare del numero delle caratteristiche considerate, al fine di mantenere un tasso di errore costante (*curse of dimensionality*). Si può dire, perciò, che, a parità di *feature* utilizzate, il sistema goda di una maggiore *sesnsibilità* (*sesnsitivity*) ai dati della sequenza di addestramento, rispetto ad un metodo di classificazione "piatto." Si noti inoltre che l'abbattimento del numero di variabili p comporti nella maggior parte dei casi un notevole miglioramento delle prestazioni. Nel caso della QDA sopra riportato, ad esempio, la complessità è dominata dal fattore p^2.

Un'altra possibile applicazione di questa libertà di scelta è il confinamento dell'utilizzo delle *feature* più onerose da calcolare, o che si rendono disponibili in un secondo momento (si pensi alle caratteristiche relative al segmento di sostegno, o di rilascio), nei livelli più bassi della gerarchia. Così facendo si pone l'accento sull'efficienza del sistema, specie se realizzato su un'architettura parallela, a scapito forse dell'efficacia.

Emulazione del processo di classificazione umano

Secondo gli studi di Rosch [84], un modello gerarchico è più vicino ai meccanismi di classificazione attuati dal cervello umano. Non ci si aspetta necessariamente che la gerarchia utilizzata coincida con quella della classificazione degli strumenti musicali della tradizione occidentale. Al contrario, si

ritiene che questa ipotesi sia un punto debole della precedente tesi di Martin, come è stato motivato a pagina 29.

Adattatività e robustezza

La possibilità di costruire la gerarchia autonomamente conferisce al sistema una notevole adattatività rispetto all'introduzione di nuove timbriche. Inoltre, le classi più ampie (quelle più in alto nella gerarchia) avranno in comune le caratteristiche percettivamente più rilevanti, anche grazie alla scelta di un *front-end* e di caratteristiche *perception-driven*, e questo conferisce alla struttura una notevole robustezza rispetto agli inserimenti. Si potrebbe perciò pensare, raggiunta una certa massa critica di strumenti, di aggiungere quelli nuovi sulla base di misure di distanza rispetto agli elementi del livello inferiore della gerarchia, senza doverla necessariamente ricostruire.

La robustezza rispetto alla degradazione del segnale, invece, deriva dalla capacità di selezionare oculatamente in ogni contesto le caratteristiche migliori da una rosa molto ricca e articolata.

Il modello presentato non è comunque privo di inconvenienti. Ne sono stati isolati due. L'ipotesi di multinormalità delle classi, se è verificata al livello inferiore, diventa sempre meno verosimile via via che il livello di astrazione sale. Tuttavia, si ritiene che questo comporti una minima degradazione delle prestazioni, a cui in ogni caso si può ovviare adoperando, nei livelli più alti, metodi che non assumano il modello delle misture normali, come quelli presentati nel paragrafo 3.2.9. La costruzione del dendrogramma e della gerarchia vengono effettuati sulla base di *tutte* le variabili, dando la possibilità alla "sciagura delle dimensioni" di corrompere la qualità dei raggruppamenti. Per cautelarsi da questo problema è consigliabile una preventiva selezione delle caratteristiche sulla base del potere discriminante rispetto all'insieme di tutti gli strumenti.

4.3 Sfruttamento dell'informazione relativa al *pitch*

In prima approssimazione, trascurando le possibilità di glissandi e vibrati esagerati, la variabile casuale del *pitch* delle note emesse da un determinato strumento musicale può essere considerata *discreta* e distribuita *uniformemente* per tutta la sua estensione. Per questo motivo, essa si distingue notevolmente dalle altre caratteristiche estratte, che si assumono continue e

normali; impiegarla nello stesso modo significherebbe con ogni probabilità compromettere i risultati.

Nel classificatore realizzato, l'informazione relativa al *pitch* viene utilizzata nel seguente modo: durante la fase di addestramento, vengono registrati i valori di ogni nota, e si stima l'estensione di quello strumento memorizzando il valore minimo e quello massimo. A valle della fase di classificazione, in cui il *pitch* non viene considerato, si controlla se i valori rientrano nell'intervallo previsto per lo strumento identificato. In caso contrario, il sistema avvisa l'utente dell'incongruenza. Con "utente" può intendersi anche un modulo cliente ad un livello di astrazione superiore, come spiegato nel paragrafo 6.2.1). Questo comportamento corrisponde, rozzamente, a quello di un essere umano che, non essendo a conoscenza dell'esistenza del violoncello, riconosca nel suono quello di un violino, pur ammettendo di non averne mai sentito una nota così grave.

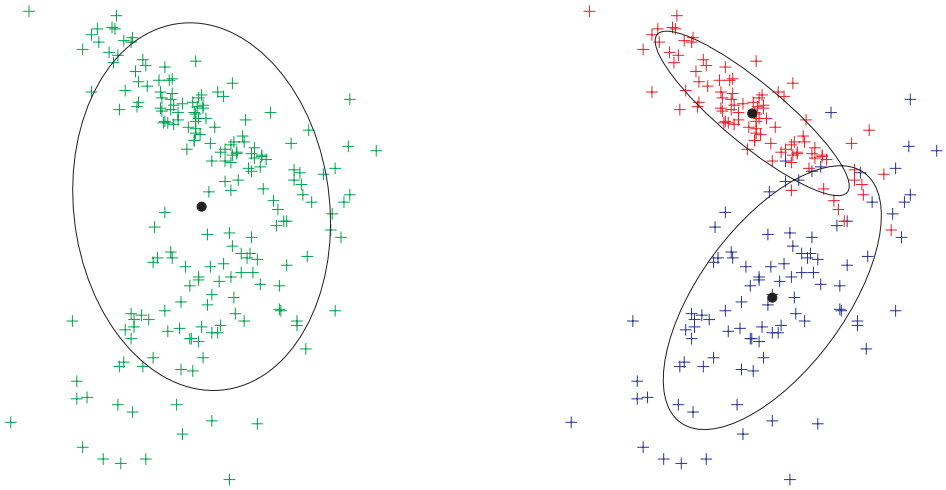
Il pericolo che la dipendenza dal *pitch* delle diverse *feature* possa compromettere i risultati è, nel modello considerato, in gran parte scongiurato dal fatto che le dipendenze lineari vengono automaticamente considerate nelle semplici statistiche di media e matrice di covarianza. Eventuali non-linearità, non necessariamente dettate dalla dipendenza dal *pitch*, possono essere aggirate spezzando lo strumento in esame in due o più classi, che possono ad esempio corrispondere ai diversi registri utilizzati (figura 4.7).

4.4 Strumenti offerti al ricercatore

Il sistema deve fornire al ricercatore un insieme di strumenti che gli consentano di esplorare le caratteristiche delle *feature* utilizzate, delle classi di strumenti studiate, e dei diversi algoritmi di raggruppamento e classificazione adottati. I possibili servizi offerti sono elencati di seguito, cominciando con quelli relativi ai dati delle singole classi. Alcuni di essi si basano sui test esposti nelle sezioni 3.2.10 e 3.2.11, e forniscono una quantificazione delle proprietà ad essi associati calcolandone il *p-value*[3].

Test di definita positività La matrice di covarianza campionaria è indispensabile per effettuare un gran numero di calcoli critici, tra cui quello di classificazione. Si verifica che essa sia definita positiva; in caso contrario, il ricercatore è invitato a fornire un maggior

[3]Il *p-value* di un test è definito come il massimo valore di α per cui il test, applicato ai dati in esame, porta ad una accettazione dell'ipotesi nulla H_0.



(a) Il semplice modello multinormale può non essere adeguato; conviene spezzare lo strumento in più timbriche distinte.

(b) La popolazione viene modellata più fedelmente da una mistura di due distribuzioni multinormali.

Figura 4.7. Due diversi modelli per una popolazione multimodale. Le ellissi rappresentano i luoghi dei punti a distanza standard stimata unitaria dalla media.

numero di osservazioni linearmente indipendenti, che, per la precisione, devono essere in numero almeno pari a $p + 1$.

Misure di asimmetria (*skewness*) e di curtosi

Applicando la (3.91) e la (3.93) ai dati appartenenti ad uno strumento musicale, si ottengono, rispettivamente, una misura della loro asimmetria e della curtosi. Valori elevati di questi parametri sono indice di una morfologia irregolare del gruppo all'interno dello spazio multidimensionale, e possono suggerire di fornire un maggior numero di osservazioni, cambiare alcune *feature*, o di suddividere la classe in due o più timbriche differenti.

Misura di multinormalità

I test 4 e 7 danno la possibilità di quantificare il grado di multinormalità della classe. Misure alternative possono essere calcolate dai test 8 e 9. Valori elevati di queste misure portano allo stesso genere di considerazioni del punto precedente. Per analizzare diverse ragionevoli suddivisioni

della classe in timbriche indipendenti, si possono utilizzare le tecniche di ripartizione trattate nella sezione 3.3.2, test di multimodalità studiati *ad hoc*, o l'algoritmo iterativo di Expectation-Minimization (EM) per misture normali, illustrato in [24, sezione 9.3]. La figura 4.7 evidenzia i vantaggi di un'analisi in questa direzione.

I valori dei seguenti indici servono a valutare le qualità delle variabili considerate.

Determinazione dei valori erratici e tipici

Calcolando la distanza standard, introdotta nel paragrafo 3.2.1, delle singole osservazioni dalla media è facile isolare i valori erratici (*outlier*) e quelli tipici. Risalendo agli eventi che hanno originato quelle osservazioni, il ricercatore può individuare i punti deboli di alcune *feature* o del processo di segmentazione automatica.

Misura di omoschedasticità e di uguaglianza delle medie

Ci si aspetta che le caratteristiche estratte siano invarianti rispetto alle condizioni di registrazione (esecutore, microfonazione, riverberazione, realizzazione dello strumento, rumorosità, intenzione, etc.), ma anche rispetto all'intensità assoluta e alla durata degli eventi sonori. Per verificarlo, si considerino diversi insiemi di osservazioni relative ad uno stesso strumento, registrato in condizioni differenti—ad esempio prendendoli da librerie differenti, o, anche all'interno della stessa libreria, considerando campioni aventi dinamiche differenti. Sotto le ipotesi di invarianza, queste popolazioni soddisfano i test di omoschedasticità (uguaglianza delle matrici di covarianza) e di uguaglianza delle medie esposti nella sezione 3.2.10.

Si noti che, in alcuni dei test per l'uguaglianza della media si assume l'omoschedasticità delle popolazioni, che deve dunque essere verificata per prima. Inoltre, ciascun insieme individuato deve includere un numero sufficiente di osservazioni linearmente indipendenti, in modo che le matrici di covarianza campionarie siano definite positive.

Analisi di ridondanza delle variabli

Attraverso una qualsiasi tecnica illustrata nella sezione 3.2.11, si possono selezionare le *feature* più significative all'interno di un insieme predefinito. Supponendo di utilizzare una tecnica di selezione in avanti, ad esempio, si può pensare di fermare l'analisi dopo aver individuato

un numero fissato di caratteristiche, o quando nessuna tra le variabili non ancora considerate apporta un sufficiente contenuto informativo, sulla base di una soglia prefissata.

Lo stesso tipo di analisi può essere condotto per selezionare le caratteristiche maggiormente invarianti rispetto alle diverse condizioni di registrazione, per lo stesso strumento, alternativamente o congiuntamente all'altro metodo sopra riportato. In questo caso, però, si isolano via via le *feature* che apportano il *minore* contenuto informativo.

4.5 Dettagli tecnici

In questa sezione si illustra come sono state realizzate e impiegate le tecniche esposte nel capitolo precedente, soffermandosi in particolare sulla costruzione del dendrogramma e della gerarchia decisionale.

4.5.1 Calcolo delle probabilità a priori

Le quantità π_j presenti nella (3.9), e in tutte le funzioni discriminanti relative ad approcci statistici, quali la dicriminante quadratica (3.31), o la discriminante canonica (3.67), rappresentano la probabilità che l'osservazione che deve essere classificata faccia parte della classe j-esima, prima che vengano effettuate misurazioni su di essa.

In mancanza di informazioni specifiche, il classificatore è costretto ad assumere l'equiprobabilità di appartenenza, utilizzando la possibilità numero 1 a pagina 36. L'alternativa numero 2 è stata scartata, in quanto non significativa in questo contesto. Il problema è che l'informazione cercata si trova ad un livello di astrazione superiore, e nella sezione 6.2.1 vengono presentate, nel contesto di un'integrazione del classificatore con un sistema esperto, alcune semplici regole che possono portare alla determinazione di questi valori.

4.5.2 Calcolo rapido delle statistiche per gli agglomerati

In più di un'occasione, nella progettazione del sistema, ci si imbatte nel seguente problema: dato un certo numero di classi, di cui siano note la media e la matrice di covarianza campionarie, calcolare la media e la matrice di covarianza campionarie della popolazione risultante dall'accorpamento di queste classi. Questo accade, ad esempio, nei metodi agglomerativi utilizzati per generare il dendrogramma e la gerarchia, ma anche nel caso in cui si voglia implementare un apprendimento incrementale, in cui la sequenza di *training*

di ogni strumento è composta da più sottosequenze, corrispondenti a diverse unità logiche di addestramento—nella fattispecie brani audio custoditi in file PCM, vedi paragrafo 5.2.

Per evitare di prendere nuovamente in considerazione la matrice dei dati, e calcolare a partire da essa le statistiche cercate attraverso le definizioni (3.22) e (3.24), si sfruttano le equazioni MANOVA (3.27–3.29), che richiedono la sola conoscenza della cardinalità, delle medie e delle matrici di covarianza relative ai singoli gruppi, eliminando dalla complessità il fattore relativo al numero totale di osservazioni.

4.5.3 Normalizzazione rapida delle statistiche

Sia $\mathbf{m}$ la media campionaria e $\mathbf{S}$ la matrice di covarianza campionaria relative ad una popolazione. A seguito di una traslazione dei dati di entità $\mathbf{t}$, e di uno stiramento degli assi (*scaling*), di coefficienti $\mathbf{c}$, ad esempio a seguito di una normalizzazione dei dati, si desidera calcolare le nuove statistiche $\hat{\mathbf{m}}$ ed $\widehat{\mathbf{S}}$ relative ai dati trasformati in maniera efficiente, ovvero senza trasformare la matrice dei dati e procedere attraverso le definizioni (3.22) e (3.24).

La media non presenta particolari problemi, essendo

$$\hat{\mathbf{m}} = \mathbf{m} + \mathbf{t}. \tag{4.1}$$

Per la matrice di covarianza le cose si complicano. Si definisce *coefficiente di correlazione* di due variabili casuali $\mathbf{Y}_i$ e $\mathbf{Y}_j$ la quantità

$$\rho_{ij} \stackrel{\text{def}}{=} \frac{\text{Cov}[\mathbf{Y}_i \mathbf{Y}_j]}{\sqrt{\text{Var}[\mathbf{Y}_i]\text{Var}[\mathbf{Y}_j]}} = \rho_{ji}, \tag{4.2}$$

e si definisce *matrice di correlazione* la matrice simmetrica $\mathbf{R} \stackrel{\text{def}}{=} [\rho_{ij}]$. La matrice di correlazione è invariante rispetto alle traslazioni e agli stiramenti degli assi, ed è possibile sfruttare questo risultato per ottenere i valori cercati. Sia

$$\mathbf{D} \stackrel{\text{def}}{=} [\text{diag}\,(\text{diag}(\mathbf{S}))]^{1/2} \tag{4.3}$$

la matrice diagonale[4] delle *deviazioni standard* campionarie (corrispondenti alle radici quadrate delle rispettive varianze, vedi pagina 40). Si dimostra [24,

[4] diag(·) è l'operatore diagonale. Se applicato ad una matrice quadrata, ne estrae la diagonale in un vettore. Se applicato ad un vettore, fornisce la matrice diagonale avente i valori del vettore sulla diagonale.

Per una matrice $\mathbf{A}$ non diagonale, si ha $\text{diag}(\text{diag}(\mathbf{A})) \neq \mathbf{A}$.

esercizio 2.9] che il corrispettivo campionario della matrice di correlazione è dato da

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}, \tag{4.4}$$

e, quindi,

$$\mathbf{S} = \mathbf{D}\mathbf{R}\mathbf{D}. \tag{4.5}$$

È sufficiente perciò calcolare le nuove deviazioni standard, moltiplicandole per i fattori di scalamento

$$\hat{\mathbf{D}} = \mathbf{D} \cdot \mathrm{diag}(\mathbf{c}), \tag{4.6}$$

ottenendo poi

$$\hat{\mathbf{S}} = \hat{\mathbf{D}}\mathbf{R}\hat{\mathbf{D}}. \tag{4.7}$$

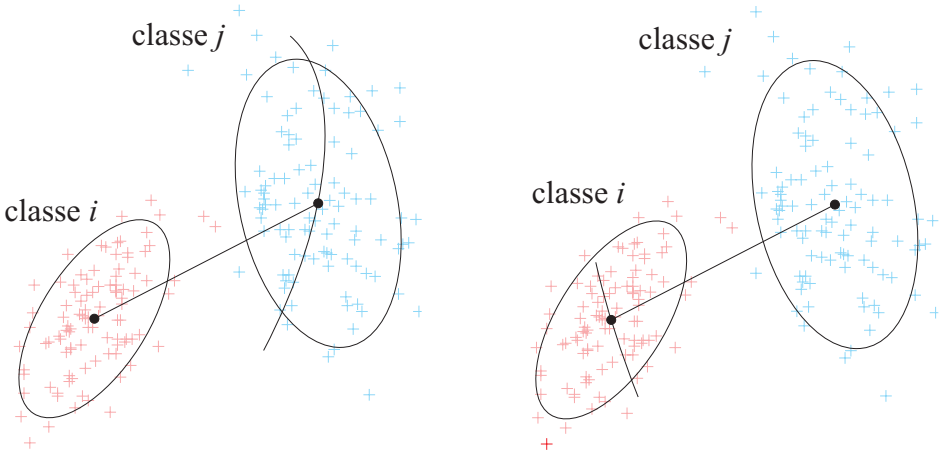
4.5.4 Rappresentazione di popolazioni p-dimensionali

Se la popolazione è immersa in uno spazio p-dimensionale, con $p > 3$, sorge il problema della sua rappresentazione grafica. Occorre proiettare lo spazio in tre dimensioni, in modo che le classi si mostrino più compatte e separate possibili. Fortunatamente, il problema è già stato affrontato nella minimizzazione del criterio (3.54) legato all'analisi discriminante canonica (paragrafo 3.2.6). La soluzione è perciò da ricercarsi nelle prime tre variate canoniche, definite nelle (3.63–3.65).

4.5.5 Costruzione del dendrogramma

L'insieme di tecniche esposte nel paragrafo 3.3.3.1 rappresenta un buon punto di partenza per la costruzione del dendrogramma a partire dai dati relativi alle singole classi. Il problema è che tutte le misure di somiglianza esposte sono applicabili a singole osservazioni, mentre in questo caso gli oggetti elementari che si vogliono organizzare in un dendrogramma, gli strumenti, sono composti da più osservazioni.

Al fine di ricavare la matrice di somiglianza, quindi, è necessaria una misura di distanza o di similitudine tra le singole classi. Una serie di possibilità è offerta dagli stessi metodi di raggruppamento illustrati, che suggeriscono ad esempio di calcolare il minimo, il massimo o la media delle distanze tra tutte le coppie di osservazioni. Oltre che oneroso computazionalmente, questo metodo non tiene in debito conto l'informazione relativa alla morfologia delle classi, in parte custodita nella matrice di covarianza campionaria.



(a) $(\mathbf{m}_i - \mathbf{m}_j)'\mathbf{S}_j^{-1}(\mathbf{m}_i - \mathbf{m}_j)$ (b) $(\mathbf{m}_i - \mathbf{m}_j)'\mathbf{S}_i^{-1}(\mathbf{m}_i - \mathbf{m}_j)$

Figura 4.8. Visualizzazione dei due termini della (4.8). I due segmenti, pur avendo la stessa lunghezza euclidea, rappresentano le due diverse distanze standard tra le due medie campionarie.

Una possibile alternativa è rappresentata dalla misura

$$r_{ij}^{\mathrm{M}} \stackrel{\mathrm{def}}{=} (\mathbf{m}_i - \mathbf{m}_j)'\mathbf{S}_i^{-1}(\mathbf{m}_i - \mathbf{m}_j) + (\mathbf{m}_i - \mathbf{m}_j)'\mathbf{S}_j^{-1}(\mathbf{m}_i - \mathbf{m}_j), \tag{4.8}$$

ovvero la somma delle due distanze standard (paragrafo 3.2.1) tra le medie della generica coppia di classi (figura 4.8).

In [66, sezione 1.12] vengono proposte numerose misure di distanza tra due gruppi generati da funzioni densità di probabilità $f_1(\cdot)$ e $f_2(\cdot)$, fra cui spicca la celebre misura di affinità di Bhattacharyya

$$\rho \stackrel{\mathrm{def}}{=} \int_{\mathbb{R}^p} [f_1(\mathbf{y}) f_2(\mathbf{y})]^{1/2}\, d\mathbf{y}, \tag{4.9}$$

che, per due distribuzioni multinormali, vale

$$-\log \rho = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\log\left[\frac{\det \boldsymbol{\Sigma}}{\sqrt{\det \boldsymbol{\Sigma}_1 \det \boldsymbol{\Sigma}_2}}\right], \tag{4.10}$$

dove

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}. \tag{4.11}$$

Quantificare la distorsione che il dendrogramma opera sulla rappresentazione multidimensionale dei dati può guidare il ricercatore, o il sistema stesso, in maniera automatica, nella scelta del coefficiente di somiglianza o del metodo di raggruppamento. A questo scopo, generalmente, si confrontano la matrice di somiglianza e la matrice cofenetica elemento per elemento, utilizzando una misura di correlazione. Una scelta possibile è rappresentata dalla (3.121).

4.5.6 *Multilayer clustering*

Il metodo studiato appositamente per questa applicazione per passare da un dendrogramma (figura 4.4) a una struttura di decisione gerarchica più articolata e compatta (figura 4.5), è già stato illustrato a grandi linee nei paragrafi 4.2.1 e 4.2.2. In questa sezione si considerano alcune procedure che possono essere applicate per collassare in un unico nodo decisionale le biforcazioni più vicine.

L'idea più semplice è quella di fissare una soglia $0 < \xi < 1$, che viene moltiplicata per il valore più alto del coefficiente cofenetico del dendrogramma in esame, ottenendo una soglia relativa

$$\xi_{\text{r}} \stackrel{\text{def}}{=} \xi \cdot \max_{i,j}\{c_{ij}\}. \tag{4.12}$$

Il dendrogramma viene percorso dalla radice verso il basso, ricorsivamente, e i rami che collegano il nodo attuale a nodi non terminali aventi un coefficiente cofenetico c_{f} tale per cui

$$(c_{\text{a}} - c_{\text{f}}) < \xi_{\text{r}}, \tag{4.13}$$

dove c_{a} rappresenta il coefficiente cofenetico del nodo attuale, vengono eliminati dalla gerarchia, e i due nodi vengono collassati. Nella figura 4.9 l'algoritmo viene applicato ad un dendrogramma esemplificativo.

Possibili generalizzazioni di questo metodo prevedono una somma pesata dei valori cofenetici dei nodi appartenenti non solo al livello immediatamente successivo, ma anche ad un certo numero prefissato (se applicabile) di livelli sottostanti. Similmente, si può pensare di includere nei calcoli anche i valori dei coefficienti cofenetici dei nodi "parenti," o rendere la soglia adattiva. Il pacchetto statistico[5] di MATLAB suggerisce un indice, detto di *inconsistenza*

[5] *Statistics Toolbox*, a partire dalla versione 2.2.

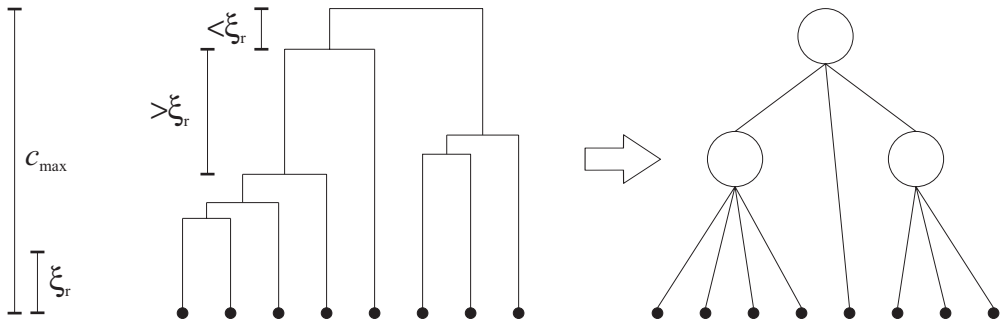


Figura 4.9. Applicazione esemplificativa dell'algoritmo di *multilayer clustering* utilizzando una soglia assoluta $\xi = 0,2$. I nodi figli non terminali sufficientemente vicini al nodo considerato vengono accorpati.

$$I \stackrel{\text{def}}{=} \frac{(c_a - \bar{c})}{\sigma_c}, \tag{4.14}$$

dove $\bar{c}$ rappresenta la media, e σ_c la deviazione standard dei valori del coefficiente cofenetico nei nodi sottostanti all'attuale, fino ad una profondità prefissata.

4.5.7 Decisione relativa ad un brano

Un brano monotimbrico si compone, come detto, di diversi eventi sonori. Volendo classificare l'intero brano, è necessario stabilire il criterio con cui i risultati relativi ai singoli eventi vengono riepilogati in un'unica decisione finale. In assenza di altre informazioni, si è obbligati ad effettuare banalmente una media dei punteggi. Un notevole miglioramento consiste nell'effettuare una media pesata, dove i coefficienti vengono forniti dall'algoritmo di segmentazione, a rappresentare l'affidabilità e la significatività degli eventi individuati. Altri coefficienti di affidabilità possono provenire dai blocchi di estrazione delle caratteristiche della figura 4.1, o, ad un livello di astrazione più alto, da un sistema esperto che sovraintenda l'intero processo (sezione 6.2.1).

capitolo 5

Progetto e realizzazione dell'applicazione

A corredo del presente lavoro è stata sviluppata un'applicazione di classificazione automatica orientata al riconoscimento di timbriche musicali, le cui caratteristiche desiderate e tecniche statistiche adottate sono già state trattate nei precedenti capitoli 3 e 4. In questo capitolo si desidera mettere in luce le varie fasi della progettazione e della realizzazione del software sviluppato, illustrando le tecniche adottate. È stato adottato il modello di ciclo di vita a cascata, riservandosi naturalmente il diritto di effettuare piccole modifiche *in itinere* ai documenti relativi a fasi già concluse, nel momento in cui fossero emersi nuovi risvolti, grazie ad una visione più ampia del problema, piuttosto che per semplici correzioni.

Come nota cautelativa, si ritiene opportuno precisare che l'applicazione fornisce funzionalità più ristrette rispetto a quelle potenziali illustrate nei capitoli precedenti. Ad esempio, sono stati abbandonati alcuni aspetti dell'automatizzazione, come il calcolo del tasso d'errore secondo le formule riportate nel paragrafo 3.2.7. Grazie alla facilità di modifica del *dataset*, comunque, è possibile calcolare agevolmente delle stime soddisfacenti del tasso di errore, effettuando delle validazioni incrociate manualmente. Come conseguenza, il sistema non è attualmente in grado di scegliere automaticamente la tecnica di classificazione migliore ad ogni nodo decisionale, che rimane come importante, ma immediata estensione futura.

5.1 Tecniche e metodi utilizzati

Nelle varie fasi del progetto sono state impiegate diverse tecniche, alcuni formalismi di specifica e strumenti CASE.

Durante le fasi di specifica dei requisiti e di progetto, ci si è avvalsi dello *Unified Modeling Language* (UML [22, 82]). In particolare, sono stati impiegati *state diagram*, *class diagram*, *sequence diagram*, e *component diagram*. Per la realizzazione dei diagrammi è stato utilizzato il pacchetto Rational Rose 98.

La fase di sviluppo è stata preceduta da una importante scelta di progetto relativa ai linguaggi di programmazione. Come sarà descritto più in dettaglio nel paragrafo 5.4.1, si è optato per uno sviluppo in C++ e MATLAB, coordinato dalle C++ Math Libraries e dal MATLAB Compiler. Attraverso questa tecnica, è stato possibile effettuare un *fast prototyping* delle funzioni di calcolo, e riusare alcuni algoritmi MATLAB disponibili in Internet.

La grammatica dei formati proprietari dei file utilizzati dall'applicazione è stata specificata formalmente grazie a due BNF [11].

Il codice e la documentazione sono stati scritti grazie all'*editor* di testo `emacs`, che prevede una vasta gamma di moduli aggiuntivi per l'evidenziazione della sintassi dei diversi linguaggi, per la gestione integrata delle configurazioni, e per l'invocazione dei comandi di compilazione e di *making*. Per la gestione delle configurazioni è stato adottato il popolare pacchetto `rcs`. La documentazione è stata redatta grazie al sistema di *typesetting* gratuito TEX e le macro offerte da LATEX 2_ε [53, 34].

5.2 Specifica dei requisiti

L'analisi del dominio applicativo e dei requisiti dell'applicazione sono distribuiti nei precedenti capitoli, e non saranno ripetuti in questa sede. Può tuttavia essere utile approfondire le caratteristiche dell'interfaccia utente che permetteranno al ricercatore di sperimentare diverse sequenze di addestramento, ad esempio per studiare la rilevanza delle caratteristiche disponibili in diversi contesti, o classificare diversi brani in blocco (*batch*), ad esempio per valutare le prestazioni del classificatore.

L'oggetto atomico che costituisce una osservazione è l'evento sonoro, che si assume per semplicità essere una singola nota (*tone*). Le note, però, non si prestano ad essere acquisite singolarmente; sono invece disponibili in grandi quantità brani (*excerpts*) di esecuzione monofonica dei diversi strumenti, direttamente in file audio PCM (vedi paragrafo 4.1.1). Nella sequenza di *training*, ogni timbrica musicale è rappresentata da un certo numero di brani,

che contengono a loro volta un numero arbitrario di note. Le note sono rappresentate dal *pitch*, dalla durata e dall'insieme dei valori delle caratteristiche estratte, che si assume essere lo stesso per tutto il *dataset*; esse sono identificate univocamente dalla sessione di registrazione, dal nome del file PCM da cui provengono e dalla loro posizione all'interno di esso. L'identificazione delle note è importante, poiché memorizzare l'audio non compresso insieme ai vettori delle variabili che lo rappresentano è improponibile: farebbere infatti crescere lo spazio occupato dalla sequenza di *training* di almeno un ordine di grandezza. Nondimeno, è desiderabile risalire all'audio delle singole note, in modo che il ricercatore possa isolare gli eventi spuri o atipici, studiando i punti deboli del modulo di segmentazione, elidendoli manualmente se necessario dalla sequenza.

Per poter valutare la rilevanza delle diverse *feature*, come detto, l'utente deve poter facilmente escludere dalla base di conoscenza, temporaneamente o definitivamente, determinati strumenti, oppure tutti i brani appartenenti ad una o più sessioni di registrazione. Similmente, devono essere previste funzioni di inserimento di nuovi strumenti e brani, ad esempio per poter valutare la robustezza del sistema al variare dei dati disponibili.

Le modifiche al *dataset* comportano l'alterazione dei coefficienti di normalizzazione e della struttura gerarchica ottima. Ricalcolare il tutto ad ogni modifica è un eccesso, e quindi si preferisce separare il processo di apprendimento in tre fasi distinte

- Modifica della sequenza di addestramento
- Normalizzazione e calcolo delle statistiche
- Calcolo dell'albero decisionale

Per meglio illustrare questi stadi, ci si avvale di uno *state diagram* di UML, rappresentato nella figura 5.1. Inizialmente il *dataset* viene caricato in memoria da un file. Successivamente, nello stato di Dirty, possono essere effettuate tutte le modifiche del caso, senza che debbano essere calcolate le statistiche ogni volta. Quando l'utente è soddisfatto della sequenza ottenuta, la manda ad effetto attraverso il commit. Se uno o più strumenti non dispongono di un numero sufficiente di osservazioni linearmente indipendenti, e quindi la loro matrice di covarianza non è definita positiva, l'utente viene invitato a fornire maggiori informazioni, rimanendo nello stato Dirty. In caso contrario, i dati vengono normalizzati, e vengono calcolate le statistiche e le misure associate ai test. Nello stato di Normalized si possono già effettuare delle classificazioni "piatte," ovvero non gerarchiche. Si possono inoltre scegliere i parametri e gli algoritmi usati nella successiva fase di costruzione della struttura gerarchica,

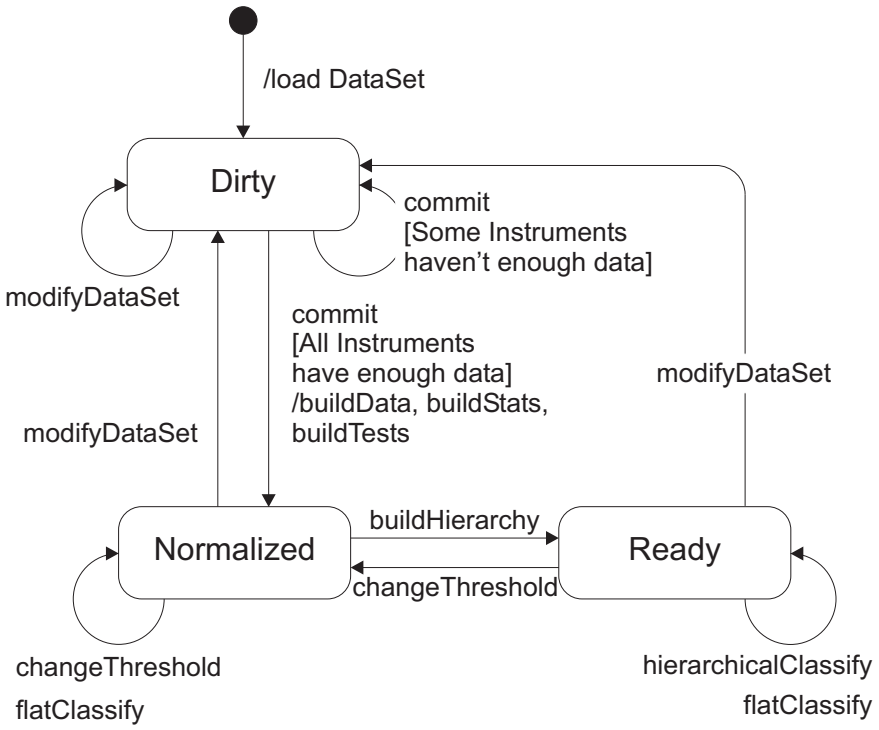


Figura 5.1. Diagramma degli stati del sistema relativo alla fase di acquisizione e modifica della sequenza di addestramento.

senza con questo inficiare la normalizzazione. Si può ad esempio modificare il valore della soglia assoluta ξ relativa al *multilayer clustering*, discusso nel paragrafo 4.5.6. Calcolato l'albero decisionale, il sistema è pronto (Ready) per effettuare classificazioni gerarchiche e tradizionali, fino a che non venga modificata la sequenza di addestramento o le opzioni relative al *clustering*.

Un'altra caratteristica importante che l'applicazione deve presentare è un elevato grado di interoperabilità, in quanto i sistemi che dovranno interagire con essa (vedi figura 4.1) saranno sviluppate separatamente, forse con altri linguaggi e su altre piattaforme. Il formato dei file di interfaccia, contenenti i valori delle variabili estratte, deve perciò essere il più semplice e universale possibile, come il formato testo ASCII.

5.3 Specifica di progetto

Supponendo di realizzare il sistema con un linguaggio orientato agli oggetti, è stata stesa una prima versione di *class diagram*, che ha indirizzato le suc-

cessive fasi del progetto. Il diagramma è poi passato attraverso raffinamenti successivi, guidando i necessari stadi di *re-engineering*. Quella illustrata nella figura 5.2 è la versione più aggiornata. In essa non sono rappresentate le funzioni membro di interfaccia sui dati privati, i costruttori e i distruttori standard, e i dettagli implementativi. Segue una breve descrizione relativa a ciascuna classe.

`Pitch` e `PCMLength` Sono stati evidenziati attraverso una definizione globale i tipi che rappresentano il *pitch*, le posizioni e le durate all'interno dei file PCM. Le unità, e di conseguenza i tipi adottati per misurare e rappresentare queste quantità, sono infatti suscettibili di variazioni, non essendo ancora noto il *front-end* che sarà utilizzato dagli altri moduli.

`FeatureSet` L'insieme delle caratteristiche significative è comune a tutte le osservazioni (di addestramento o di test) memorizzate nel sistema. Questa classe associa biunivocamente un indice numerico, adottato nelle rappresentazioni matriciali dei dati, ad una descrizione testuale della variabile relativa. Un oggetto di questo tipo è presente in ciascun oggetto `Classifier`, `IntrumentPool`, `AbstractInstrument` ed `Excerpt`. Per evitare di duplicare inutilmente l'informazione, in queste classi sono previsti dei *puntatori* ad un oggetto `FeatureSet` unico per ogni istanza di `Classifier`.

`Tone` Come anticipato nella sezione precedente, una nota custodisce l'informazione relativa al *pitch*, il vettore dei valori delle singole variabili[1], posizione e durata all'interno del brano da cui è stata estratta. Il *flag* `inDataSet`, comune anche alle classi `Excerpt` e `Instrument`, viene abbassato se si intende escludere temporaneamente un oggetto dal *dataset*.

`Excerpt` I brani contengono un insieme di `Tone`, e sono caratterizzati da due stringhe, che si riferiscono alla sessione di registrazione e al nome del file PCM. Il costruttore indicato legge un file di testo con estensione `.exc`, il cui formato verrà illustrato nel paragrafo 5.4.5.

[1] `mwArray` è una classe fornita dalla libreria MATLAB di cui si parlerà più approfonditamente nella sezione 5.4. Per ora basti sapere che con essa vengono rappresentati indistintamente vettori e matrici di numeri in virgola mobile.

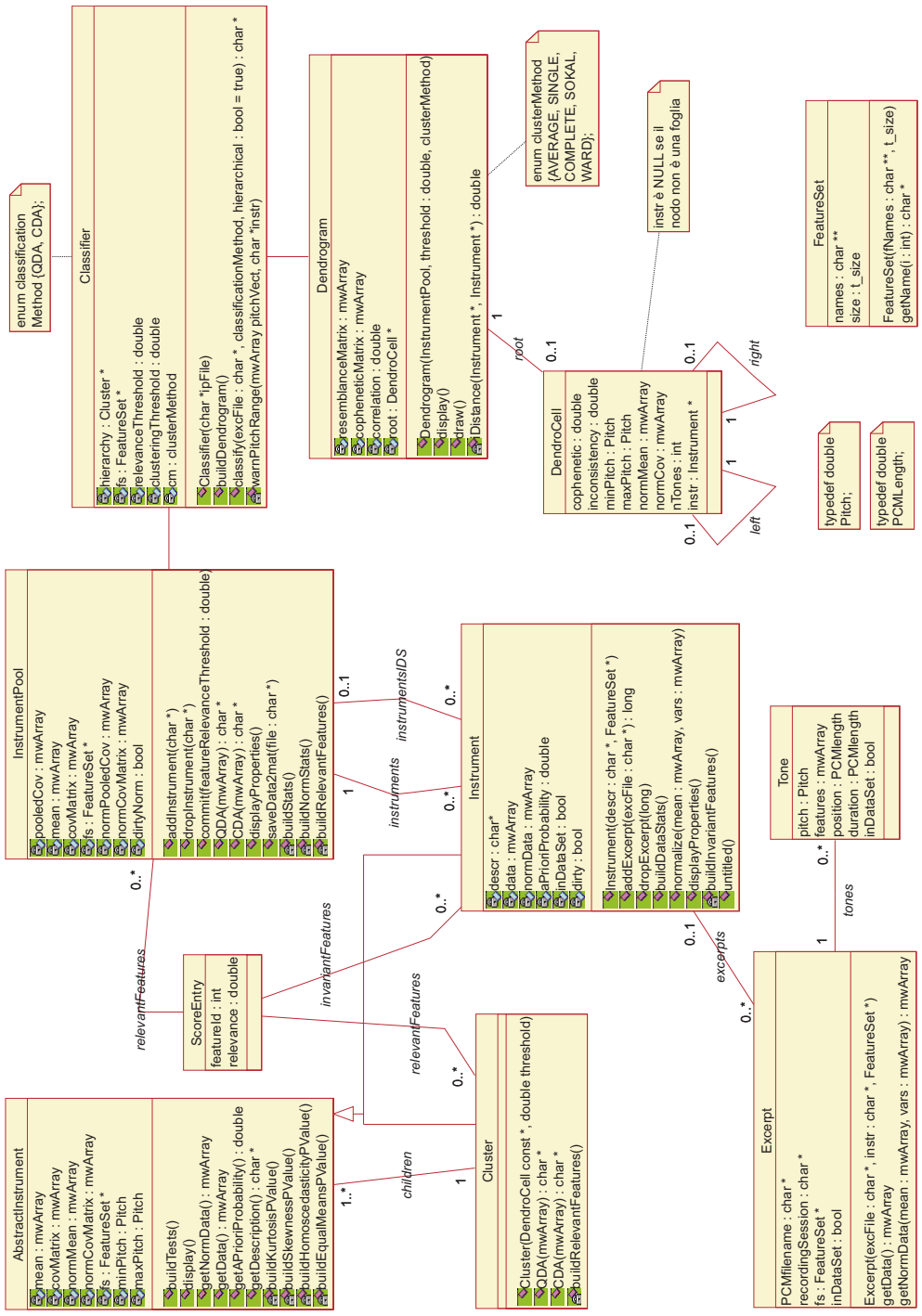


Figura 5.2. *Class diagram* dell'applicazione realizzata. Dove non specificato, i dati si intendono privati (icona con il lucchetto), e le funzioni membro pubbliche.

Le funzioni `getData()` e `getNormData()` restituiscono le matrici, normalizzate e non, dei dati relativi alle note contenute nel brano. Questo fa emergere una problema che si è dovuto affrontare fin dalle prime fasi del progetto: la gestione delle *cache* per i dati aggregati. Si propone infatti un *trade-off* tra la rapidità di accesso ai dati, specie quelli normalizzati, e la duplicazione delle informazioni. Il dilemma è stato risolto valutando i numeri in gioco, mantenendo una sola *cache* dei dati a livello di `Instrument`.

`Abstract Instrument` Come suggerisce il nome, si tratta di una classe astratta, in quanto i metodi sono tutti virtuali puri, da cui derivano le classi `Instrument` e `Cluster` (quest'ultima rappresenta un nodo decisionale nella gerarchia finale). I dati comuni a queste due strutture sono il vettore media, la matrice di covarianza e i valori del massimo e del minimo *pitch* registrati, oltre ai diversi *p-value* relativi ai test, non mostrati nel diagramma.

`Instrument` Identificato da una stringa, ogni strumento conserva le due matrici (normalizzata e non) di tutte le osservazioni ad esso relative che fanno parte del *dataset* e il *flag* `dirty` notifica la validità di queste matrici. La funzione `buildDataStats()` aggiorna le matrici e calcola le statistiche relative. Il metodo `buildInvariantFeatures()` aggiorna la lista delle *feature* che presentano la maggiore invarianza intra-classe. I punteggi associati a ciascuna variabile sono raccolti in semplici strutture dati, `ScoreEntry`.

Per il calcolo e la memorizzazione delle medie e delle matrici di covarianza si è riproposto il problema della politica di *caching*, in quanto le formule MANOVA esposte a pagina 42 consentono un notevole risparmio computazionale per grosse moli di dati. La soluzione adottata ricalca quella relativa alle matrici dei dati: si conservano le statistiche solo negli oggetti di tipo `Instrument`.

`Cluster` Ogni oggetto di questo tipo rappresenta un nodo decisionale nella gerarchia illustrata nella figura 4.5. Essi conservano i puntatori agli oggetti del livello immediatamente inferiore, che possono essere a loro volta di tipo `Cluster`, o di tipo `Instrument`, ovvero nodi foglia. Generalizzando, `Cluster` contiene un vettore di puntatori ad oggetti di tipo `AbstractInstrument`. Tra le

funzioni membro spiccano quelle di classificazione (`QDA` e `CDA`, le uniche implementate attualmente) ed il costruttore, che riceve in ingresso un dendrogramma (o una sua parte) e una soglia assoluta.

`InstrumentPool` Questa classe raccoglie l'insieme di tutti gli strumenti presenti nel sistema e le relative proprietà. Anche qui si trovano le funzioni `QDA` e `CDA`, che effettuano la classificazione "piatta" sull'intero insieme di strumenti. La lista `instrumentsIDS` contempla solo gli strumenti che sono effettivamente nel *dataset*. Questa struttura si rivela comoda per il processo di costruzione della gerarchia, che altrimenti avrebbe dovuto controllare ogni volta il *flag* `inDataSet` di `Instrument`. Il prezzo da pagare è un costante allineamento tra `instruments` e `instrumentsIDS` durante la modifica della sequenza di *training*.

`Classifier` Si tratta della classe centrale dell'applicazione. Essa contiene l'`InstrumentPool`, il `Dendrogram` e la struttura gerarchica (`hierarchy`), nonché tutte le opzioni e i parametri configurabili dall'utente, come il metodo di raggruppamento, la soglia per il *multilayer clustering*, etc. Il costruttore suggerisce che il caricamento del *dataset* avviene attraverso un file di formato proprietario, che verrà illustrato nel paragrafo 5.4.5.

`Dendrogram` Vengono memorizzate in questa classe tutte le strutture dati necessarie alla costruzione (`resemblanceMatrix`) e alla rappresentazione (`copheneticMatrix` e `root`) del dendrogramma. Il fattore di correlazione è un indice della distorsione introdotta dal processo di agglomerazione binario (paragrafo 4.5.5). La funzione `iDistance` calcola la distanza tra gli strumenti, al fine di ottenere la matrice di somiglianza.

`DendroCell` Ogni oggetto di questa classe rappresenta un nodo della rappresentazione ad albero del dendrogramma. Si ricorda che la matrice cofenetica e l'albero sono due strutture dati dall'equivalente contenuto informativo. Il coefficiente di inconsistenza è stato introdotto nel paragrafo 4.5.6.

Le interazioni fra i vari oggetti sono generalmente piuttosto semplici e lineari. L'unica eccezione è rappresentata dal processo di commit, che dà effetto a tutte le modifiche apportate al *dataset*. Esso, infatti, coinvolge i *flag* di `dirty` dei vari oggetti, e deve tenere conto delle precedenze di alcune

operazioni rispetto ad altre: ad esempio i valori dei test vengono calcolati sui dati già normalizzati. Per questo motivo è stato approntato un *sequence diagram* (figura 5.3) che catturasse questo meccanismo, sgombrando il campo da eventuali incomprensioni.

5.4 Dettagli di realizzazione

In questa sezione verranno esaminate più da vicino le tecniche e le scelte implementative adottate.

5.4.1 Scelta dei linguaggi di programmazione

Il progetto si presta ad essere realizzato attraverso un linguaggio di programmazione orientato agli oggetti, e necessita al contempo di una libreria matematica molto ricca, che copra non solo gli operatori matriciali elementari, ma anche funzioni matriciali e statistiche piuttosto avanzate. Fortunatamente, è stata messa a disposizione dell'autore un pacchetto di MATLAB che risolve questi problemi e presenta altri vantaggi, che verranno esposti nel seguito. L'utilizzo di questo pacchetto rende il C++ una scelta obbligata come linguaggio di sviluppo principale.

Le MATLAB C++ Math Libraries sono delle librerie per C++ che arricchiscono il linguaggio con un gran numero di funzioni proprie dell'ambiente MATLAB, e con la classe `mwArray`. Quest'ultima rappresenta il tipo polimorfico utilizzato nel linguaggio interpretato da MATLAB, che può contenere scalari (in doppia precisione), numeri complessi, vettori, matrici, persino stringhe, *struct* e *cell array*. Il suo impiego all'interno di un programma fortemente tipizzato come il C++ deve quindi essere cauto e circostanziato. Attraverso l'utilizzo delle Math Libraries, è possibile combinare la potenza e la flessibilità del C++ con i servizi di alto livello ed estremamente efficienti forniti da MATLAB per la manipolazione delle matrici.

Il MATLAB Compiler rappresenta un ulteriore passo in avanti. Esso consente di tradurre funzioni MATLAB qualsiasi in codice C++, purché si utilizzino le Math Libraries, effettuando se necessario una ricorsione sulle eventuali funzioni ausiliarie. In questo modo è stato possibile scrivere il codice critico direttamente in MATLAB, e, dopo averlo testato, tradurlo e integrarlo nel progetto, realizzando prototipi veloci delle funzioni ad alto contenuto matematico. Questo meccanismo, inoltre, ha permesso di riutilizzare il codice gentilmente fornito dal Prof. Flury per la classificazione a discriminante canonico e quadratico [24].

L'uso delle Math Libraries si è rivelato utile anche sotto altri aspetti.

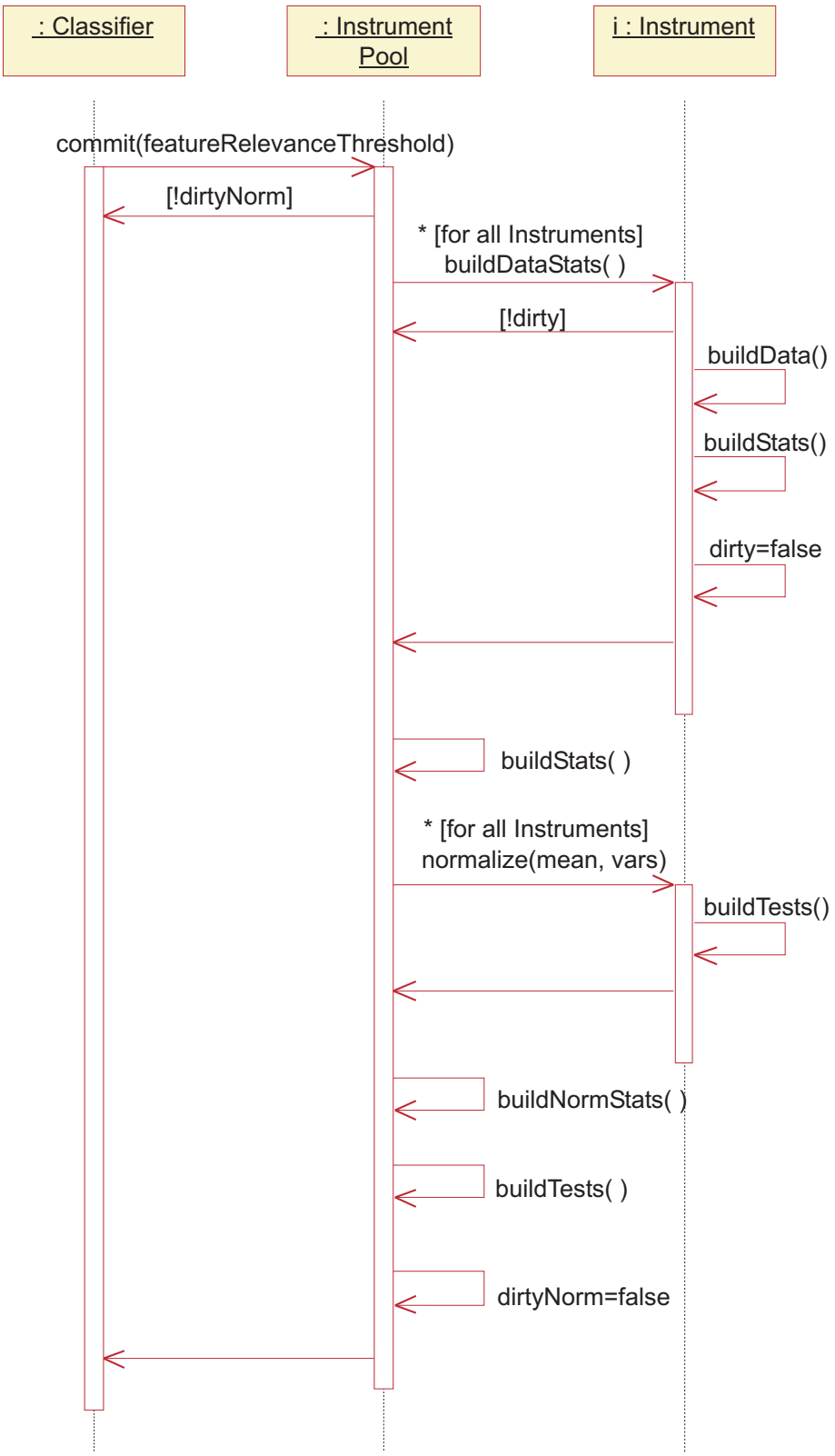


Figura 5.3. *Sequence diagram* relativo al processo di commit.

Salvando i dati relativi alla sequenza di addestramento in file `.mat`, formato proprietario di MATLAB, è stato possibile visualizzarli utilizzando il motore di *rendering* tridimensionale incluso nel pacchetto matematico. Lo stesso dicasi per i dendrogrammi, che necessitano di una visualizzazione in modalità grafica per apprezzarne il colpo d'occhio. È stata comunque fornita una rappresentazione alternativa LISP-*like* in modalità testo per gli utenti che non disponessero del sistema MATLAB sulla propria macchina.

5.4.2 Utilizzo delle librerie standard del C++

Ove possibile, sono state sfruttate le librerie standard fornite con il compilatore. In particolare, si è fatto largo uso delle *Standard Template Libraries* (STL) per rappresentare le collezioni di oggetti evidenziati nel *class diagram* della figura 5.2. Ad esempio, la lista di diramazioni del nodo decisionale `Cluster` è stata realizzata attraverso un `vector<AbstractInstrument *>`, e le associazioni `instruments` e `instrumentsIDS` attraverso delle `map<string, Instrument *>`.

Purtroppo, il compilatore utilizzato (Borland C++ 5.0) ha mostrato parecchi punti deboli. Ad esempio, non gestisce correttamente la classe standard `string`, e quindi si è spesso fatto ricorso al vecchio `char *`. Un altro limite enigmatico di questo tipo è stato rilevato con il tipo `vector<mwArray>`. Per questi motivi il codice può presentare in alcuni punti delle incoerenze, o delle apparentemente assurde perifrasi sintattiche, atte ad aggirare questi problemi.

5.4.3 Modularizzazione, *information hiding* e leggibilità del codice

Il C++ facilita e incoraggia una programmazione basata sui principi dell'*information hiding* e della modularizzazione. Nel caso dell'applicazione in esame, la scomposizione in moduli e quindi in file sorgente del programma rispecchia la specifica di progetto, facendo corrispondere semplicemente un modulo ad ogni classe. In alcuni casi si è avvertita la necessità di spezzare i moduli di implementazione in più sottomoduli, al fine di agevolare il processo di compilazione. La figura 5.4 rappresenta le dipendenze funzionali dei diversi componenti. Si è spesso sentita la necessità di raccogliere a fattor comune alcune funzioni statistiche frequentemente usate, come quelle per il calcolo rapido delle statistiche (paragrafo 4.5.2), o semplicemente utilizzate in due contesti diversi, come `QDA` e `CDA`, o le funzioni che determinano il grado di rilevanza o invarianza delle varie *features*. Esse si trovano nel modulo commonStats. Lo stesso dicasi per il modulo common, in cui sono state raccolte le funzioni e le definizioni globali comuni a quasi tutti gli altri moduli. In

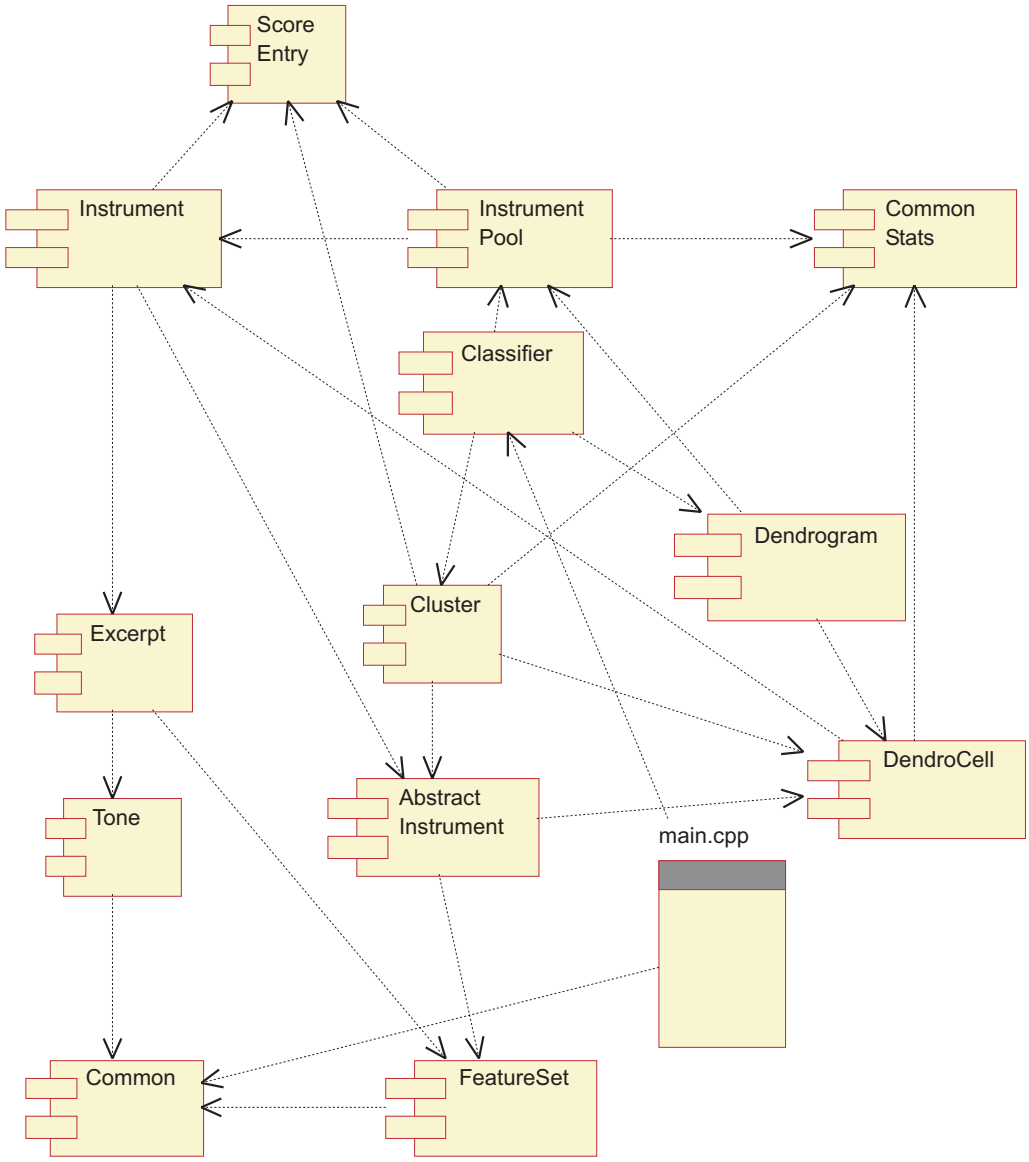


Figura 5.4. *Component diagram* che evidenzia la modularizzazione e le dipendenze funzionali dei moduli del sistema. Nel diagramma non sono evidenziate le dipendenze dalle librerie standard del C++ e dalle MATLAB C++ Math Libraries.

questo modo, e grazie anche al corretto uso dei costrutti del linguaggio, si è minimizzata, se non annullata, la deprecabile attività di "copia/incolla" che rende i progetti non banali poco trasparenti e difficili da mantenere. Altri piccoli accorgimenti, come la definizione di tipi enumerativi simbolici nelle intestazioni dei moduli e la dichiarazione dei loro valori di default, hanno favorito una maggiore leggibilità del codice.

Le funzioni MATLAB tradotte automaticamente in C++ sono state isolate in moduli separati, al fine di poter modificare i sorgenti in MATLAB e sovrascrivere semplicemente le vecchie traduzioni con le nuove. Questi moduli non sono stati riportati nel diagramma per evitare di appesantirlo ulteriormente.

I cicli di interdipendenza (ad esempio tra `Cluster`, `Instrument` e `DendroCell`) sono stati risolti attraverso la tecnica standard della dichiarazione vuota delle classi interessate prima della loro definizione, al fine di spezzare il ciclo di `#include`, che avrebbe altrimenti reso impossibile la compilazione.

Non si ritiene opportuno, in questa sede, scendere più in dettaglio per quanto riguarda le dipendenze funzionali dei vari moduli. In caso di necessità ci si può comunque riferire ai commenti diligentemente riportati all'inizio di ogni modulo e di ogni funzione, in cui si specificano le direzioni e i significati dei flussi di informazione.

5.4.4 Gestione degli errori

Gli errori e le situazioni inaspettate sono stati gestiti attraverso il meccanismo di *exception handling* fornito dal C++. Sono state identificate tre tipologie di eccezioni: quelle dovute a problemi di I/O (file inesistenti, corrotti, o dal formato scorretto), a errori di calcolo (perlopiù divisioni per zero e matrici di covarianza non definite positive), e infine quelle dovute a chiavi di accesso errate per gli *abstract container* o ad indici fuori dagli intervalli consentiti. La gestione delle eccezioni sollevate internamente dalle librerie di MATLAB non viene approfondita in questo paragrafo.

Nella tabella 5.1 sono riportate i vari tipi di eccezioni, cui corrispondono delle semplici classi, i moduli che possono sollevarle, ed il loro significato.

5.4.5 Formati proprietari dei file

L'applicazione comunica con altri moduli, come illustrato nella figura 4.1. Si è già detto che il formato di queste comunicazioni deve essere il più possibile semplice ed indipendente dal linguaggio di programmazione e dalla piattaforma. Il formato di puro testo è quindi da preferirsi, seppure più ingombrante, ai formati binari.

Eccezione	Moduli che la sollevano	Commenti
`TooFewClasses (int)`	commonStats	Le ststistiche per gli agglomerati (paragrafo 4.5.2) vengono calcolate a partire da due o più sottogruppi.
`CantOpenFile (char *)`	Excerpt, Classifier	File inesistente o protetto in scrittura.
`BadFileFormat (char *)`	Excerpt, Classifier	Il formato del file è diverso da quello specificato (paragrafo 5.4.5).
`Unknown Classification Method(char *)`	Classifier	Il metodo di classificazione specificato non è valido.
`TargetTooFar()`	InstrPool, Cluster	Si è verificata una divisione per zero nelle formule (3.35) o (3.68): probabilmente il campione da classificare è troppo lontano.
`DepthTooLow(int)`	DendroCell	La profondità per il calcolo del coefficiente di inconsistenza (4.14) è troppo bassa, e genera una divisione per zero. Si suggerisce di provare con valori più alti.
`MessedIndices()`	Dendrogram	Errore interno irrecuperabile, dovuto ad un'inconsistenza degli indici durante la costruzione del dendrogramma.
`UnknownKey`	Excerpt, Instrument, InstrumentPool	La chiave con cui si desidera accedere ad un elemento di un *abstract container* (`tones`, `excerpts`, `instruments` e `instrumentsIDS`) non è associata ad alcun elemento.
`DuplicateKey`	Excerpt, Instrument, InstrumentPool	La chiave con cui si desidera inserire un nuovo elemento in un *abstract container* è già associata ad un altro elemento.
`WrongInstrument (char *)`	Excerpt	Si è cercato di aggiungere ad uno strumento A i dati relativi ad un brano eseguito da uno strumento B≠A.
`WrongFeatures()`	Excerpt	Le *feature* del brano che si è cercato di classificare o di aggiungere al *dataset* sono diverse da quelle del `FeatureSet` corrente.
`IndexOutOfRange (int)`	FeatureSet	L'indice non è associato ad alcuna *feature*.
`CovNotPosDef (char *)`	Instrument	La matrice di covarianza relativa allo strumento descritto nel parametro non è definita positiva. L'utente è invitato a fornire un maggior numero di osservazioni linearmente indipendenti o ad escludere lo strumento dal *dataset*.

Tabella 5.1. Eccezioni sollevate nell'ambito dell'applicazione realizzata.

Lo stesso può dirsi dei file di salvataggio della sequenza di addestramento. Infatti, permettere all'utente di accedere e modificare manualmente questi file non può che aumentare la flessibilità dello strumento. Si è quindi in presenza di due formati proprietari di file di testo, che sono di seguito specificati attraverso la notazione *Backus-Naur Form* (BNF). A scanso di equivoci, comunque, la distribuzione dell'applicazione è stata corredata di file esemplificativi.

5.4.5.1 File delle informazioni relative ad un brano (`*.exc`)

Ogni brano è caratterizzato dal nome del file PCM da cui sono stati estratti i dati, dalla sessione di registrazione e dallo strumento che lo ha suonato, che può anche essere ignoto. Seguono il numero e la descrizione delle variabili estratte, e i vettori dei dati associati ad ogni nota, di cui vengono registrate anche il *pitch*, la durata e la posizione all'interno del file audio.

All'interno delle seguenti BNF si dà per scontata la sintassi di alcuni simboli non-terminali comuni, come `<identifier>`, `<integer>` e `<string>`. Con `<CRLF>` si intende la sequenza di nuova linea adottata dal sistema operativo in uso.

```
<excFile> ::= <Instrument_name> <PCM_file_name>
              <Recording_session> <FeatureSet>
              <List_of_values>

<Instrument_name> ::= <identifier> <CRLF>

<PCM_file_name> ::= <filename> <CRLF> ; completo di path

<Recording_session> ::= <string> <CRLF>

<FeatureSet> ::= NumberOfFeatures: <ws> <integer> <CRLF>
                 *(<identifier> <ws>) <CRLF>
                 ; il numero di <identifier> deve essere
                   pari al valore di <integer>

<List_of_values> ::= *<value>

<value> ::= <Position> <ws> <Duration> <ws> <Pitch> <ws>
            <Feature_vector> <CRLF>

<Feature_vector> ::= *(<real> <ws>)
```

```
                    ; il numero di ripetizioni deve
                      essere pari al valore di <integer>
                      in <FeatureSet>

<Position> ::= <real>

<Duration> ::= <real>

<Pitch> ::= <real>

<ws> ::= *(" " | \t | <CRLF>) ; whitespace
```

5.4.5.2 File di salvataggio del *dataset* (`*.ip`)

Questi file sono molto più semplici, e conservano, oltre alle informazioni relative al `FeatureSet`, i nomi dei file `.exc`, completi di percorso, che costituiscono il *dataset*, e il nome dello strumento a cui vanno associati.

```
<ipFile> ::= <FeatureSet> <Excerpts_List>

<Excerpts_List> ::= *<Excerpt>

<Excerpt> ::= <Instrument_name> <ws> <exc_file_name>
              <ws> <CRLF>

<exc_file_name> ::= <filename>
```

A questo punto è possibile avere un quadro più completo dell'architettura dell'applicazione, anche in relazione alle interazioni con gli altri sistemi. Nella figura 5.5 sono riassunti i diversi flussi informativi di cui ci si è occupati in questa sezione, e i diversi tipi di file ad essi associati. Si tratta di un diagramma qualitativo senza alcuna valenza formale, che visualizza intuitivamente alcune tecniche adottate e l'utilizzo dei file di interscambio.

Nella parte superiore è illustrato il processo di *fast prototyping*, che consiste nello sviluppare e testare le *routine* statistiche nel linguaggio di MATLAB, e tradurle in seguito in codice C++. Nella parte centrale è evidenziata la creazione dei file `.exc` da parte del sistema di prossima realizzazione che si occupa della segmentazione dei file audio e dell'estrazione delle caratteristiche per ciascun evento individuato. Questi file possono far parte della sequenza di addestramento del sistema o essere classificati basandosi sul *dataset* esistente, rappresentato nella figura dal file `data1.ip`. Attraverso la creazione

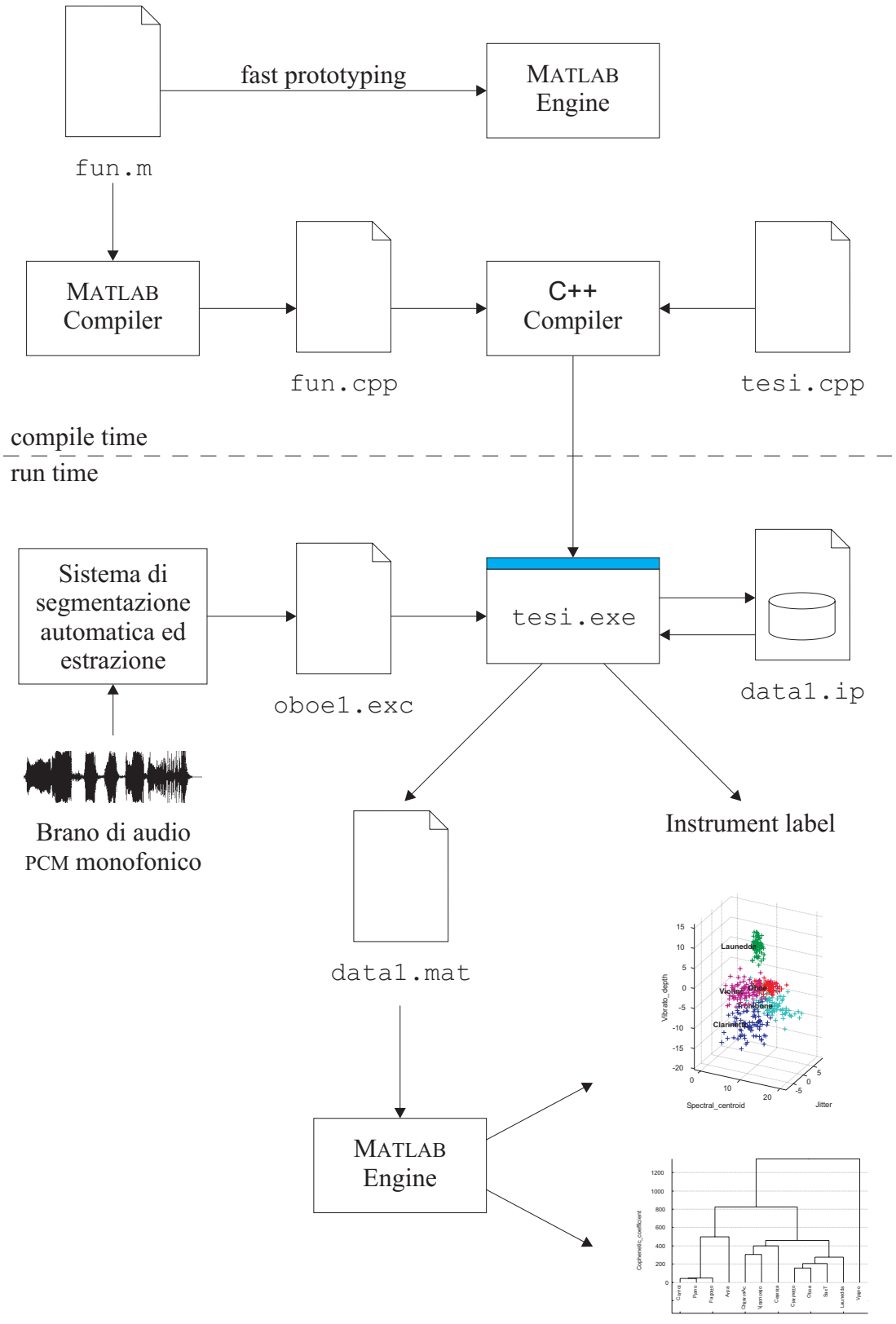


Figura 5.5. Diagramma riassuntivo dell'architettura del sistema.

dei file `.mat`, in formato proprietario MATLAB, è stato possibile visualizzare il *dataset* e i dendrogrammi sfruttando il *viewer* di MATLAB.

5.5 Modalità di test

I singoli moduli sono stati testati attraverso tecniche di tipo *black-box*, generando manualmente dei casi di test appartenenti alle diverse classi di equivalenza. Per i test di integrazione e per il testing globale dell'applicazione sono state realizzate delle procedure di generazione automatica dei casi di test, anche perché durante questa fase non si avevano a disposizione dati "reali." Le popolazioni di prova, perciò, consistono in realizzazioni di misture multinormali, attraverso tecniche di simulazione statistica, che rappresentano, idealmente, le classi associate ad ogni strumento. I dati relativi ai brani da classificare sono stati generati con la stessa tecnica.

Non si è sentita la necessità di adottare la tecnica del *built-in self-test*, in quanto gli invarianti di classe più immediati, come la definita positività delle matrici di covarianza degli oggetti di tipo `Instrument`, possono non essere soddisfatti negli stati di inconsistenza, ovvero prima dell'operazione di commit (stato Dirty, figura 5.1). Inoltre, è stato fatto ampio utilizzo delle STL e delle altre librerie standard, nonché di algoritmi statistici dalle solide basi teoriche, che non necessitano di uno *stress-testing* in questo senso.

Sono stati sintetizzati diversi tipi di sequenze di *training*, variando, oltre alle medie e alle matrici di covarianza delle singole classi, la cardinalità dei brani e il numero di dimensioni. Nelle figure 5.6–5.8 sono stati riportate le rappresentazioni grafiche di popolazioni a due, tre e cinque dimensioni, quest'ultima proiettata sulle tre variate canoniche principali, come descritto nel paragrafo 4.5.4. I nomi associati alle classi sono volutamente poco evocativi, in quanto non sono stati estratti da brani reali.

Il *dataset* della figura 5.6, il più interessante ed al contempo il più facilmente interpretabile tra quelli mostrati, dà luogo ai dendrogrammi rappresentati nelle figure 5.9–5.12, originati da diversi metodi di raggruppamento, illustrati nel paragrafo 3.3.3.1. Nella didascalia è riportato il coefficiente di correlazione, indice della distorsione indotta sui dati (paragrafo 4.5.5). Le differenze tra questi dendrogrammi riflettono le diverse interpretazioni che possono essere attribuite alla figura 5.6, e si può intuire che daranno luogo a diverse strutture gerarchiche, anche per uno stesso valore della soglia ξ (paragrafo 4.5.6).

Nella figura 5.13 si riporta la struttura gerarchica di decisione ottenuta a partire dal dendrogramma della figura 5.9. La rappresentazione grafica, al contrario delle altre figure di questo paragrafo, non è stata generata au-

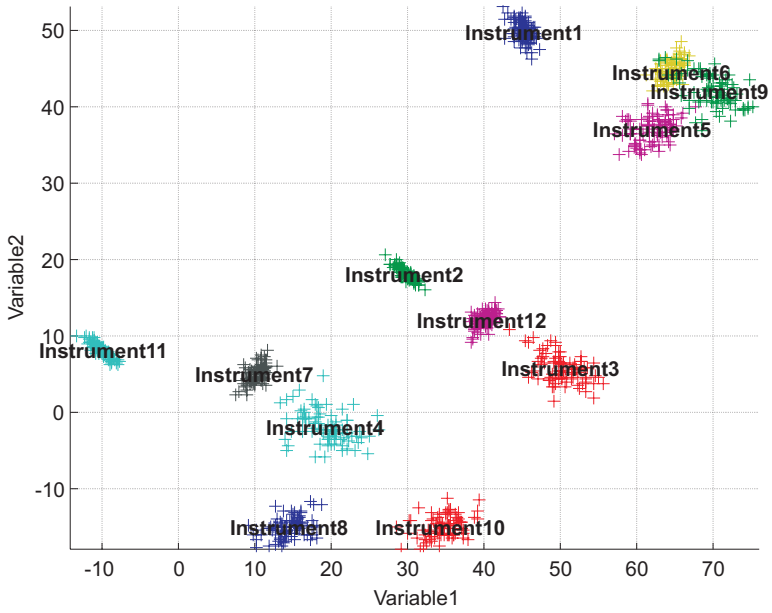


Figura 5.6. Rappresentazione grafica di una popolazione bidimensionale di test generata automaticamente.

tomaticamente. Nella figura 5.14 si riportano i risultati della classificazione gerarchica relativa ad una matrice di dati

$$\mathbf{D}_{\text{unkn}} = \begin{bmatrix} 20,1938 & -7,51625 \\ 16,4656 & -6,73588 \\ 17,5604 & -7,247939 \end{bmatrix}. \tag{5.1}$$

È stata adottata in ogni nodo la tecnica dell'analisi discriminante quadratica (QDA, paragrafo 3.2.4). Nella figura sono stati colorati i nodi decisionali attraversati, ed è stato riportato il punteggio relativo ad ogni nodo valutato, in termini di probabilità a posteriori. Essendo il *dataset* sintetizzato con medie e varianze arbitrarie, non si ritiene opportuno dettagliare ulteriormente i risultati della classificazione. Un'analisi più approfondita su dati reali sarà effettuata nella sezione 6.1.

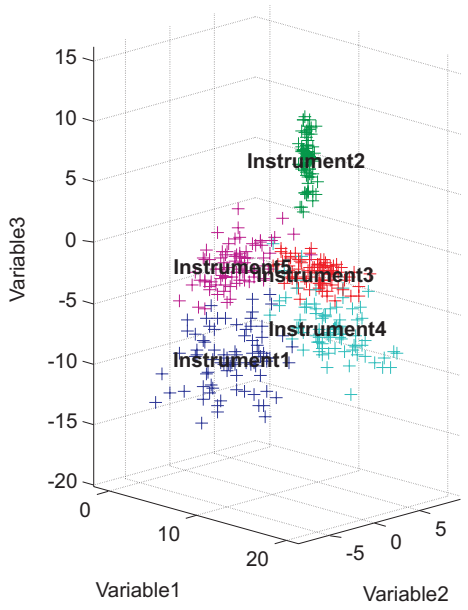


Figura 5.7. Rappresentazione grafica di una popolazione tridimensionale di test generata automaticamente.

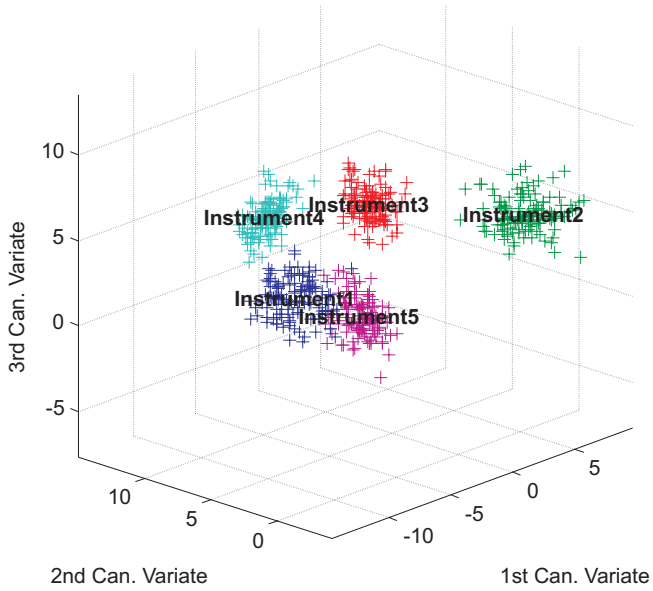


Figura 5.8. Rappresentazione grafica di una popolazione di test a cinque dimensioni, proiettata sulle prime tre variate canoniche.

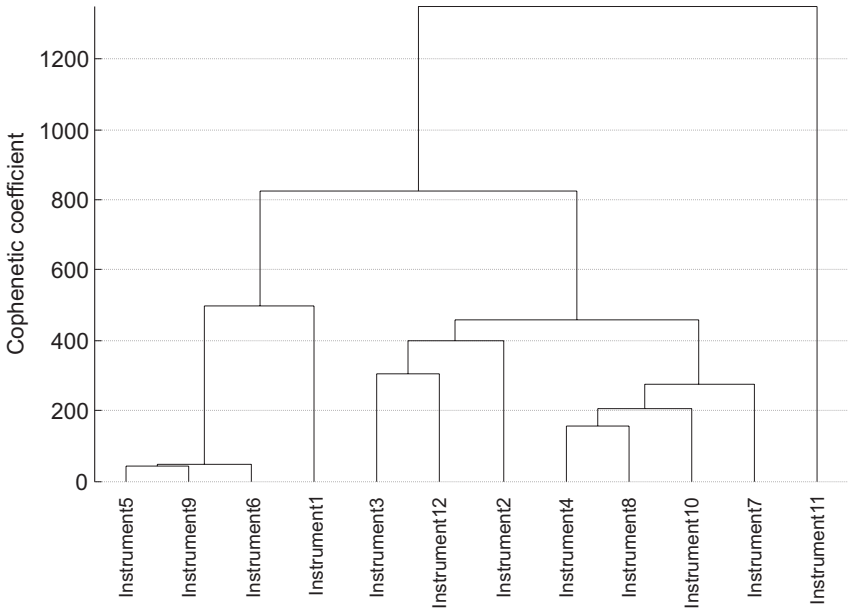


Figura 5.9. Dendrogramma ottenuto dal *dataset* della figura 5.6 utilizzando il metodo di *single linkage*. L'indice di correlazione è pari a $0,56$.

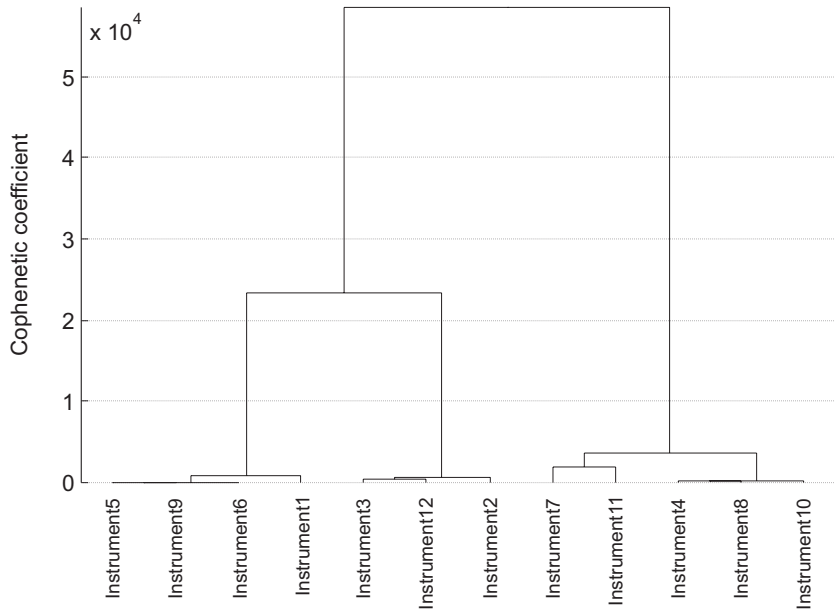


Figura 5.10. Dendrogramma ottenuto dal *dataset* della figura 5.6 utilizzando il metodo di *complete linkage*. L'indice di correlazione è pari a $0,32$.

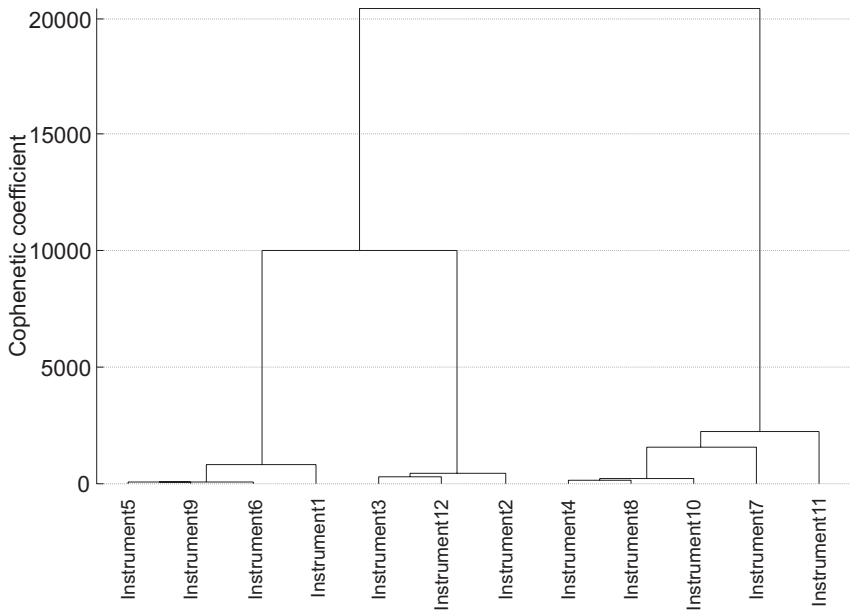


Figura 5.11. Dendrogramma ottenuto dal *dataset* della figura 5.6 utilizzando il metodo di Sokal e Sneath. L'indice di correlazione è pari a $0,33$.

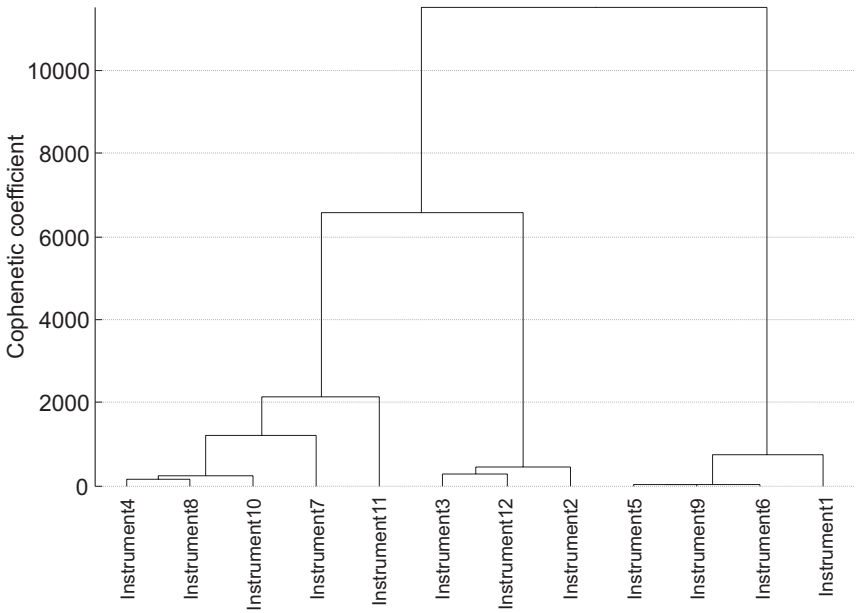


Figura 5.12. Dendrogramma ottenuto dal *dataset* della figura 5.6 utilizzando il metodo di *average linkage*. L'indice di correlazione è pari a $0,35$.

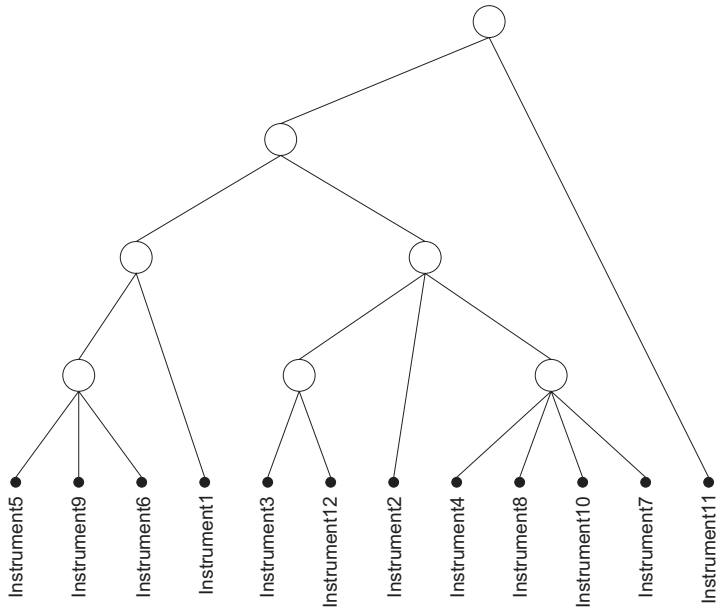


Figura 5.13. Struttura gerarchica di decisione ottenuta a partire dal dendrogramma della figura 5.9 con una soglia $\xi = 0,05$.

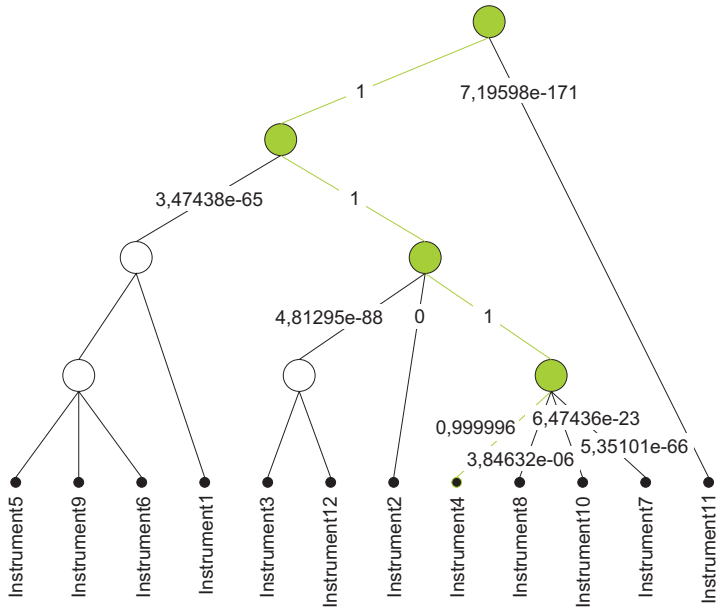


Figura 5.14. Rappresentazione grafica del processo di classificazione gerarchica per la (5.1).

capitolo 6

Risultati e conclusioni

Il presente lavoro adotta un nuovo approccio alla classificazione delle timbriche musicali e fornisce al ricercatore uno strumento preciso e flessibile per lo studio delle caratteristiche significative di uno strumento o gruppo di strumenti. Il classificatore sviluppato introduce la tecnica del *multilayer clustering*, illustrata nel paragrafo 4.2.1, che consente di ottenere la struttura gerarchica di decisione automaticamente, al contrario degli altri studi recenti in questa direzione [20, 63, 62] in cui è prefissata. Questo comporta numerosi vantaggi, riportati nel paragrafo 4.2.3, quali una maggiore efficienza, la modularità delle diverse tencniche di classificazione e la conseguente estendibilità del sistema, l'adattatività rispetto a nuove classi introdotte successivamente.

Il maggiore punto di forza e di originalità del classificatore consiste nella modalità con cui è stato aggirato il problema della dimensionalità, associando ad ogni nodo decisionale un insieme sub-ottimo di *feature*, sulla base di test di rilevanza statistica.

Durante le prime prove effettuate su misture multinormali, è subito stata individuata l'inadeguatezza di alcune misure statistiche proposte in letteratura per il particolare problema affrontato. In dettaglio, la misura di affinità di Bhattacharyya (4.9) dà origine a dendrogrammi troppo "schiacciati," mentre l'utilizzo dell'indice di inconsistenza proposta nel pacchetto statistico di MATLAB (4.14) per il processo di *multilayer clustering* fornisce strutture gerarchiche più frammentarie e difficilmente interpretabili rispetto alla semplice regola euristica esposta nel paragrafo 4.5.6.

Il *framework* di ricerca di questo progetto promette la realizzazione di un sistema di riconoscimento timbrico a partire dai dati audio monotimbrico nel breve periodo, non appena saranno disponibili dati estratti da un

corpus sufficientemente ricco di esecuzioni di strumenti musicali. L'applicazione consentirà altresì, grazie agli strumenti offerti al ricercatore elencati nella sezione 4.4, di raggiungere risultati di per sé interessanti riguardo la significatività di ciascuna caratteristica analizzata, singolarmente o congiuntamente ad altre, per determinati strumenti o famiglie di strumenti, nonché l'invarianza delle stesse *feature* in questi contesti. La stessa costruzione automatica della gerarchia di timbriche può trovare facilmente applicazione, per esempio, in strumenti di supporto alla composizione musicale.

La tecnica di classificazione ideata, nonostante sia stata progettata espressamente per le timbriche musicali, ha validità generale, e può essere utilizzata anche in altri ambiti. Per esempio, sempre all'interno dell'*audio analysis*, si può facilmente immaginare un suo impiego per problematiche di *speaker id* (paragrafo 1.1).

6.1 Esperimento di riconoscimento basato su dati reali

In questa sezione saranno illustrati i risultati relativi a un esperimento condotto su un *dataset* "reale," ovvero ricavato a partire da sequenze audio di esecuzioni di strumenti musicali.

6.1.1 I dati utilizzati

Allineandosi alla maggior parte degli esperimenti resi noti in letteratura, per le fonti sonore sono stati utilizzati i McGill University Master Samples (MUMS [76]). Si tratta di registrazioni di esecuzioni di scale cromatiche per 37 strumenti musicali, con diverse tecniche di esecuzione.

Alcuni di questi brani sono stati segmentati, individuando le note, e sono state estratte alcune caratteristiche, come illustrato nella figura 4.1. Le note dei MUMS sono lunghe e ben distinte, ed è stato quindi adottato un algoritmo di segmentazione relativamente semplice, basato su una soglia del valore relativo di RMS (vedi nota a pagina 25).

Nella tabella 6.1 sono riassunte e commentate le *feature* estratte dal file PCM, tutte note dalla letteratura (sezione 2.5) e già adottate in numerosi sistemi di classificazione (sezione 2.6). Molte di esse sono basate sul valore del *pitch*, che è stato ottenuto attraverso il *pitch-tracker* realizzato da Meroni [69]. Ciascuna caratteristica è stata calcolata su finestre di segna-

Feature	Commento
Energia delle prime quattro parziali	Sia f_0 la frequenza fondamentale, risolta cercando un picco dell'energia spetrale in un'intorno della frequenza stimata per il *pitch*. L'energia $E_{\mathrm{p}0}$ relativa alla parziale di ordine zero viene calcolata integrando l'energia su un'intervallo simmetrico centrato in f_0 e normalizzando per l'energia totale. Siano ora $f_i \stackrel{\text{def}}{=} (i+1)f_0$ le frequenze relative alle armoniche di ordine superiore. Le frequenze delle parziali successive $f_{\mathrm{p}i}$ (vedi nota a pagina 12) e le energie associate $E_{\mathrm{p}i}$ vengono calcolate allo stesso modo.
Scostamento tra parziali e armoniche	$\delta \stackrel{\text{def}}{=} \sum_{i=1}^{4} \lvert f_i - f_{\mathrm{p}i} \rvert$
Armonicità	$h \stackrel{\text{def}}{=} \sum_{i=1}^{4} \lvert f_i - f_{\mathrm{p}i} \rvert E_{\mathrm{p}i}$
Centroide spettrale	$c \stackrel{\text{def}}{=} \frac{\sum_{f=f_{\min}}^{f_{\max}} fE(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)}$
Ampiezza di banda	$B \stackrel{\text{def}}{=} \frac{\sum_{f=f_{\min}}^{f_{\max}} \lvert c-f \rvert E(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)}$
Tasso di attraversamento dello zero (*zero crossing rate*)	Determinato dalla frequenza con cui il valore dell'ampiezza istantanea cambia segno. Questo valore tende a dare valori anomali in segnali rumorosi.

Tabella 6.1. Caratteristiche rappresentative estratte dai campioni.

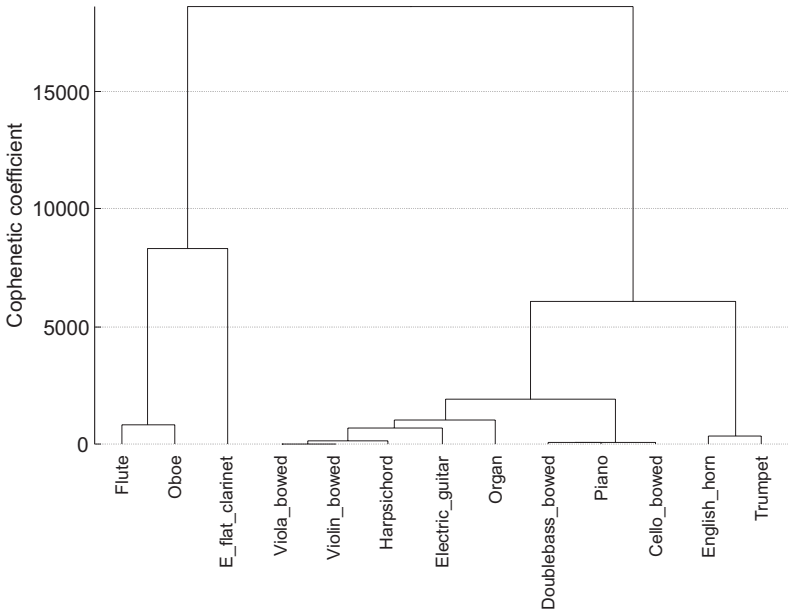


Figura 6.1. Dendrogramma ottenuto utilizzando il metodo di legame completo e il coefficiente di somiglianza dato dalla (4.8).

le parzialmente sovrapposte[1]; al fine di ottenere un singolo valore per ogni nota, è stato necessario utilizzare alcune tecniche di riepilogo descritte nel paragrafo 2.5.5, in particolare la media e la deviazione standard. Il numero di variabili disponibili è quindi pari al doppio di quello delle caratteristiche estratte, ovvero diciotto. Il numero di note analizzate varia da strumento a strumento, e in media sono state trenta. I tredici strumenti analizzati sono: tromba, flauto, clarinetto, oboe, corno inglese, clavicembalo, organo, pianoforte, chitarra elettrica, viola, violino, violoncello e contrabbasso (suonati con l'archetto).

6.1.2 La costruzione della gerarchia

Il dendrogramma relativo al *dataset* descritto è riportato nella figura 6.1. Balza subito all'occhio che la gerarchia differisce considerevolmente da quella tradizionalmente applicata agli strumenti occidentali, e spesso è in contra-

[1]Le finestre utilizzate hanno dimensioni pari a 2048 campioni, e sono state sovrapposte di 1024 campioni, adottando la funzione di *windowing* di Hamming.

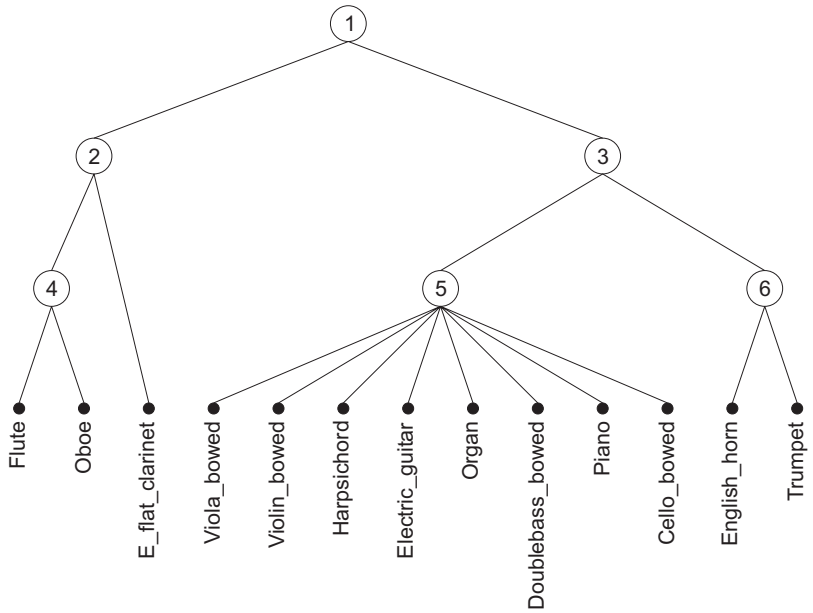


Figura 6.2. Struttura gerarchica ricavata dal dendrogramma della figura 6.1 attraverso la tecnica di *multilayer clustering* descritta nel paragrafo 4.5.6, utilizzando una soglia $\xi = 0,2$.

sto con la percezione umana di somiglianza timbrica. Ad esempio, gli archi sono stati separati in due gruppi distinti, e si può dire che l'oboe e il corno inglese siano agli antipodi. Questo non stupisce, infatti le *feature* riportate nella tabella 6.1 ricadono tutte nella classe delle caratteristiche spettrali (paragrafo 2.5.3), tenendo peraltro conto delle loro variazioni all'interno della nota calcolandone semplicemente la deviazione standard. L'apparato uditivo umano, invece, distingue le timbriche anche in base alla forma dell'inviluppo e alle modulazioni in ampiezza e in frequenza (sezione 2.5). Ecco quindi che la percussività e l'assenza di modulazioni periodiche nel suono del pianoforte, estraendo solo queste variabili, vengono definitivamente perse. L'estrazione di altri tipi di caratteristiche, specie quelle relative all'inviluppo, richiede studi più approfonditi, che consentano una segmentazione più accurata di quella fornita dagli algoritmi attualmente utilizzati.

Dal dendrogramma è stata automaticamente derivata la gerarchia della figura 6.2. Per poterla meglio comprendere, si riportano le rappresentazioni dei dati di ciascun nodo decisionale nelle figure 6.3–6.8. La maggior parte di essi presenta due sole sottoclassi, e quindi è disponibile solo una variata canonica ($k-1=1$, vedi paragrafo 3.2.6). In questi casi, la tipica proiezione

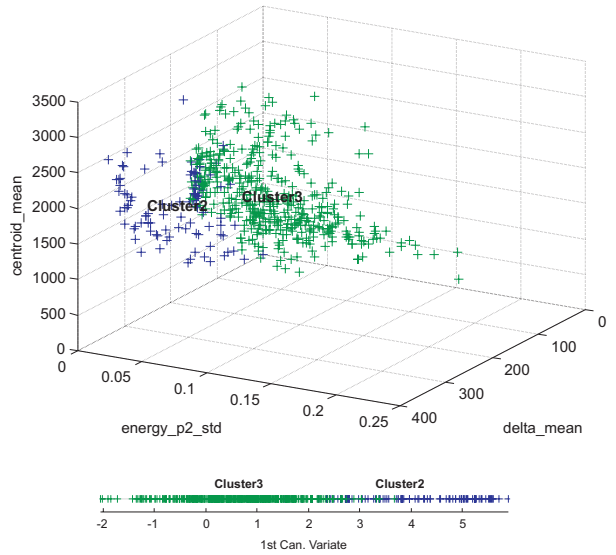


Figura 6.3. Proiezione sulle tre variabili maggiormente discriminanti e sull'unica variata canonica del nodo 1 della figura 6.2.

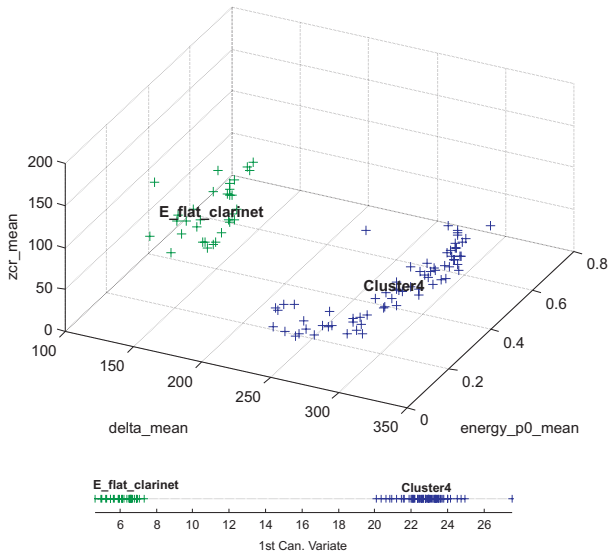


Figura 6.4. Proiezione sulle tre variabili maggiormente discriminanti e sull'unica variata canonica del nodo 2 della figura 6.2.

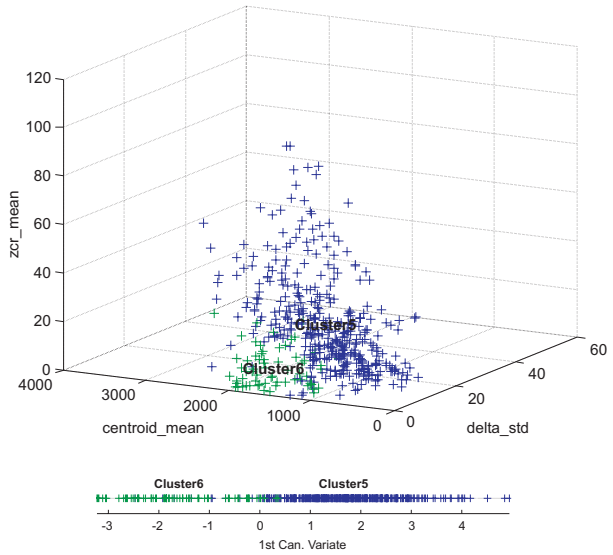


Figura 6.5. Proiezione sulle tre variabili maggiormente discriminanti e sull'unica variata canonica del nodo 3 della figura 6.2.

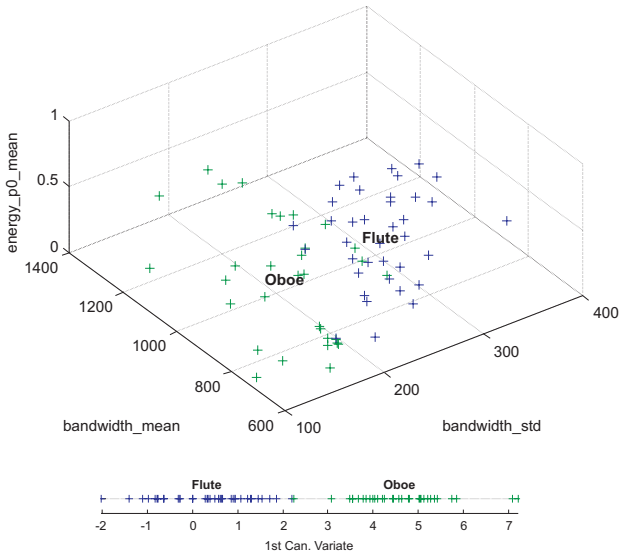


Figura 6.6. Proiezione sulle tre variabili maggiormente discriminanti e sull'unica variata canonica del nodo 4 della figura 6.2.

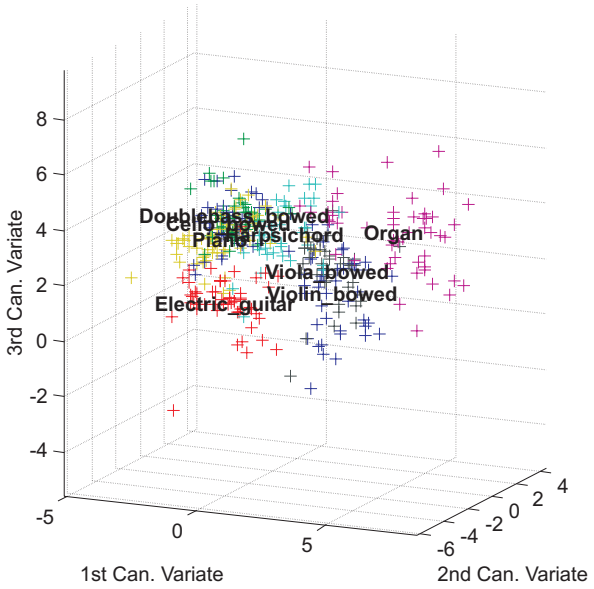


Figura 6.7. Proiezione sulle prime tre variate canoniche del nodo 5 della figura 6.2.

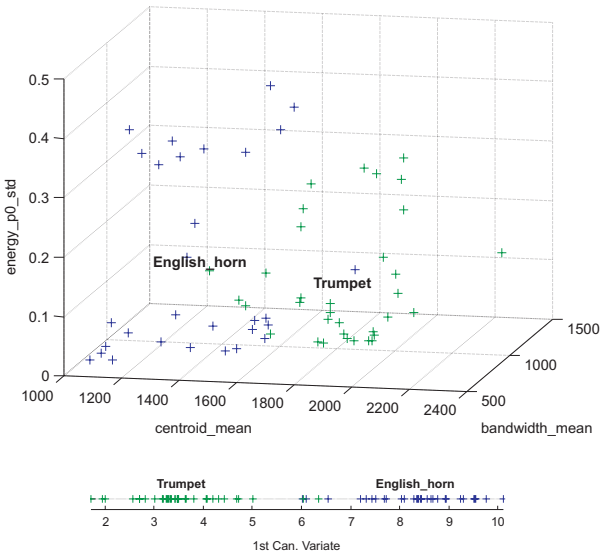


Figura 6.8. Proiezione sulle tre variabili maggiormente discriminanti e sull'unica variata canonica del nodo 6 della figura 6.2.

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`delta_mean`	1
2	`centroid_mean`	0,247
3	`delta_std`	0,205
4	`centroid_std`	0,133
5	`bandwidth_std`	0,110
6	`zcr_mean`	0,0994
7	`energy_p0_mean`	0,0861
8	`bandwidth_mean`	0,0937
9	`energy_p2_std`	0,0829
10	`harmonicity_std`	0,0853

Tabella 6.2. Lista delle dieci caratteristiche maggiormente rappresentative per l'intera popolazione.

sulle variate canoniche illustrata nel paragrafo 4.5.4 ed esemplificata nella figura 5.8 risulta quindi monodimensionale e poco suggestiva, ed è stata integrata con un grafico tridimensionale che riporta sugli assi le variabili che sono maggiormente rilevanti per la decisione all'interno del nodo studiato (cfr. tabelle 6.3–6.8). Nei *cluster* che contengono un gran numero di scelte, questi grafici possono non essere immediatamente interpretabili, a causa della limitatezza delle tre dimensioni—si ricorda che il numero di variabili originarie è $p = 18$.

Nella tabella 6.2 sono elencate le caratteristiche maggiormente discriminanti all'interno dell'intero *dataset*. Nelle tabelle 6.3–6.8 si riportano invece quelle relative ai singoli nodi di decisione. Il valore riportato a fianco di ogni *feature* è l'indice di rilevanza normalizzato dato dalla (3.106), nel contesto di una selezione in avanti (*forward selection*, paragrafo 3.2.11, test 12). In pratica, esso può essere interpretato come misura dell'informazione introdotta da quella variabile oltre a quella già presente grazie alle variabili che la precedono nella lista. Per questo motivo, può accadere che una variabile più in basso presenti valori più alti. Nella tabella 6.2, ad esempio, dopo che le prime sei variabili sono state selezionate, la media dell'ampiezza di banda apporta maggiore informazione in coppia con l'energia della parziale fondamentale, che non da sola. Inoltre, dovendone scegliere una soltanto tra le due, la seconda è da preferirsi.

Dalla lettura delle tabelle emergono parecchie informazioni interessan-

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`delta_mean`	1
2	`energy_p2_std`	0,453
3	`centroid_mean`	0,315
4	`harmonicity_std`	0,115
5	`energy_p2_mean`	0,0923
6	`energy_p0_mean`	0,109

Tabella 6.3. Lista delle sei caratteristiche maggiormente rappresentative per il nodo di decisione 1 nella figura 6.2.

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`delta_mean`	1
2	`energy_p0_mean`	0,538
3	`zcr_mean`	0,657
4	`delta_std`	0,603
5	`energy_p3_std`	0,313
6	`energy_p2_std`	0,244

Tabella 6.4. Lista delle sei caratteristiche maggiormente rappresentative per il nodo di decisione 2 nella figura 6.2.

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`delta_std`	0,579
2	`centroid_mean`	0,786
3	`zcr_std`	1
4	`bandwidth_mean`	0,620
5	`energy_p2_std`	0,274
6	`delta_mean`	0,250

Tabella 6.5. Lista delle sei caratteristiche maggiormente rappresentative per il nodo di decisione 3 nella figura 6.2.

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`bandwidth_std`	0,960
2	`bandwidth_mean`	0,450
3	`energy_p0_mean`	1
4	`delta_std`	0,496
5	`centroid_mean`	0,717
6	`energy_p3_mean`	0,218

Tabella 6.6. Lista delle sei caratteristiche maggiormente rappresentative per il nodo di decisione 4 nella figura 6.2.

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`delta_mean`	1
2	`bandwidth_std`	0,340
3	`bandwidth_mean`	0,335
4	`energy_p0_mean`	0,192
5	`zcr_mean`	0,154
6	`centroid_mean`	0,197

Tabella 6.7. Lista delle sei caratteristiche maggiormente rappresentative per il nodo di decisione 5 nella figura 6.2.

i	Variabile	$\Lambda_i / \max_i\{\Lambda_i\}$
1	`centroid_mean`	1
2	`bandwidth_mean`	0,292
3	`energy_p0_std`	0,175
4	`harmonicity_std`	0,360
5	`energy_p0_mean`	0,250
6	`zcr_std`	0,151

Tabella 6.8. Lista delle sei caratteristiche maggiormente rappresentative per il nodo di decisione 6 nella figura 6.2.

ti. Anzitutto, a conferma dei risultati disponibili in letteratura, il centroide spettrale e lo scostamento delle parziali dalle armoniche δ si sono rivelate molto spesso buoni discriminanti. Al contrario, come ci si aspettava, i valori dell'energia delle parziali di ordine superiore sono determinanti solo in alcuni gruppi. A questo proposito, le tabelle evidenziano una consistente discrepanza di contenuti, a significare che l'approccio della classificazione gerarchica, e in particolare il meccanismo di selezione automatica delle *feature* in ogni nodo di decisione, è in certi casi indispensabile. Ad esempio, l'ampiezza di banda riveste un ruolo centrale nella classificazione tra flauto e oboe (tabella 6.6), ma essa compare solo al quinto e ottavo posto nella "graduatoria generale" della tabella 6.2. Quindi una classificazione tradizionale, ovvero "piatta," basata solo sulle *feature* migliori rischierebbe di trascurare questo fattore, e conseguentemente di confondere con maggiore probabilità i due strumenti. Questo, per inciso, è il motivo per cui le prestazioni dei classificatori non gerarchici peggiorano considerevolmente all'aumentare del numero di classi.

6.1.3 Convalida del sistema

Per verificare la validità del classificatore e delle *feature* adottate, è stata implementata la stima *leave-one-out* del tasso di errore, descritta nel paragrafo 3.2.7. Essa consiste nell'escludere, una per una, le osservazioni dalla sequenza di *training*, addestrare il sistema, e cercare di classificarla. In questo modo, mediando su tutti i risultati, si ottiene una misura dell'efficacia molto accurata e priva del fenomeno dell'*overfitting*.

Attraverso questa stima, è stato possibile controllare le prestazioni del sistema attraverso le intuitive[2] *matrici di confusione* (*confusion matrix*), in cui vengono tabulate in percentuale le risposte fornite ai vari tipi di stimolo. Gli elementi diagonali, evidenziati in grassetto, sono idealmente pari a 100. Nella figura 6.9 è riportata la matrice di confusione relativa ad una classificazione "piatta," che conferma i raggruppamenti ricavati per altra via nel dendrogramma della figura 6.1 e nella gerarchia. La tecnica di classificazione adottata per questa stima è stata la QDA, che ha fornito di gran lunga i risultati migliori.

Adottando la stessa stima, è stato possibile scegliere la migliore tra le due tecniche realizzate (QDA, paragrafo 3.2.4, e CDA, paragrafo 3.2.6) in ogni nodo decisionale. I risultati hanno in tutti i casi suggerito l'utilizzo dell'analisi discriminante quadratica, ma questo non significa che, con altri strumenti o

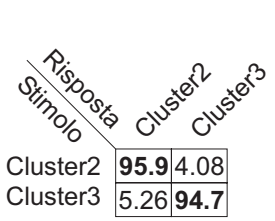
[2] a dispetto del nome...

Risposta / Stimolo	Flute	Oboe	E_flat_clarinet	Viola_bowed	Violin_bowed	Harpsichord	Electric_guitar	Organ	Doublebass_bowed	Piano	Cello_bowed	English_horn	Trumpet
Flute	**100**												
Oboe	3.23	**96.8**											
E_flat_clarinet			**100**										
Viola_bowed				**95.1**	4.88								
Violin_bowed				18.2	**81.8**								
Harpsichord						**100**							
Electric_guitar							**100**						
Organ								**100**					
Doublebass_bowed									**72.1**		27.9		
Piano									6.25	**82.8**	10.9		
Cello_bowed									4.35		**95.7**		
English_horn												**96.6**	3.45
Trumpet													**100**

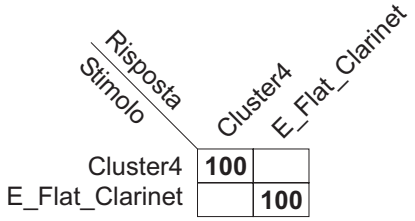
Figura 6.9. Matrice di confusione relativa alla classificazione "piatta" sull'intera popolazione.

con diverse *feature*, la CDA possa in alcuni nodi avere prestazioni superiori rispetto alla QDA (paragrafo 3.2.8). La modularità del classificatore rispetto ad altre tecniche di classificazione, inoltre, lascia supporre che, una volta codificati altri algoritmi, i nodi in cui la QDA sembra avere prestazioni peggiori, tipicamente a causa del mancato soddisfacimento delle ipotesi di multinormalità. La selezione automatica delle variabili maggiormente significative non è stata ancora implementata, e quindi dalla classificazione gerarchica ci si aspetta un comportamento peggiore.

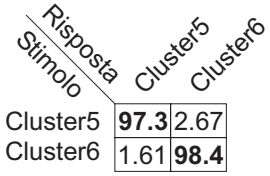
Le matrici di confusione relative ad ogni nodo sono illustrate nella figura 6.10, mentre la matrice riepilogativa della classificazione gerarchica è riportata nella figura 6.11, direttamente confrontabile con la figura 6.9. Confermando le aspettative, le prestazioni globali sono lievemente peggiorate. Si noti, comunque, che i tassi di errore nei livelli bassi sono diminuite—in particolare, si confrontino i risultati dei nodi 4 e 6 con i rispettivi valori della classificazione "piatta." I non trascurabili tassi di errore nei livelli alti sono responsabili di un "inquinamento" delle colonne relative a flauto, clarinetto e corno inglese. Si ritiene che l'imminente implementazione della selezione automatica delle *feature* migliori all'interno di ogni nodo farà in modo che le prestazioni della classificazione gerarchica aumentino, superando quelle della classificazione tradizionale.



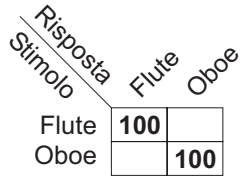
(a) Nodo 1



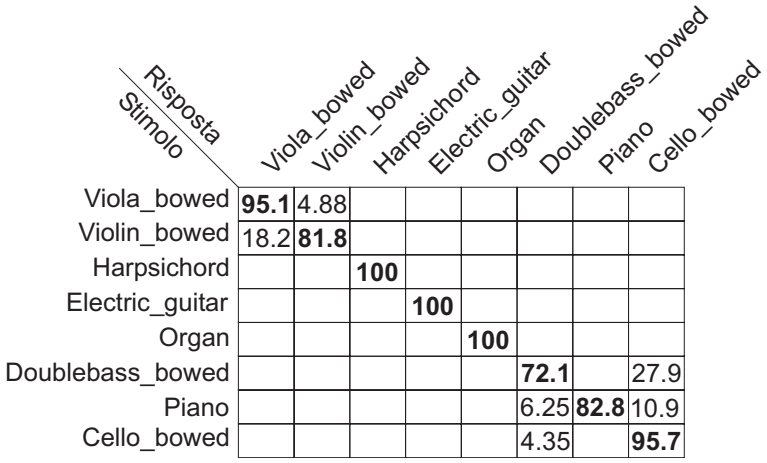
(b) Nodo 2



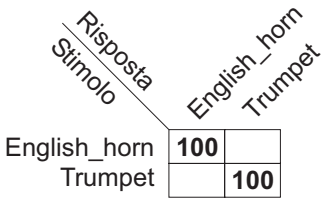
(c) Nodo 3



(d) Nodo 4



(e) Nodo 5



(f) Nodo 6

Figura 6.10. Matrici di confusione relative ai singoli nodi (*cluster*) della gerarchia.

Risposta / Stimolo	Flute	Oboe	E_flat_clarinet	Viola_bowed	Violin_bowed	Harpsichord	Electric_guitar	Organ	Doublebass_bowed	Piano	Cello_bowed	English_horn	Trumpet
Flute	**100**												
Oboe	3.23	**93.5**						3.23					
E_flat_clarinet			**83.9**	3.23	9.68							3.23	
Viola_bowed	4.88		2.44	**82.9**	4.88							2.44	2.44
Violin_bowed	4.55		4.55	15.9	**72.7**							2.27	
Harpsichord						**100**							
Electric_guitar							**94.1**					5.88	
Organ								**100**					
Doublebass_bowed	2.33								**69.8**		27.9		
Piano									6.25	**82.8**	10.9		
Cello_bowed	2.17								2.17		**95.7**		
English_horn			13.8									**86.2**	
Trumpet	3.03		12.1										**84.8**

Figura 6.11. Matrice di confusione relativa alla classificazione gerarchica.

Complessivamente, ci si può ritenere più che soddisfatti dei risultati ottenuti: non si ha notizia, per un sistema con 13 classi, di un tasso medio di errore migliore di quello ottenuto, pari al 6,1%. Confrontando i risultati ottenuti con quelli di Martin in un esperimento simile [62, sezione 6.5], si può ritenere di aver raggiunto prestazioni comparabili, se non superiori. Si consideri ad esempio la famiglia degli archi: mentre nel lavoro del MIT si riportano [62, tabella 28] percentuali di successo che vanno dal 15% al 50%, mentre nella figura 6.9 si leggono valori minimi del 72%. Il fatto che Martin abbia utilizzato un *dataset* con un maggior numero di strumenti (ventisette, per la precisione) non giustifica il divario tra i due risultati. Infatti, tranne nel caso della viola, la sequenza di addestramento adottata negli esperimenti trattati in questa sezione comprendono gli strumenti (tromba e flauto, oltre agli stessi membri della famiglia degli archi) con i quali sono stati confusi gli stimoli nell'esperimento citato. Si suppone che tutto questo sia dovuto al fatto che Martin effettua la classificazione prefissando rigidamente la gerarchia, e non sfrutta la correlazione tra le variabili, ipotizzandole indipendenti.

6.2 Sviluppi futuri

Va da sé che l'architettura del sistema presentato nei precedenti capitoli può essere migliorata sotto molti punti di vista. Alcuni banali, e facilmente realizzabili nel breve periodo, come l'aggiunta di un'interfaccia grafica all'applicazione, in modo da rendere più facile l'opera del ricercatore; altri più impegnativi, e alcuni addirittura irrealizzabili allo stato dell'arte.

Codifica di altre tecniche di classificazione

Il paradigma di classificazione gerarchica illustrato prevede l'utilizzo, in ogni nodo, della tecnica di classificazione avente il minimo tasso di errore stimato. La codifica di altre tecniche statistiche, come la k-NN, le *kernel rules*, o le *support vector machines*, di cui si è accennato nel paragrafo 3.2.9, ma anche di tecniche neurali, come le mappe auto organizzanti, migliorerebbe sicuramente le prestazioni del sistema, dal momento che, come visto nel paragrafo 3.2.8, ogni tecnica di classificazione ha il suo campo di applicabilità. Ad esempio, una tecnica non basata sull'ipotesi di multinormalità fornirà risposte migliori nei livelli più alti della gerarchia, dove questa assunzione è più difficilmente verificata.

Estensione a suoni articolati

Attualmente, il classificatore si basa sul concetto di "nota," o comunque di "evento sonoro." Un brano viene rigidamente suddiviso in note, e i valori delle *feature* relativi ai singoli eventi vengono presentati separatamente al sistema. Gli esperimenti psicoacustici esposti nel paragrafo 2.3.2 confermano che l'orecchio umano è molto sensibile alle fasi di transizione tra nota e nota, in presenza dei quali il tasso di corretto riconoscimento aumenta notevolmente. Incorporare nel modello anche queste caratteristiche significa considerare atomico il brano in ingresso nella figura 4.1, e quindi modificare, migliorare, ed accorpare i blocchi di segmentazione ed estrazione delle caratteristiche, che produrranno *un solo* vettore per ogni brano. In un certo senso, ci si sposterebbe da un'approccio riduzionistico ad uno più olistico e sintetico, vicino alla scuola di pensiero CASA (paragrafo 2.2).

Identificazione automatica di nuove timbriche

In alcuni casi (paragrafo 4.4) l'enucleazione di diverse timbriche all'interno di uno stesso strumento, dovute ad esempio a diverse tecniche di eccitazione, o ai diversi registri, può giovare al modello (figura 4.7). L'algoritmo di Expectation-Minimization (EM [24, sezione 9.3]) o più semplici algoritmi di

clustering non gerarchico (paragrafo 3.3.2) consentono l'automatizzazione di questo processo, che può essere attivato in presenza di valori troppo alti dei *p-value* relativi ai test di multinormalità e curtosi, o in presenza di un numero "preoccupante" di osservazioni atipiche.

Classificazione di sorgenti sonore arbitrarie

Lo studio effettuato, e di conseguenza il sistema realizzato, è imperniato sulla problematica della classificazione di timbriche di strumenti musicali. Estenderlo a sorgenti sonore arbitrarie, come versi di animali o rumori stradali, richiede cambiamenti minimi al classificatore, ma probabilmente un arricchimento notevole dell'insieme di caratteristiche da considerare. In questo caso, il meccanismo di selezione automatica delle caratteristiche diventa una stringente necessità.

Caratteristiche distribuite non normalmente

Assumere la normalità della distribuzione delle *feature* può portare a modelli poco aderenti alla realtà, e quindi a risultati insoddisfacenti. Martin ha cercato di affrontare questo problema nella sua tesi di dottorato [62, sezione 5.2], ma è stato costretto a ipotizzare che le variabili fossero ortogonali, rinunciando alla preziosa informazione legata alla loro correlazione.

Per alcune caratteristiche, tuttavia, come quelle legate alla modulazione in frequenza e in ampiezza, si sente l'esigenza di modelli più accurati di quello normale, data ad esempio da una mistura di due normali (paragrafo 2.5.5). Altre possibilità sono date da distribuzioni discrete o booleane, attualmente non rappresentabili nel sistema.

Coefficienti di attendibilità

Un'estensione abbastanza immediata, ma interessante è rappresentata dall'introduzione di valori di attendibilità (*reliability*) associati alle singole caratteristiche estratte. Alla fine della fase di addestramento basterebbe mediare sulle diverse osservazioni, e tenere conto di questi indici in fase di classificazione, con modalità che variano naturalmente al variare della tecnica di classificazione adottata.

Sorgenti polifoniche e politimbriche

Passare da un contesto monofonico ad uno polifonico, ma monotimbrico non sembra eccessivamente complicato, in questa area di ricerca, a patto di abbandonare l'approccio analitico basato sulla segmentazione e il *pitch tracking* e concentrarsi

su proprietà "di gruppo" che riguardino invariabilmente note singole e accordi. Purtroppo, gli studi effettuati finora in questa direzione hanno portato pochi risultati concreti, anche se il filone CASA (paragrafo 2.2) e l'utilizzo di *front-end* più accurati e fedeli al comportamento dell apparato uditivo umano, come il correlogramma, lasciano ben sperare per il prossimo futuro.

Per quanto riguarda l'estensione alle sorgenti polifoniche e politimbriche, eventualmente corrotte da rumore, non si è ancora in grado di stimare quando e come il problema possa essere affrontato con strumenti sufficientemente affidabili e accurati, a causa delle innumerevoli problematiche, già citate nella sezione 2.1, riguardanti l'identificazione dei flussi uditivi (*auditory streams* [65]) all'interno di audio complesso e il fenomeno di fusione (*blending* [85]) tra le diverse sorgenti.

Miglioramento delle caratteristiche

È auspicabile il miglioramento degli algoritmi di estrazione attualmente disponibili, specie per quelli relativi alle diverse fasi dell'inviluppo, basate ancora su soglie determinate empiricamente ed euristiche non meglio precisate. Ci si aspettano inoltre risultati positivi dall'arricchimento di questo insieme con nuove *feature* più complesse, più vicine alle proprietà fisiche di risonanza degli strumenti, come ad esempio le frequenze delle formanti.

6.2.1 Integrazione in un sistema CASA

Con riferimento allo spazio di astrazione introdotto da Bregman e McAdams in [65], il sistema implementato è collocabile al livello sensoriale/percettivo. Come illustrato nella figura 6.12, esso è idealmente integrabile con un sistema ad un più elevato grado di astrazione che sfrutti queste ed altre informazioni provenienti da altri sottosistemi (ad esempio un *pitch tracker*) per estrapolarne informazioni a livello cognitivo, sfruttando eventualmente anche le informazioni relative al contesto [17, 46, 65]. Questo sistema fornirebbe quindi un *modello di spiegazione* per il brano in ingresso che, come suggerisce Ellis in [18], può essere utilizzato per effettuare delle previsioni sui dati successivi (frecce verso il basso nel diagramma).

Ad esempio, se, analizzando un brano polifonico e politimbrico, il sistema esperto osserva una linea melodica di sassofono accompagnata da un contrabbasso pizzicato, si può ragionevolmente "aspettare" che le note con il *pitch*

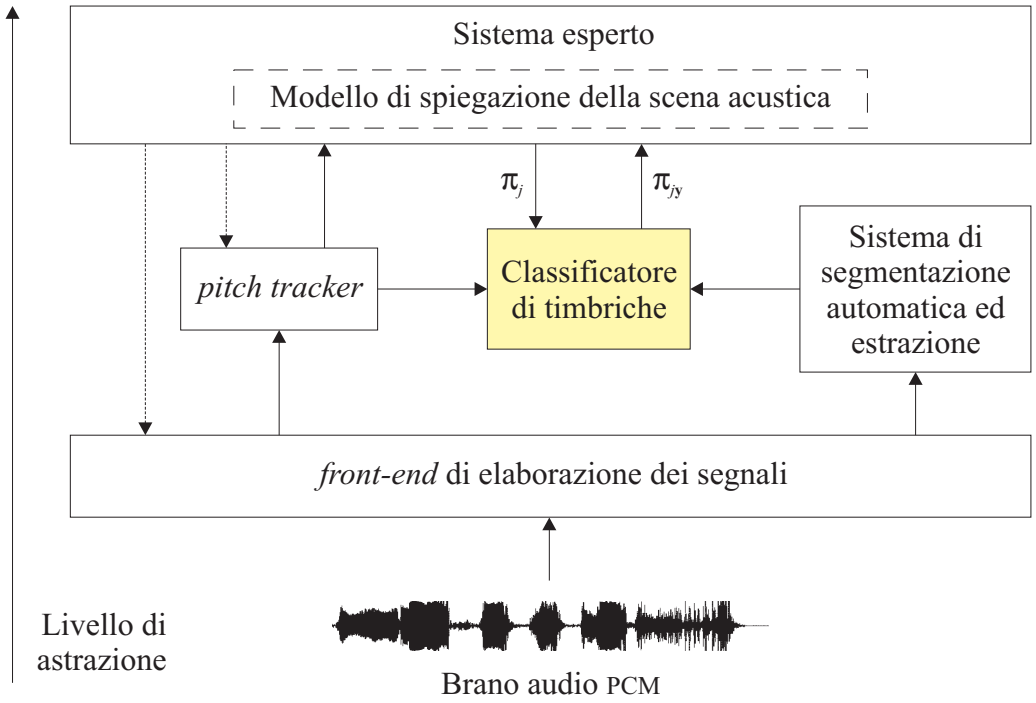


Figura 6.12. Architettura di una possibile integrazione di un sistema di riconoscimento timbrico con un sistema esperto. Le frecce verso il basso indicano eventuali flussi informativi (ad esempio valori di parametri dei blocchi di destinazione) atti ad influenzare i processi sensoriali e percettivi, sulla base delle "esperienze cognitive" del sistema esperto.

più alto apparterranno con buona probabilità al primo strumento. Ecco allora che l'informazione relativa alle probabilità a priori π_j, che altrimenti il classificatore è costretto ad assumere ugualmente pari a $\frac{1}{N}$, con N numero degli strumenti noti, viene fornita dall'alto, sulla base di un'insieme di regole codificate da esperti e musicologi.

Sarà proprio il sistema esperto ad effettuare la decisione finale riguardo al timbro di una nota avente il *pitch* fuori dagli intervalli consueti dello strumento che presenta valore massimo di probabilità a posteriori ($\pi_{j\mathbf{y}}$, equazione 3.6), e quindi è il migliore candidato. Come suggerisce Martin in [61], per la realizzazione di un tale sistema possono essere adottate numerose ipotesi parallele, che vengono risolte quando si ritenga di possedere sufficienti evidenze in favore di una di esse. Analogamente, il sistema può risolvere le possibili situazioni di incoerenza in cui un fraseggio sia solo parzialmente attribuito ad uno strumento, e per la rimanente parte ad un altro. Un siste-

ma esperto di questo tipo, quindi, potrebbe fornire una descrizione di alto livello della scena sonora, riconoscendo gli strumenti coinvolti, o quantomeno le loro famiglie di appartenenza, il loro ruolo musicale (solista, coro, accompagnamento, etc.) e magari, in collaborazione con altri moduli studiati *ad hoc*, il risultato complessivo o addirittura il genere (brano ritmato, melodico, concitato, etc.)

Gli esempi più importanti in letteratura di sistemi di questo tipo provengono dallo stesso Martin [60], che propone un'architettura a *blackboard* per la trascrizione automatica di musica polifonica, e da Klassner [54], che con il paradigma della elaborazione e comprensione integrate del segnale (*Integrated Processing and Understanding of Signals*, IPUS), mira ad una interazione tra i processi di comprensione e quelli di analisi del segnale volta ad adattare dinamicamente il *front-end* in risposta ai cambiamenti di scenario, rianalizzando i dati incerti o corrotti.

Una variante semplificata del problema della trascrizione automatica, o più in generale della descrizione della scena sonora, infine, consiste nel fornire esplicitamente al sistema le informazioni relative all'orchestrazione del brano audio da analizzare, attraverso gli identificativi delle timbriche coinvolte, o addirittura attraverso dei brani audio monotimbrici, rappresentativi di ciascuno strumento, possibilmente registrati nelle stesse condizioni ambientali e di esecuzione. Un'altra variante potrebbe consistere nel fornire al sistema le informazioni relative alla partitura. Problemi di questo tipo sono senza dubbio affrontabili con gli strumenti a disposizione attualmente, e, a parere dell'autore, risolvibili con risultati soddisfacenti nel medio periodo. Essi possono inoltre rappresentare un buon punto di partenza per traguardi più ambiziosi, o fornire idee per sistemi più evoluti. Non si fatica, ad esempio, a pensare ad un insieme di regole o di euristiche che consentano di isolare fraseggi o caratteristiche spettrali inequivocabili dei diversi strumenti all'interno del brano, riducendo il problema al suo corrispettivo "assistito."

Bibliografia

[1] E. J. Anderson. *Limitations of Short-Time Fourier Transforms in Polyphonic Pitch Recognition.* Ph.D. Thesis, University of Washington, 1997.

[2] American National Standards Institute. *American National Psychoacoustical Terminology.* S3.20. American Standards Association, New York, 1973.

[3] K. E. Atkinson. *An Introduction to Numerical Analysis.* Seconda Edizione, John Wiley & Sons, New York, 1989.

[4] B. Blankertz. "A robust constant Q spectrum for polyphonic pitch tracking." CCRMA DSP Seminar, 1999.
http://wwwmath.uni-muenster.de/math/inst/logik/org/staff/blankertz/DspPapers.html

[5] A. Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, MA, 1990.

[6] D. R. Brillinger, R. A. Irizarry. "An investigation of the second- and higher-order spectra of music." Berkeley Tech Report.

[7] J. C. Brown. "Calculation of a constant Q spectral transform." *Journal of the Acoustical Society of America* **89**(1), 425–434, 1991.

[8] J. C. Brown. "Musical instrument identification using pattern recognition with cepstral coefficients as features." *Journal of the Acoustical Society of America* **105**(3), 1933–1941, 1999.

[9] C. Burges. "A tutorial on Support Vector Machines for pattern recognition." *Data Mining and Knowledge Discovery* **2**(2), 1998.

[10] G. R. Charbonneau. "Timbre and the perceptual effects of three types of data reduction." *Computer Music Journal* **5**(2), 10–19, 1981.

[11] S. Crespi-Reghizzi. *Sintassi, semantica e tecniche di compilazione.* CLUP, Milano, 1990.

[12] E. Dedò, A. Varisco. *Algebra Lineare—Elementi ed esercizi.* CittàStudi, Milano, 1988.

[13] G. De Poli, P. Prandoni, P. Tonella. "Timbre clustering by self-organizing neural networks." *Atti del X Colloquio di Informatica Musicale*, a cura di G. Haus e I. Pighi, 102–108, 1993.

[14] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[15] A. J. Dobson. *An Introduction to Generalized Linear Models.* Chapman and Hall, New York, 1990.

[16] R. O. Duda, P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[17] D. P. W. Ellis. "Prediction-driven computational auditory scene analysis for dense sound mixtures." International Computer Science Institute, Berkeley, 1996.
http://www.icsi.berkeley.edu/~dpwe/research/pubs/

[18] D. P. W. Ellis. *Prediction-Driven Computational Auditory Scene Analysis.* Ph.D. Thesis, Massachussets Institute of Technology, 1996.
http://sound.media.mit.edu/~dpwe/pdcasa/

[19] D. P. W. Ellis. "The weft: a representation for periodic sounds." International Computer Science Institute, Berkeley, 1997.
http://www.icsi.berkeley.edu/~dpwe/research/pubs/

[20] A. Eronen. "Automatic musical instrument identification." Tampere University of Technology, 1999.
http://www.cs.tut.fi/~eronen/research.html

[21] R. A. Fisher. "The use of multiple measurements in taxonomic problems." *Ann. Eugen.* **7**, 179–188, 1936.

[22] M. Fowler, K. Scott. UML *Distilled: Applying the Standard Object Modeling Language.* Addison-Wesley, 1997.

[23] B. Flury. *Common Principal Components and Related Multivariate Models.* John Wiley & Sons, New York, 1988.

[24] B. Flury. *A First Course in Multivariate Statistics.* Springer-Verlag, New York, 1997.
Tabelle di dati e *routine* di classificazione disponibili presso ftp://129.79.94.6/pub/flury.

[25] B. Flury, H. Riedwyl. *Multivariate Statistics—A Practical Approach.* Chapman and Hall, New York, 1988.

[26] J. Foote. "An overview of audio information retrieval." National University of Singapore, 1997.
http://www.cs.princeton.edu
courses/archive/spr99/cs598b/foote_over.pdf

[27] J. Foote. "A similarity measure for automatic audio classification." National University of Singapore, 1997.

[28] A. Fraser, I. Fujinaga. "Toward real-time recognition of acoustic musical instrument." Johns Hopkins University, Baltimore, MD, 1997.

[29] G. Frazzini, G. Haus, E. Pollastri. "Cross automatic indexing of score and audio sources: Approaches for music archive applications." Presentato al ACM-SIGIR 1999, University of California, Berkeley, 1999.

[30] J. H. Friedman. "Regularized Discriminant Analysis." *Journal of American Statistical Association* **84**, 165–175, 1989.

[31] *La nuova enciclopedia della musica.* Garzanti, 1983.

[32] D. Gibson. "Name that clip: Music retrieval using audio clips." University of California, Berkeley, 1999.

[33] J. L. Goldstein. "Auditory spectral filtering and monaural phase perception." *Journal of the Acoustical Society of America* **41**, 458–478, 1967.

[34] Michel Goossens, Frank Mittelbach, Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, 1994.

[35] J. M. Grey. "Multidimensional perceptual scaling of musical timbres." *Journal of the Acoustical Society of America* **61**(5), 1270–1277, 1977.

[36] J. M. Grey. "Timbre discrimination in musical patterns." *Journal of the Acoustical Society of America* **64**(2), 467–472, 1978.

[37] J. M. Grey, J. W. Gordon. "Perceptual effects of spectral modifications on musical timbres." *Journal of the Acoustical Society of America* **63**(5), 1493–1500, 1978.

[38] J. M. Grey, J. A. Moorer. "Perceptual evaluations of synthesized musical instrument tones." *Journal of the Acoustical Society of America* **62**(2), 454–462, 1977.

[39] J. Hajda. "A new model for segmenting the envelope of musical signals: the relative salience of steady state versus attack, revisited." Proceedings of the 101st Convention of the Audio Engineering Society, Los Angeles, 1996.
http://www.ethnomusic.ucla.edu/systematic/Students/Hajda/backgroun.htm

[40] K. Han, Y. Par, S. Jeon, G. Lee, Y. Ha. "Genre classification system of TV sound signals based on a spectrogram analysis.." IEEE *Transactions on Consumer Electronics*, **44**(1), 33–42, 1998.

[41] J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, New York, 1975.

[42] D. M. Hawkins. "A new test for multivariate normality and homoscedasticity." *Technometrics* **23**, 105–110, 1981.

[43] H. L. F. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* F. Vieweg & Sohn, Braunschweig, 1863.

[44] H. Hotelling. "Relations between two sets of variates." *Biometrika* **28**, 321–377, 1936.

[45] H. Hotelling. "A generalized T test and measure of multivariate dispersion." Proceedings of the Second Berkeley Symposium, Berkeley. University of California Press, 23–41, 1951.

[46] A. J. M. Houtsma. "Pitch and timbre: Definition, meaning and use." *Journal of New Music Research* **26**, 104–115, 1997.

[47] C. Hourdin, G. Charbonneau, T. Moussa. "A multidimensional scaling analysis of musical intruments' time-varying spectra." *Computer Music Journal*, 40–55, 1997.

[48] P. Iverson, C. L. Krumhansl. "Isolating the dynamic attributes of musical timbre." *Journal of the Acoustical Society of America* **94**(5), 2595–2603, 1993.

[49] T. Jehan. *Musical Signal Parameter Estimation.* MS Thesis, CNMAT, Berkeley, 1997.

[50] I. Kaminskyj, A. Materka. "Automatic source identification of monophonic musical instrument sounds." Proceedings of the 1995 IEEE International Conference on Neural Networks, 189–194, 1995.

[51] K. Kashino, K. Nakadai, T. Kinoshita, H. Tanaka. "Organization of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism." Proceedings of the International Conference on Artificial Intelligence, Montréal, 1995.

[52] R. A. Kendall. "The role of acoustic signal partitions in listener categorization of musical phrase." *Music Perception* **4**(2), 185-214, 1986.

[53] Leslie Lamport. *LaTeX—A Document Preparation System—User's Guide and Reference Manual.* Addison-Wesley, Reading, 1994.

[54] V. R. Lesser, S. H. Nawab, F. I. Klassner. "IPUS: An architecture for the integrated processing and understanding of signals." *Artificial Intelligence Journal* **77**(1), 1995.
http://mas.cs.umass.edu/research/ipus/ipus.html

[55] D. Luce. *Physical Correlates of Nonpercussive Musical Instruments Tones.* Ph.D. Thesis, Massachussets Institute of Technology, 1963.

[56] P. C. Mahalanobis. "On the generalized distance in statistics." *Proceedings of the National Institute of Sciences India* **2**, 49–55, 1936.

[57] V. Maniezzo. *Algoritmi di apprendimento automatico.* Esculapio, Bologna, 1995.

[58] K. V. Mardia. "Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies." *Sankhyā B* **36**, 115–128, 1974.

[59] J. Marques, P. J. Moreno. "A study of musical instrument classificarion using gaussian mixture models and support vector machines."

Cambridge Research Laboratory Tech Report, 1999.
http://crl.research.compaq.com/
publications/techreports/techreports.html

[60] K. D. Martin. "A blackboard system for automatic transcription of simple polyphonic music." MIT Media Lab Technical Report No. 385, 1996.
http://sound.media.mit.edu/~kdm/research/

[61] K. D. Martin. "Toward automatic sound-source recognition: Identifying musical instruments." NATO *Computational Hearing Advanced Study Institute*, Italy, 1998.
http://sound.media.mit.edu/~kdm/research/

[62] K. D. Martin. *Sound-Source Recognition: A Theory and Computational Model.* Ph.D. Thesis, Massachussets Institute of Technology, 1999.
ftp://sound.media.mit.edu/pub/Papers/kdm-phdthesis.pdf

[63] K. D. Martin, Y. E. Kim. "Musical instrument identification: A pattern-recognition approach." MIT *Media Lab Machine Listening Group*, 1998.
http://sound.media.mit.edu/~kdm/research/

[64] K. D. Martin, E. D. Scheirer, B. L. Vercoe. "Music content analysis through models of audition." Proceedings of the 1998 ACM Multimedia Workshop on Content-Based Processing of Music. Bristol, UK, 1998.
ftp://sound.media.mit.edu/pub/Papers/ACMMM98.pdf

[65] S. McAdams, A. Bregman. "Hearing musical streams." *Computer Music Journal* *3*(4), 26–43, 1979.

[66] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley & Sons, New York, 1992.

[67] R. Meddis, M. J. Hewitt. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. Part I: Pitch identification." *Journal of the Acoustical Society of America* **89**(6), 2866–2882, 1991.

[68] R. Meddis, M. J. Hewitt. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. Part I: Phase sensitivity." *Journal of the Acoustical Society of America* **89**(6), 2883–2894, 1991.

[69] A. Meroni. "Modelli ed algoritmi per l'identificazione delle note musicali in un segnale audio." Tesi di laurea, Politecnico di Milano, 1999.

[70] M. Minsky. *The Society of Mind.* Simon & Schuster, 1986.

[71] A. M. Mood, F. A. Graybill, D. C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., 1974.

[72] J. A. Moorer, J. Grey. "Lexicon of analyzed tones, part 1: A violin tone." *Computer Music Journal* **1**(2), 39–45, 1977.

[73] J. A. Moorer, J. Grey. "Lexicon of analyzed tones, part 2: Clarinet and oboe tones." *Computer Music Journal* 12–29, June 1977.

[74] J. A. Moorer, J. Grey. "Lexicon of analyzed tones, part 3: The trumpet." *Computer Music Journal* **2**(2), 23–31, 1977.

[75] M. Norris. "Design decisions in an oscillatory model of primitive auditory segregation." University of Queensland, 1996.
http://www.cs.uq.edu.au/personal/michaeln/casa.html

[76] F. Opolko, J. Wapnick. "McGill University Master Samples." McGill University, Montreal, Quebec, 1987.
http://lecaine.music.mcgill.ca/newHome/mums/html/mums.html

[77] K. Pearson. "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine* ser. 6, **2**, 559–572, 1901.

[78] S. Pfeiffer, S. Fischer, W. Effelsberg. "Automatic audio content analysis." Mannheim Universität Technical Report, 1996.

[79] M. Piszczalski. "Spectral surfaces from performed music, part 1." *Computer Music Journal* **3**(1), 18–24, 1979.

[80] M. Piszczalski. "Spectral surfaces from performed music, part 2." *Computer Music Journal* **3**(3), 25–27, 1979.

[81] R. Plomp, J. M. Steeneken. "Effect of phase on the timbre of complex tones." *Journal of the Acoustical Society of America* **46**, 409–421, 1969.

[82] *UML Notation Guide* Rational Software, 1997.
http://www.rational.com/uml

[83] C. H. Romesburg. *Cluster Analysis for Researchers.* Robert E. Krieger Publishing, Malabar, Florida, 1990.

[84] E. Rosch. "Principles of categorization." *Cognition and Categorization*, a cura di E. Rosch e B. B. Lloyd, Lawrence Erlbaum, 1978.

[85] G. J. Sandell. "Roles for spectral centroid and other factors determining 'blended' instrument pairings in orchestration." *Music Perception* **13**(2), 209–246, 1996.

[86] E. D. Scheirer. *Extracting expressive performance information from recorded music.* MS Thesis, Massachussets Institute of Technology, 1995. http://sound.media.mit.edu/~eds/papers.html

[87] E. D. Scheirer. "The MPEG-4 Structured Audio standard." Proceedings of the 1998 IEEE ICASSP, Seattle, 1998. http://sound.media.mit.edu/~eds/papers.html

[88] E. D. Scheirer. *Music Perception Systems.* Ph.D. Thesis proposal, Massachussets Institute of Technology, 1998. http://sound.media.mit.edu/~eds/papers/phdprop/

[89] E. D. Scheirer. "The MPEG-4 Structured Audio Orchestra Language." Proceedings of the 1998 ICMC, Ann Arbor, MI, 1998. http://sound.media.mit.edu/~eds/papers.html

[90] E. D. Scheirer. "Structured Audio and effects processing in the MPEG-4 multimedia standard." *Multimedia Systems* **7**(1), 11–22, 1999. http://sound.media.mit.edu/~eds/papers.html

[91] E. D. Scheirer, Y. E. Kim. "Generalized audio coding with MPEG-4 Structured Audio." Proceedings of the Audio Engineering Society 17th International Conference on High-Quality Audio Coding, Firenze, 1999. http://sound.media.mit.edu/~eds/papers.html

[92] E. D. Scheirer, L. Ray. "Algorithmic and wavetable synthesis in the MPEG-4 multimedia standard." Proceedings of the 105th Meeting of the AES, San Francisco, 1998. http://sound.media.mit.edu/~eds/papers.html

[93] E. D. Scheirer, B. L. Vercoe. "SAOL: The MPEG-4 Structured Audio Orchestra Language." *Computer Music Journal* **23**(2), 31–51, 1999.

[94] M. Slaney. "Auditory toolbox." Apple Technical Report #45, Apple Computer, 1994. http://web.interval.com/papers/1998-010/

[95] M. Slaney. "Multi-model estimation and classification as a basis for computational timbre understanding." 1996.

[96] M. Slaney. "A critique of pure audition." *Readings in Computational Auditory Scene Analysis*, a cura di H. Okuno e D. Rosenthal, Erlbaum Publishing, Inc., 1996.
http://www.interval.com/
frameset.cgi?papers/1997-056/index.html

[97] M. Slaney, D. Naar, R. F. Lyon. "Auditory model inversion for sound separation." Proceedings of the 1994 IEEE ICASSP, Adelaide, Australia, 1994.

[98] H. Späth. *Cluster Analysis Algorithms.* Ellis Horwood Ltd., Chichester, 1980.

[99] D. T. Teaney, V. L. Moruzzi, F. C. Mintzer. "The tempered Fourier transform." *Journal of the Acoustical Society of America* **67**(6), 2063–2068, 1980.

[100] P. Toiviainen, M. Kaipainen, J. Louhivuori. "Musical timbre: Similarity ratings correlate with computational feature space distances." *Journal of New Music Research* **24**, 282–298, 1995.

[101] B. L. Vercoe, W. G. Gardner, E. D. Scheirer. "Structured Audio: Creation, transmission and rendering of parametric sound representations." *Proceedings of the* IEEE **86**(5), 922–940, 1998.

[102] P. J. Walmsley, S. J. Godsill, P. J. W. Rayner. "Bayesian modelling of harmonic signals for polyphonic music tracking." Cambridge Music Processing Colloquium, 1999.

[103] D. L. Wessel. "Timbre space as a musical control structure." *Computer Music Journal* **3**(2), 45–52, 1979.

[104] D. L. Wessel, J. Risset. "Exploration of timbre by analysis and synthesis." In *The Psychology of Music*, Academic Press, 1982.

[105] E. Wold, T. Blum, D. Keislar, J. Wheaton. "Content-based classification, search, and retrieval of audio." IEEE *Multimedia* **3**(3), 27–36, 1996.
http://www.musclefish.com/ieeemm96/index.html

[106] E. Wold, T. Blum, D. Keislar, J. Wheaton. "Classification, search, and retrieval of audio." CRC Handbook of Multimedia Computing, 1999.
http://www.musclefish.com/crc/index.html

[107] X. Yang, K. Wang, S. A. Shamma. "Auditory representations of acoustic signals." IEEE *Transactions on Information Theory* **38**(2), 1992.

[108] T. Y. Young, T. W. Calvert. *Classification, Estimation and Pattern Recognition.* American Elsevier Publishing Company, New York, 1974.

[109] T. Zhang. "Content-based classification and retrieval of audio." University of Southern California, 1998.
http://viola.usc.edu/
extranet/Projects/database-audio/default.htm

[110] T. Zhang. "Hierarchical system for content-based audio classification and retrieval." University of Southern California, 1998.
http://viola.usc.edu/
extranet/Projects/database-audio/default.htm