UNIVERSITÀ DEGLI STUDI DI VERONA

Dipartimento di Informatica

Corso di Laurea in Bioinformatica

TESI DI LAUREA

Analisi esplorativa di *soundscapes* con approcci di *Pattern Recognition* e *Machine Learning*

Relatore:
Manuele BICEGO

Candidato:
Giambattista Pomari

Anno Accademico 2024/2025

Indice

Sommario

Il seguente studio si colloca nell'ambito della *Pattern Recognition* e del *Machine Learning*, in particolare alla sua applicazione nell'analisi di *soundscape*. Per *soundscape* si intende l'ambiente sonoro composto da suoni naturali e artificiali, relativi ad uno specifico luogo geografico.

L'analisi si propone di caratterizzare *soundscapes* mediante tecniche di classificazione, i.e. metodi in grado di assegnare un oggetto ad un insieme predefinito di categorie.

E' stata implementata una *pipeline* di *Pattern Recognition*. Il segnale audio è stato caratterizzato attraverso classiche tecniche d'estrazione delle *features* - o caratteristiche (ad esempio *features* legate al contenuto spettrale, alla forma, o al timbro). Mediante tecniche di classificazione si è provato a misurare la capacità discriminativa di queste caratteristiche per classificare diverse categorie, come ad esempio il giorno dalla notte oppure la presenza o assenza di un temporale.

Sulla base di questo studio, in una fase successiva si è effettuata un'analisi esplorativa mediante algoritmi di *anomaly detection* (algoritmi in grado di evidenziare pattern anomali) per inferire informazioni e analizzare eventuali singolari pattern emersi.

La pipeline proposta è stata testata su dati raccolti in ambito di un monitoraggio acustico nella *Riserva Naturale Los Yátaros*, nel dipartimento di *Boyacá* in *Colombia*. La riserva presenta una biodiversità acustica molto particolare.

I risultati ottenuti sono incoraggianti e il loro contributo potrà migliorare la conoscenza nello studio acustico degli ecosistemi.

Capitolo 1

Introduzione

In questo capitolo saranno introdotte le nozioni fondamentali per comprendere l'ambito su cui si è sviluppato questo studio. Saranno definiti i concetti relativi a *Pattern Recognition* e *Machine Learning*, al loro utilizzo nell'analisi di *soundscape*, e in generale nell'analisi di audio nel campo dell'ecologia. Nel paragrafo successivo, si andrà ad esplicitare il fine che ha suggerito lo sviluppo di questo studio.

1.1 *Pattern Recognition* e *Machine Learning* per l'analisi di soundscapes e in generale di audio per l'ecologia

Pattern recognition (PR) e *Machine learning* (ML) [1] rappresentano una branca fondamentale dell'intelligenza artificiale, in particolare un insieme di tecniche utilizzate per estrarre informazioni dai dati tramite il riconoscimento automatico di specifici schemi, definiti *pattern*. In modo approssimato, si può dire che sono equivalenti, poiché condividono obiettivi, strumenti e approcci.

Il loro impiego è noto in molteplici ambiti: dal riconoscimento vocale o di immagini, all'elaborazione del linguaggio naturale, dai sistemi di raccomandazione, al monitoraggio in tempo reale e molti altri [1]. Tra questi emerge un contesto poco analizzato, che in letteratura si presenta come una sfida ancora aperta: l'analisi di soundscape. Prima di esaminare nel dettaglio come i *soundscape* sono stati affrontati nella PR/ML, per chiarezza, si desidera spiegare cosa si intende con tale definizione, le motivazioni per cui merita attenzione e le varie problematiche annesse.

Quinn *et al.* identificano i *soundscape* come una "particolare combinazione di suoni in un paesaggio" considerandola come "una caratterizzazione ecologica dei paesaggi" [2]. Gli autori ritengono che la composizione di un soundscape si divide in

quattro elementi principali: l'antropofonia (ANT: indica l'attività antropogenica), la biofonia (BIO: intesa come le vocalizzazioni della fauna selvatica), la geofonia (GEO: descrive i suoni dei fenomeni meteorologici) [2] e infine la quiete (indicata come il suono dell'ambiente).

A tal proposito, è molto interessante la caratterizzazione fornita da Farina *et al.* [3]. L'argomentazione descritta espone una visione alternativa più mirata e strutturata: separa il concetto di *sonoscape* da *soundscape*. Con *sonoscape* intende "il mosaico di tutte le non interpretate informazioni sonore all'interno di un landscape" [3]. Da questa definizione si deduce per esclusione l'interpretazione che l'autore attribuisce al *soundscape*, ossia "un *sonoscape* che è stato cognitivamente interpretato in un mosaico di categorie di ANT, BIO e GEO semioticamente interpretate da un organismo" [3]. Un'ulteriore suddivisione separa gli elementi in quelle che definisce unità sonore, i *sonotope* per gli *sonoscape*, e i *soundtope* per i *soundscape*. I primi vengono definiti dall'autore come una *patch* spazialmente unica di suoni non interpretati, mentre i secondi come suoni di ANT, BIO e GEO semioticamente interpretati da un organismo [3]. Rispetto ad un'umana suddivisione in ANT/BIO/GEO [3], questi concetti appena espressi consentirebbero una classificazione con maggiore dettaglio e specificità. Ciononostante, i termini sonotope e soundtope sono tuttora relegati a mere speculazioni a causa di una scarsità di evidenze empiriche [3].

Sebbene la definizione di *soundscape* possa risultare complessa, molto chiara è invece la sua importanza. Il ruolo che ricopre nell'ambiente naturale rappresenta un segnale della salute dell'ecosistema. Tale segnale può essere utilizzato per studi ecologici [3], diviene un significativo approfondimento della biodiversità e dell'impatto umano [2], può evidenziare cambiamenti negli habitat dove la qualità acustica è fondamentale per la dimensione vitale e il rumore umano risulta deleterio sulla biodiversità [2].

I vantaggi appena descritti supportano e incoraggiano l'analisi degli *soundscape*. PR/ML possono dare un grande contributo in tale processo. Sviluppare sistemi automatici mediante tecniche di PR/ML permetterebbe di supportare le sfide riguardanti l'analisi dei dati e il monitoraggio in tempo reale dell'ecosistema. La classificazione di *soundscape* consente l'identificazione automatica di suoni indesiderati su grandi quantità di dati, inoltre permette di modellare gli effetti e le interazioni di suoni diversi, e utilizzare poi tali modelli per identificare pattern spazio-temporali nell'attività sonora [2]. Introdurre un sistema euristico in grado di monitorare la presenza o l'abbondanza di particolari specie potrebbe aiutare la gestione di tale specie in una determinata zona, o addirittura evitarne l'estinzione. Allo stesso modo, può essere utile per prevenire situazioni di pericolo come il bracconaggio. I dati ricavati sarebbero fonte di studio per molti comportamenti animali in specifici periodi dell'anno, come il corteggiamento.

Questa innovazione nell'analisi degli *soundscape* non è priva di problematiche. Pochi studi di ecoacustica hanno provato a classificare soundscapes utilizzando intere categorie sonore come ANT/BIO/GEO e quiete [2]. Tale difficoltà si sviluppa su due elementi. In primo piano, l'identificazione manuale delle sorgenti sonore è altamente dispendiosa in termini di tempo [2]. Ciò è dovuto all'enorme quantità di dati da visionare manualmente che servono a censire un *dataset* di addestramento per i sistemi di PR/ML. Tanto più il *dataset* risulta ampio e dettagliato, maggiore sarà la qualità del sistema sviluppato. Oltre al tempo impiegato si possono sottintendere anche i costi di tale opera. Il secondo punto riguarda le competenze specifiche del settore. Infatti, per censire i dati di addestramento è richiesta una conoscenza della vocalizzazione degli animali del contesto, determinando per necessità la scelta di sviluppare *dataset* di addestramento di piccole dimensioni [2]. Tali *dataset* non riescono a spiegare nel complesso il problema, limitando così la qualità dei sistemi realizzabili.

1.2 Obiettivo della tesi: classificazione e analisi preliminare di *anomaly detection*

Il seguente studio si propone di studiare approcci di PR/ML per la caratterizzazione di un soundscape.

E' stata implementata una *pipeline* di *Pattern Recognition*. Il segnale audio è stato caratterizzato attraverso classiche tecniche d'estrazione delle *features* - o caratteristiche (ad esempio features legate al contenuto spettrale, alla forma, o al timbro). Mediante tecniche di classificazione si è provato a misurare la capacità discriminativa di queste caratteristiche per classificare diverse categorie, come ad esempio il giorno dalla notte oppure la presenza o assenza di un temporale.

Sulla base di questo studio, in una fase successiva si è effettuata un'analisi esplorativa mediante algoritmi di *anomaly detection* (algoritmi in grado di evidenziare pattern anomali) per inferire informazioni e analizzare eventuali singolari pattern emersi.

Capitolo 2

Background

L'obiettivo di questo capitolo è di fornire le conoscenze necessarie per poter comprendere l'analisi affrontata nei capitoli successivi. Esse comprendono l'introduzione all'analisi di un segnale (con relativo concetto di spettrogramma), la sua rappresentazione mediante le caratteristiche selezionate, la standardizzazione dei dati, la classificazione ed infine l'*anomaly detection*.

2.1 Dall'analisi del suono allo spettrogramma

Il suono nasce dalla vibrazione o oscillazione di un corpo sonoro. Queste vibrazioni creano delle onde sonore, cioè variazioni di pressione del mezzo che le propaga, per esempio l'aria.

L'onda sonora è definita da tre caratteristiche: l'ampiezza, la frequenza [4] e il timbro [5].

L'ampiezza (o intensità) dell'onda è associata a quanto il suono è percepito intenso (il volume) ed è misurata in *Decibel*.

La frequenza (o altezza) identifica il numero di oscillazioni in un secondo, esprime un valore minore o maggiore in base a che il suono risulti più grave o più acuto, determinando così il tono, ed è calcolata in *Hertz*. La frequenza più bassa è definita come fondamentale e corrisponde al tono percepito del suono. Oltre alla fondamentale, ogni suono complesso contiene armoniche, che sono frequenze multiple della fondamentale. L'insieme di queste frequenze definisce lo spettro del suono che caratterizza un onda sonora.

Il timbro, infine, è la qualità percepita del suono che ci permette di distinguere due strumenti musicali che stanno eseguendo la stessa nota (quindi stessa frequenza) alla stessa intensità (stessa ampiezza). Il timbro è influenzato dalla forma dell'onda sonora e dallo spettro.

Nel quotidiano utilizzo del mondo digitale, è comune visualizzare un segnale audio come un'onda, senza sapere che questa prospettiva rappresenta graficamente l'andamento dell'ampiezza (sull'asse delle ordinate) in funzione del tempo (l'asse delle ascisse). Questo tracciato è il risultato di un processo, il campionamento, effettuato sul segnale analogico, ovvero sulla forma originale del suono rilevato, che trasforma i campioni, ad intervalli regolari, in segnale digitale. Poiché la struttura digitale non è in grado di cogliere il segnale originale nella sua forma continua, nella sua reale interezza, il campionamento registra in forma discreta, ovvero traccia un valore numerico, discreto, a intervalli regolari. Maggiore è la frequenza di campionamento quindi, il numero di campioni analizzati per ogni secondo, maggiore è la qualità del risultato. Gli intervalli di campionamento sono come delle istantanee che misurano e registrano digitalmente il valore dell'ampiezza del segnale in precisi istanti di tempo.

Il segnale campionato viene rappresentato nel *dominio del tempo*, cioè lo spazio che misura la variazione dell'ampiezza rispetto al tempo. Questo punto di vista è molto utile per evidenziare la durata dei suoni, la durata delle pause e la struttura temporale del segnale. Tuttavia, per poter analizzare nel dettaglio le componenti frequenziali, si deve effettuare un cambio di prospettiva. Si applica la *Trasformata di Fourier* [7], una funzione in grado di suddividere l'onda complessa nelle sue sottocomponenti sinusoidali, permettendo quindi di visualizzare il segnale nel *dominio delle frequenze*, dove l'ampiezza è in rapporto con le frequenze.

Se invece di eseguire la trasformata sull'intero segnale, lo si suddivide in blocchi, o finestre temporali, e su ognuna si applica separatamente la funzione, si ottengono più spettri di frequenza, uno per ogni intervallo. Questi spettri, combinati in un'unica rappresentazione, formano lo spettrogramma, un grafico che mostra l'andamento delle frequenze in funzione del tempo. Tale prospettiva ci permette di cogliere in combinazione le informazioni temporali e frequenziali. In questa processo è importante la dimensione della finestra temporale e del passo, che definisce quanto si devono sovrapporre le finestre consecutive. Il numero di campioni analizzati per gruppo determina la dimensione della finestra. Maggiore è il numero di campioni considerati, minore sarà il numero di finestre utilizzate nel calcolo dello spettrogramma. Una finestra maggiore determina una migliore risoluzione delle frequenze, ma una peggiore risoluzione temporale. Il passo, tipicamente, viene impostato ad un valore uguale alla metà del numero di campioni utilizzati per la finestra.

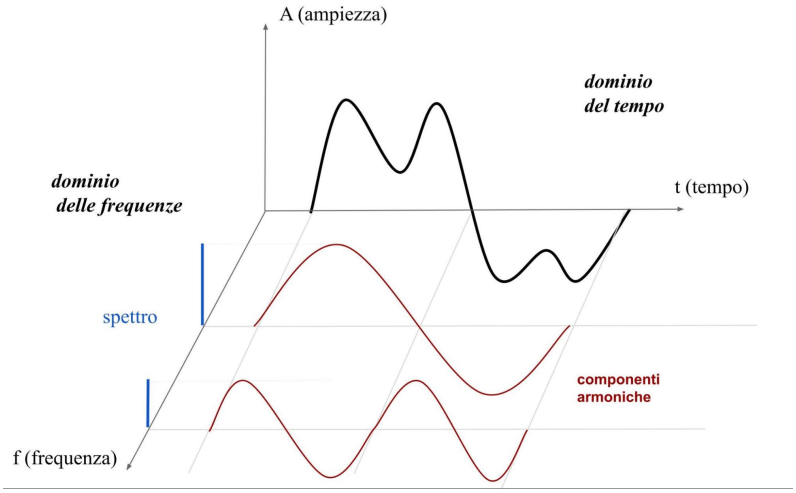

Figura 2.1: Relazione tra dominio del tempo e dominio delle frequenze. In alto a destra la forma d'onda (in nero) rappresentata nel dominio del tempo in rapporto al tempo e all'ampiezza. Sotto le varie componenti armoniche dell'onda (forme ondulate in rosso). A sinistra viene rappresentato il dominio delle frequenze che presenta le ampiezze delle frequenze armoniche di cui è composta l'onda.

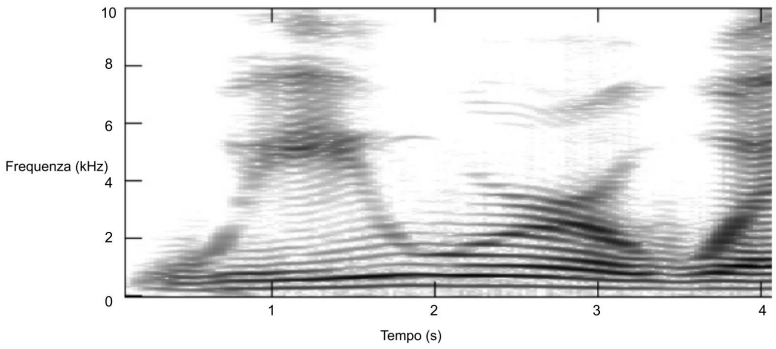

Figura 2.2: Spettrogramma di un file audio. La scala di colori va dal bianco, che indica un'intensità sonora bassa, al nero, che rappresenta un'intensità sonora alta

2.2 Estrazione delle *features*

Distinguere due oggetti qualsiasi, come una bottiglia e una mela, può sembrare una capacità comune, per nulla speciale. Questa abilità è frutto di un meccanismo che il nostro cervello sviluppa attraverso l'esperienza e la conoscenza. Per ogni oggetto con cui interagiamo, la mente elabora un insieme di caratteristiche in grado di descriverlo e lo esegue con una tale velocità che nemmeno ce ne accorgiamo. Il cervello estrae elementi in grado di definire l'oggetto, come il colore, la lunghezza e la forma, e con ogni senso del corpo. L'oggetto è da intendersi anche come un profumo, un suono, un'immagine o qualsiasi altra percezione.

Al fine di insegnare questa capacità ad una macchina, è necessario identificare ciò che è rilevante, discriminante e misurabile nei dati: le caratteristiche, o *features*. Le *features* forniscono le informazioni necessarie per costruire il modello in grado di individuare i *pattern* nei dati e generalizzare, ovvero la capacità di riconoscere anche oggetti mai visti.

In questo studio, che tratta di *soundscape*, l'oggetto da analizzare è un segnale audio. Per caratterizzare tale segnale sono state utilizzate delle classiche misure di *signal processing*, che si basano sui concetti illustrati nel paragrafo precedente.

Si possono distinguere tre gruppi principali di *features*: spettrali (SPE), tonali (TON) e temporali (TEM) [7].

Le SPE caratterizzano la forma dello spettro e influenzano le percezione del timbro. Tali features vengono calcolate sullo spettrogramma del segnale. Si suddividono in:

- *Spectral Centroid*: consiste nella media pesata delle frequenze nel segnale e indica il centroide, ovvero il centro di massa dello spettro. Valori più elevati indicano un suono più brillante [7]. Per brillante si intende che la maggioranza delle armoniche si trova su alte frequenze.

- *Spectral Spread*: misura la dispersione delle frequenze attorno al centroide [7]. Un basso valore indica una concentrazione maggiore delle frequenze attorno al centroide.

- *Spectral Rolloff*: misura la frequenza al di sotto della quale si trova una percentuale specifica dell'energia totale dello spettro. Valori bassi indicano una scarsa presenza di componenti ad alte frequenze [7].

- *Spectral Decrease*: misura quanto l'energia spettrale cala rapidamente all'aumentare delle frequenze. Un curva ripida indica una diminuzione rapida dell'energia spettrale, quindi un blocco ricco di basse frequenze e povero di alte frequenze [7].

- *Spectral Flux*: rileva il numero di cambiamenti nella forma dello spettro. Identifica variazioni rapide e significative nel contenuto del segnale [7].

Le TON misurano le componenti tonali del segnale rispetto al rumore. Tali feature vengono calcolate sullo spettrogramma del segnale. Sono composte da:

- *Spectral Crest Factor*: descrive il rapporto tra il valore massimo delle magnitudini dello spettro e la somma di tutte le magnitudini dello spettro (la magnitudo si riferisce all'ampiezza massima raggiunta da ogni frequenza nello spettro del segnale). Un valore basso indica un segnale molto uniforme [7].

- *Spectral Flatness*: indica quanto lo spettro è uniforme. Un valore alto suggerisce un segnale con poca struttura tonale, quindi molti rumori.

- *Spectral Tonal Power Ratio*: rapporta l'energia tonale con l'energia totale. Un valore alto indica che l'energia si concentra in componenti tonali, basso sui rumori [7].

Infine le TEM, che descrivono come il segnale varia rispetto al tempo. Tali features vengono calcolate sul segnale nel dominio del tempo. Si suddividono in:

- *Time Zero Crossing Rate*: identifica il numero di volte in cui il segnale cambia di segno quindi quando ha valore zero. Un valore alto indica una forte presenza di alte frequenze [7].

- *Time Acf Coeff*: quantifica la correlazione tra il segnale e una versione ritardata dello stesso (funzione di autocorrelazione). Questa misura è utile per identificare pattern ripetitivi [7].

- *Time Max Acf*: indica il valore massimo dell'autocorrelazione. Un valore alto può esprimere una forte periodicità del segnale [7].

2.3 Standardizzazione dei dati

La standardizzazione è un'attività di pre-processamento dei dati in grado di trasformarli in una forma indipendente dalla scala utilizzata. Per scala si intende l'intervallo in cui le varie features vivono, ed è fondamentale per confrontare i dati. Infatti, una certa misurazione può avere un rapporto diverso con gli altri dati a seconda della scala. Immaginiamo di avere i risultati di due esami scolastici diversi, fatti da due gruppi di studenti. Il primo gruppo ha ottenuto i risultati in centesimi, un intervallo da 1 a 100, invece il secondo gruppo, in trentesimi, da 1 a 30. Un voto di 30 nel primo gruppo è molto diverso da un voto 30 nel secondo gruppo: senza

standardizzare la scala il confronto è falsato. Si deve quindi riportare alla stessa scala. La standardizzazione è un insieme di tecniche dove vengono uniformate le versioni ottenendo una versione standard dei dati, "senza dimensionalità".

Una tecnica di standardizzazione molto utilizzata è lo *Z-score standardization*, qui specificata con la formula:

$$x^*_{ji} = \frac{x_{ji} - \overline{x}_j}{\sigma_j} \tag{2.1}$$

Si definisce x^*_{ji} la j-esima *feature* dell'oggetto i standardizzata, x_{ji} la j-esima feature dell'oggetto i prima della standardizzazione, $\overline{x}_j$ la media della *feature* j ed infine σ_j la deviazione standard delle *features*. Si consideri ora una matrice composta sulle righe dagli oggetti in esame e in colonna i valori delle *feature*. Per ogni elemento x in riga j e posizione i viene sottratta la media calcolata in colonna j, e il valore ottenuto si divide con la deviazione standard estratta dalla colonna j. Dopo la standardizzazione ogni feature ha media uguale a 0 e deviazione standard a 1.

2.4 Classificazione

La classificazione è uno dei *task* più utilizzati, affrontato con tecniche di PR e ML. Un classificatore rappresenta un sistema decisionale in grado di assegnare una categoria, o etichetta, ad un oggetto sulla base di un modello, tipicamente a partire da una descrizione vettoriale (vettore di *features*). In sostanza, è una funzione che prende in input un oggetto, ne elabora le *features* e restituisce un valore discreto che determina a quale categoria appartiene.

In generale gli approcci alla classificazione si suddividono in generativi e discriminativi.

Nell'approccio generativo si mira a definire un modello per ogni categoria, o classe. Questa tipologia di approcci presenta una struttura più flessibile in grado di adattarsi a nuove classi, è più rapida nell'addestramento e ottiene una migliore capacità descrittiva per la singola classe.

L'approccio discriminativo, invece, si basa sulla ricerca del migliore confine decisionale per separare le classi nello spazio. Per sua natura è più rapido, soprattutto nella fase di test, e presenta un'efficacia di classificazione migliore dato che il sistema viene costruito specificatamente per risolvere il problema di classificazione.

Un'ulteriore suddivisione nei classificatori si basa sulla loro natura parametrica o non parametrica. La parametrica si caratterizza per l'assunzione della forma di distribuzione dei dati per ogni classe (es. la distribuzione normale), e si concentra nella stima dei parametri della funzione che genera tale distribuzione. Diversamente, la non parametrica non assume nessuna forma di distribuzione, ma viene stimata

direttamente dal *training set*. Risulta più dispendiosa in termini computazionali ma non basandosi su assunzioni determina un modello maggiormente flessibile e adattabile al contesto.

2.4.1 Apprendimento supervisionato

Nella PR spesso si utilizza il paradigma dell'*apprendimento da esempi*. Questo metodo può essere visto come l'apprendimento di un bambino che sperimenta e acquisisce conoscenza da esempi, dall'esperienza. In particolare, nel contesto della classificazione, si utilizza un approccio supervisionato in cui la conoscenza viene acquisita tramite dati campionati dal problema, il *training set*, dotato di categorie, o etichette, note. Conoscendo la reale classe di appartenenza degli oggetti, il modello può addestrarsi e migliorare gradualmente la sua capacità di classificazione. In questo processo è importante che il sistema non "impari a memoria" i dati del *training set*, il cosiddetto *overfitting*, che comporta un eccessivo adattamento ai dati di addestramento. L'obiettivo infatti è di creare un modello in grado di generalizzare quindi di classificare correttamente anche oggetti sconosciuti, mai visti.

2.4.2 Validazione

Una volta costruito il modello è necessario verificare la qualità del classificatore. Per tale scopo si utilizza il *testing set*, un *dataset* che presenta elementi diversi da quelli usati in addestramento, ma dotato di categorie note da confrontare con il risultato predetto dal classificatore. Il modello classifica tali dati e si valuta la predizione in base all'errore ottenuto. Questo valore pone in rapporto le previsioni errate rispetto al numero totale di oggetti analizzati. Una previsione errata consiste in una falsa valutazione del classificatore, quindi un valore diverso dalla reale categoria di appartenenza.

Nella costruzione di un classificatore di solito si dispone di un unico *dataset* che si deve suddividere in due parti, il *training set*, per l'addestramento, e il *testing set* per i test e la validazione. Un metodo ideale sarebbe poter usare tutti dati di esempio per il *training* ed estrarre altri esempi dal problema per testare il modello, ma nella realtà potrebbe essere non fattibile o troppo dispendioso. Si preferisce quindi usare una metodica comune che consiste nella *Cross Validation*, che permette di ottenere una valutazione più valida e consistente. Esistono diverse varianti, ognuna con le sue caratteristiche. La forma più semplice è la *Holdout*, che distribuisce casualmente i dati in due insiemi di uguale dimensione. Un'alternativa simile è l'*Average Holdout*, che per essere indipendente dalle partizioni effettua più *holdout* e calcola l'errore come media dei risultati ottenuti in tutti i casi.

Infine, una delle più utilizzate è la *Leave One Out* (LOO), una variante particolare che ottiene ottimi risultati in termini di affidabilità, soprattutto con dataset ristretti. Come suggerisce il nome, consiste nell'effettuare l'addestramento con tutti gli oggetti del *dataset* meno uno, x_i, che viene invece utilizzato per validare il modello. Si ripete il procedimento lasciando fuori come oggetto di testing un diverso elemento x_i del *dataset*, e al termine si media il risultato ottenuto. Presenta un costo computazionale maggiore ma garantisce indipendenza dalla partizione e dai dati scelti del *training set* e *testing set*.

2.4.3 Metodo di classificazione

In questo studio è stato utilizzato il classificatore *K Nearest Neighbor* (KNN), un approccio supervisionato generativo non parametrico, semplice e intuitivo: il metodo si basa sul classificare un punto, un oggetto, assegnandogli la classe che più frequentemente ritroviamo tra i k oggetti più vicini. Il concetto di vicinanza si concretizza con la scelta della distanza: nel nostro caso la *distanza euclidea*, una delle misure più utilizzate. Come risulta chiaro, la scelta del valore di k è cruciale.

2.5 Anomaly Detection

L'*anomaly detection* (AD) consiste nell'identificare fenomeni ed eventi che presentano un comportamento anomalo rispetto al resto del *dataset* [8]. Tali fenomeni, denominati *outlier*, si discostano in modo significativo dagli *inlier*, il resto dei dati normali, per la loro natura anomala e la scarsa numerosità.

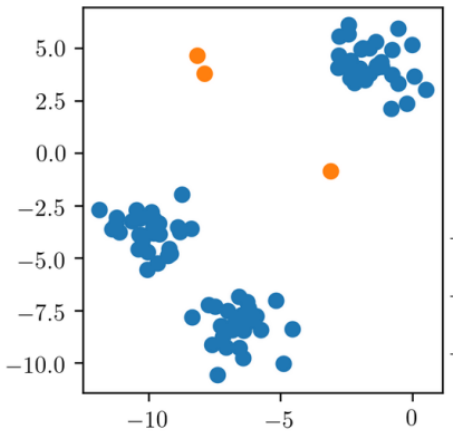

Figura 2.3: Rappresentazione di *inliers* (cerchi colore blu) e outliers (cerchi colore giallo) in un *dataset*. Si nota come gli *outlier*, in questo caso, risultano distaccati dal comportamento generale definito dagli inlier, che sono raggruppati e seguono una distribuzione più omogenea.

2.5.1 Gli *outlier*

Individuare gli *outlier*, pur essendo complicato, è una necessità. Ciò consentirebbe di inferire informazioni preziose oppure rappresentare una prevenzione per eventuali criticità. Contribuisce in modo significativo anche al data cleaning, la "pulizia dei dati", che conta di un insieme di processi che servono a rimuovere duplicati, uniformare e filtrare i dati. In questo contesto rimuovere gli outlier semplificherebbe le fasi successive di analisi migliorando la qualità del risultato.

Si distinguono varie tipologie di *outlier*:

- i *point*, che sono identificati tali indipendentemente se si trovano da soli o in gruppo;
- i *collective*, considerati outlier solo se rilevati in gruppo, altrimenti rientrano negli *inlier*;
- i *contextual*, che sono rilevati normali o anomali in base al contesto specifico in cui si trovano. Per esempio, in un sistema di monitoraggio della temperatura di una città, una valore di 30 gradi durante l'estate verrà considerato normale, diversamente lo stesso valore in inverno sarà definito anomalo.

2.5.2 Le applicazioni

L'AD ricopre un ruolo importante in molteplici campi. Per esempio, possiamo citare:

- le intrusioni di rete: in un rete informatica di sistemi che condividono informazioni monitorare attività sospette o non autorizzate è fondamentale. Tali attività, come un programma o un individuo malevole, potrebbero insinuarsi con lo scopo di rubare dati o compromettere la sicurezza. Sistemi basati su AD consentono di rilevare tali comportamenti come anomali poiché si discostano dal traffico della rete considerato normale;
- l'ambito sanitario: fondamentale nell'analisi dei dati dei pazienti per diverse ragioni come condizioni di salute anomale, errori nella strumentazione o nelle registrazioni mediche. Tale approccio potrebbe contribuire in modo significativo nel rilevamento di situazioni potenzialmente critiche, sia per intervenire in anticipo che per evitare errori nella diagnosi;
- i sistemi automatizzati, in cui prevenire un avaria o, riuscire a intervenire in tempo nel caso si manifestasse, è essenziale;
- nel processamento di immagini o testi, come rilevare *fake news*;

- nel rilevamento di frodi, all'interno delle innumerevoli transazioni bancarie prodotte ogni giorno.

Queste casistiche sono accomunate da un enorme quantità di dati dove sistemi troppo rigidi e specifici non potrebbero adattarsi al continuo mutamento delle variabili in gioco. Le problematiche nell'AD sono svariate. I contesti affrontati non sono supervisionati, non vi è modo di basarsi su esempi espliciti di *outliers* dato che non esiste una chiara definizione di ciò che rende un'anomalia tale e il rischio di determinare falsi positivi è molto alto.

2.5.3 I metodi

I diversi approcci di AD tipicamente si suddividono sulla base del metodo utilizzato per rilevare le anomalie. I vari metodi possono essere:

- Metodi basati sui concetti statistici: essi ricercano elementi che non rispecchiano la distribuzione dei dati. Una bassa probabilità di appartenenza alla distribuzione determina un'alta probabilità di essere un *outlier*.

- Metodi basati sul *clustering*: essi suddividono i dati per somiglianza in gruppi, i cluster, e le anomalie risultano evidenti poiché molto diverse dal loro gruppo di appartenenza. Oppure gli *outlier* formano piccoli cluster che presentano una dimensione o una densità inferiori alla soglia necessaria per essere considerati tra i dati normali, venendo quindi identificati come anomalie.

- Metodi basati sull'apprendimento: vengono utilizzati metodi di apprendimento automatico per ricercare *pattern* all'interno dei dati. Gli elementi che non si identificano in questi *pattern* vengono considerati come anomalie.

- Metodi basati sulla distanza o sulla densità: viene considerata la distanza tra gli elementi, o la densità locale. Nel primo caso gli *outlier* si troveranno distanti dagli altri punti, nel secondo saranno in zone a bassa densità.

- Metodi basati sulla combinazione di vari metodi, o *ensemble*: queste tipologie combinano i risultati di metodi diversi, o gli stessi con parametri differenti, per ottenere una previsione più accurata.

La maggioranza degli algoritmi ritorna come risultato un valore, denominato *Anomaly score*, che quantifica quanto un elemento è probabile che sia un'anomalia.

In letteratura esistono differenti algoritmi di AD [8], ma di seguito ne saranno descritti solo tre, quelli utilizzati nello studio. I metodi sono L'*IForest* (IF) [9], o *Isolation forest*, il *Local Outlier Factor* (LOF) [10] e l'*Ocsvm* (OCSVM) [11], o *One Class Support Vector Machine*.

IF è un algoritmo basato su un *ensemble* [8] di alberi decisionali. Per albero decisionale si intende una struttura ad albero dove ogni nodo contiene un test e i rami le possibili risposte. Procedendo dall'alto verso il basso del modello, i dati vengono indirizzati dalle varie risposte fino alle foglie, lungo il path. In IF viene costruita una foresta di alberi decisionali e ciascun albero cerca di isolare i dati mediante suddivisione. L'idea è che le anomalie avranno un path più corto poiché avranno bisogno di meno divisioni rispetto ai dati normali. In sostanza, più risulta facile isolare un oggetto e con maggiore probabilità sarà un anomalia. Questo algoritmo risulta molto scalabile, veloce e accurato, specialmente su dataset di grandi dimensioni.

LOF è un metodo basato sulla densità. La densità di un oggetto può essere calcolata guardando al suo vicinato, ovvero dalla numerosità degli elementi che gli sono vicini. Il criterio utilizzato per determinare la vicinanza ad un oggetto dipende dalla metrica scelta nell'implementazione. La più comune è la distanza euclidea, che misura la lunghezza del segmento tracciato tra due punti. In sostanza, l'algoritmo stabilisce che minore è la densità locale allora più alta è la probabilità che un determinato oggetto possa essere un'anomalia. Questo metodo è efficace con dati ad alta dimensionalità, ma al contrario, risulta molto dispendioso in termini computazionali.

Infine, OCSVM si basa sull'apprendimento. Si tratta di una variante delle *Support Vector Machine*, un approccio discriminativo applicato solitamente a problemi binari. In questa forma, a singola classe, l'algoritmo viene addestrato per definire un confine che racchiude al suo interno solo i dati normali. Così facendo i dati anomali, che risultano esterni alla distribuzione, emergono e sono quindi rilevabili. Spesso viene utilizzato il trucco del *kernel*, che consiste nel proiettare i dati in una dimensione superiore, dove può risultare più facile separare i dati.

Capitolo 3

Dataset

Lo studio condotto si è basato sui dati raccolti nell'ambito di un monitoraggio acustico passivo nella *Riserva Naturale Los Yátaros*, nel dipartimento di *Boyacá* in *Colombia* [3]. Con il termine passivo si identifica una modalità di osservazione del paesaggio incentrata sulla registrazione di un particolare luogo e solo successivamente prevede un'analisi approfondita, diversamente da quella attiva, dove si osserva e si analizza il fenomeno in tempo reale. La raccolta dati è stata commissionata dalla fondazione *Von Humboldt*, un ente colombiano che si occupa di ricerca sulla biodiversità e sulle sue relazioni con il benessere umano.

La riserva è composta da querceti e foresta subandina in diversi stadi di rigenerazione naturale, e presenta una biodiversità acustica molto particolare. Il progetto mirava a profilare l'impronta acustica della riserva campionando suoni nello spettro udibile e negli ultrasuoni.

Sono stati predisposti tre siti, denominati *YAT*, organizzati in una disposizione triangolare, lungo il sentiero principale, distanti 150 m, con due sensori acustici *AudioMoth* ciascuno, per le due forme di suono desiderate, posti ad altezza diverse. Il periodo di campionamento si è svolto dall'1 marzo al 2 maggio 2020, registrando un minuto di audio ad intervalli di trenta minuti durante tutto il giorno (dalle 00:00 alle 23:30) per lo spettro dell'udibile (0 *Hz*-16 *kHz*) e nella fascia notturna (dalle 16:30 alle 6:00) per lo spettro dell'ultrasuono (fino a 192 *kHz*). In totale si è ottenuto 12447 registrazioni di cui 9055 nell'udibile e 3392 nell'ultrasuono. In questo progetto, sono stati considerati solo i dati nello spettro udibile, per permettere l'ascolto del contenuto.

3.1 Dataset prima fase

Nella prima fase dello studio è stato utilizzato il *dataset* completo (DATA1) che presenta l'insieme originale dei dati. Il gruppo si presenta con una suddivisione per

i tre siti (YAT1, YAT2, YAT3) con quantità leggermente differenti. Gli audio sono 3018 per *yat*, a parte il primo con 3019. Ogni sito presenta 1482 file per il mese di marzo (1483 solo per YAT1), 1440 per il mese di aprile e 96 per il mese di maggio.

Data l'ingente quantità di dati disponibile, non si è potuto analizzare il contenuto, ossia ascoltare l'intero insieme di registrazioni. In una prima caratterizzazione, analizzando diversi audio in momenti diversi della giornata e del mese, si è osservato che tutti e tre i luoghi risultano molto caratterizzati dal suono del fiume e della cascata vicina. È importante notare che anche se tale suono fosse stato ad una distanza maggiore avrebbe sortito lo stesso effetto: infatti Farina *et al.* sostengono che "la geofonia può essere rilevata anche a grandi distanze in base all'ampiezza della sorgente sonora" [3]. A questo si aggiungono ulteriori problematiche dovute a periodi piovosi, il cui rumore sovrasta in diverse occasioni i suoni ambientali naturali. Entrambi gli elementi appena descritti determinano un ambiente umido che potrebbe influire anche sulla capacità del sensore.

3.2 Dataset seconda e terza fase

Il *dataset* della seconda e terza fase dello studio (DATA2) consiste in un sottoinsieme del *dataset* DATA1. Per potere inferire maggiori informazioni dal contesto si è stabilito che era necessaria una descrizione più accurata del contenuto. Quindi si è ristretto l'insieme ad un campione di dati minore che potesse essere ascoltato e studiato nel dettaglio. L'analisi ha estratto dal *dataset* originale 186 audio, focalizzandosi sul sito YAT1, nel mese di marzo con finestre temporali a intervalli di due ore, partendo dalle due del mattino, quindi nelle ore: 02:00, 06.00, 10:00, 14:00, 18:00, 22:00. Con questa modalità si può ottenere una visione abbastanza generale della varietà sonora presente nella giornata, includendo i due momenti fondamentali di alba e tramonto, caratterizzati da picchi di attività acustica.

L'interpretazione manuale del *dataset* ha classificato il contenuto assegnando delle etichette ai vari elementi distinti utilizzando i gruppi delle categorie descritte nel *cap.1.1.* In relazione all'ANT è stato individuato un unico suono, appartenente al rumore di veicoli (classe V). Nella BIO è stato individuato il verso degli uccelli e dei grilli (classi U e G). Per la GEO si è rilevato il precedentemente menzionato rumore del fiume/cascata, la pioggia e i tuoni (classi C, P e T). Infine, sono stati identificati i rumori relativi alle interferenze del sensore (classe I), ed eventuali elementi uditi ma non interpretati (sconosciuti classe S) . Rispetto a quanto specificato nell'introduzione, all'interno della GEO si è integrato anche l'insieme dei suoni relativi alla quiete. Tale scelta è derivata da una maggiore semplicità nella trattazione, ma specialmente per l'impossibilità nel poterli classificare correttamente.

Nel grafico della figura 3.1 è possibile visualizzare la distribuzione degli elementi descritti nelle fasce analizzate. Da una prima osservazione risulta evidente la significativa presenza dell'elemento C, come già esplicitato nel paragrafo precedente. Lo stesso comportamento si può osservare al suono dell'elemento U, risultato meno attivo solo nella parte centrale della giornata. Il resto è relativamente distribuito, a parte I e S diffusi con bassa intensità. Si può notare come S sia presente solo nelle due fasce pomeridiane. Dal grafico della figura 3.2 possiamo avere una prospettiva alternativa della distribuzione di ogni suono su tutto il mese di marzo, per il quale è lecito esplicitare le medesime considerazioni annotate.

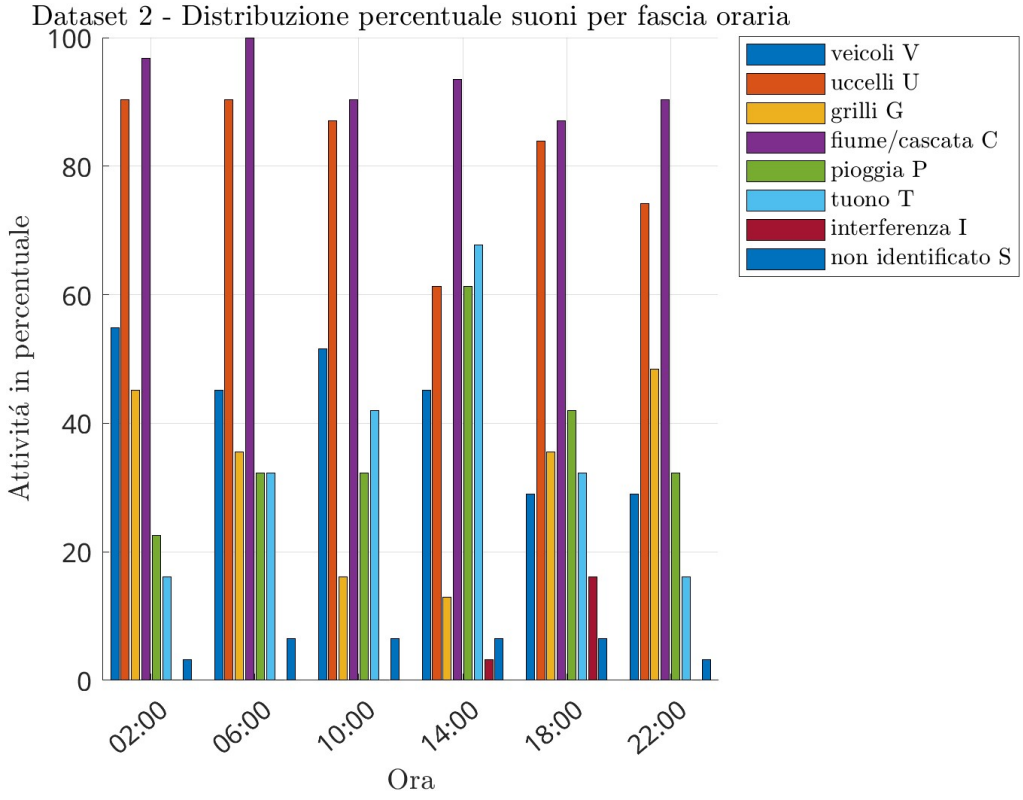

Figura 3.1: Esposizione della presenza sonora in percentuale per ogni suono nelle varie fasce orarie.

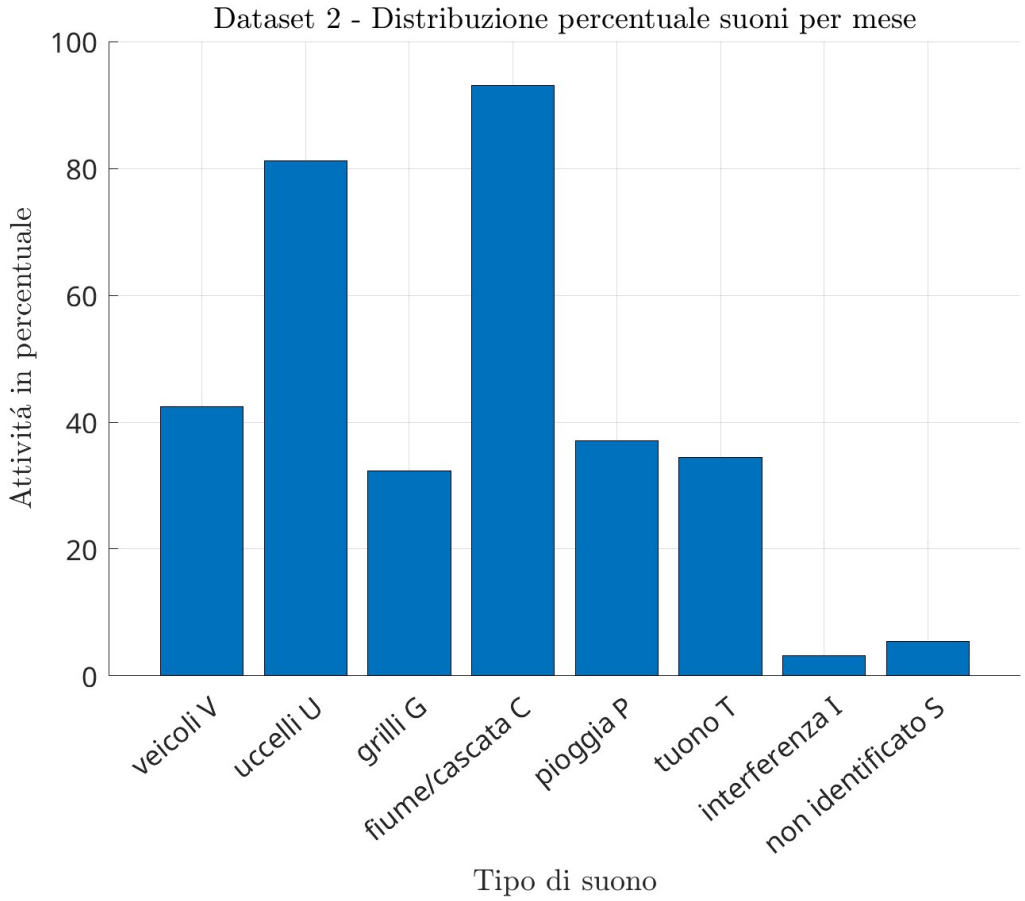

Figura 3.2: Esposizione della presenza sonora in percentuale per ogni suono in tutto il mese. Ogni suono presenta un percentuale di distribuzione su tutti gli audio analizzati e un audio può contenere più suoni.

Capitolo 4

Classificazione

In questo capitolo verranno descritti i risultati relativi alle prime due fasi dello studio basate sulla classificazione: la prima riguardante la classificazione su categorie oggettive, la seconda su categorie semantiche. Le categorie oggettive riguardano informazioni deducibili dai metadati dei file audio, come il luogo di registrazione o una fase della giornata. Le categorie semantiche invece si riferiscono a informazioni ricavate dall'analisi del contenuto dell'audio, ovvero ai vari suoni identificati presenti nella foresta.

Saranno definiti i gruppi di features utilizzati negli esperimenti, descrivendo poi la tipologia di classificatore scelto. A seguire, saranno illustrati nel dettaglio i problemi di classificazione disegnati, ed infine si andrà ad analizzare i risultati ottenuti nei due esperimenti. L'obiettivo è determinare la parametrizzazione migliore per l'approccio, quali features sono più efficaci al nostro contesto, e in quali problemi ottengono i risultati migliori. Questo studio considera gli esperimenti di classificazione precedentemente condotti dai colleghi Ilaria Ballerini e Andrea Piazza.

Inoltre, viene analizzata la possibilità di aggiungere una fase di filtraggio per alcune frequenze. L'idea consiste nel rimuovere alcune frequenze preponderanti su ogni audio, che rappresentano rumore, dovute in maggior parte alle problematiche intrinseche del contesto, ma anche alla sensibilità dello strumento di misurazione.

4.0.1 Configurazioni *features*

Nello studio sono state considerate sia le *features* nella loro forma originale, come descritta nel capitolo 2.2, sia aggregate. Le *features* originali sono descritte da un numero definito di componenti (che saranno indicate con N). Ogni feature originale è quindi descritta da un vettore riga di 1xN componenti. N dipende dai parametri usati per calcolare lo spettrogramma.

In particolare sono stati identificati 19 gruppi di *features*:

- 11 ORIGINALI (o ORIG), per ognuna delle 11 *features*, utilizzate singolarmente, 1xN elementi ciascuna;
- CONCATENAZIONE ORIGINALI (o CONC.ORIG.), formata dalla concatenazione orizzontale delle 11 *features* originali, ottenendo un totale di 11xN componenti;
- CONCATENAZIONE SPETTRALI (o CONC.SPE.), ottenuta dalla concatenazione delle componenti delle 5 *features* spettrali, quindi 5xN componenti;
- CONCATENAZIONE TONALI (o CONC.TON.), come la precedente ma considerando le 3 *features* della tonalità, ovvero 3xN componenti;
- CONCATENAZIONE TEMPORALI (o CONC.TEM.), come la precedente ma con le 3 *features* temporali, 3xN componenti;
- CONCATENAZIONE MEDIE ORIGINALI (o CONC.MED.ORIG.), formata dalla concatenazione orizzontale delle medie delle 11 *features* originali, quindi 11 componenti;
- CONCATENAZIONE MEDIE SPETTRALI (o CONC.MED.SPE.), ottenuta dalla concatenazione orizzontale delle medie delle componenti delle 5 *features* spettrali, quindi solo 5 componenti;
- CONCATENAZIONE MEDIE TONALI (o CONC.MED.TON.), come la precedente ma utilizzando le 3 *features* tonali, quindi solo 3 componenti;
- CONCATENAZIONE MEDIE TEMPORALI (o CONC.MED.TEM.), come la precedente ma utilizzando le 3 *features* temporali, in totale 3 componenti;

Ai 19 gruppi sopracitati si è tenuto conto anche della relativa versione standardizzata, ovvero ottenuta dalle componenti processate con la tecnica di standardizzazione *Z Score*, definendo quindi 38 gruppi: 19 originali e 19 standardizzati. In questo modo, si dispone anche di una rappresentazione con una scala comune indipendente dalle misurazioni originali.

In aggiunta, i dati sono stati estratti in due forme diverse (ottenendo quindi 78 gruppi di *features*) basate su diverse finestre temporali scelte per il campionamento nel calcolo dello spettrogramma: la prima, FS0X, usa una finestra di 32768 campioni, quindi meno di 1 secondo; la seconda, FS1, invece si basa su una finestra di 48000 campioni, corrispondente a 1 secondo.

Nel caso di FS1 sono state estratte 120 componenti (N=120): il segnale audio analizzato presenta una lunghezza temporale di 60 secondi e, per costruire lo spettrogramma lo si analizza con un intervallo di 1 secondo alla volta, ottenendo 60

finestre. Inoltre, considerando un passo pari alla metà dell'intervallo, 0.5 secondi, si ottengono altre 60 finestre, per un totale di 120 campionamenti. Alla stessa modo, è stato fatto per la configurazione FS0X dove, considerando una finestra più breve, si è ottenuto un maggior numero di componenti (N=176).

La tabella 4.1 riassume i vari gruppi di features considerati.

Capitolo 5

Anomaly Detection

Capitolo 6

Conclusione

Bibliografia

[1] Eco, Umberto (1977), *Come si fa una tesi di laurea*, Bompiani, Milano.