

UNIVERSITÀ DEGLI STUDI DI GENOVA

---

Scuola Politecnica

Corso di Laurea in Ingegneria Elettronica e Tecnologie dell'Informazione

Tesi di Laurea Triennale

# Estrazione di *feature* per classificazione di immagini telerilevate ad alta risoluzione mediante istogrammi di gradienti orientati

Feature extraction for high resolution remote sensing image classification using  
histograms of oriented gradients



**Relatore**

Prof. Gabriele Moser

**Laureandi**

Margherita PICCINI

**Correlatore:**

Dott. Vladimir Krylov

Simone ROSSI

Eugenio ZUCCARELLI

---

ANNO ACCADEMICO 2014-2015

*“Anyone who has never made a mistake has never tried anything new. ”*

Albert Einstein

# Sommario

La classificazione di immagini telerilevate è un metodo di analisi ed identificazione che permette di associare ad ogni pixel di una determinata immagine, una etichetta corrispondente ad una specifica classe e che permette ad un calcolatore di discriminare autonomamente le varie regioni costituenti l'immagine stessa. Le possibili applicazioni sono innumerevoli, ma questa tesi si focalizzerà su tecniche di telerilevamento di immagini ad alta risoluzione (VHRI), finalizzate all' osservazione della Terra.

Questa tesi esplora l'applicazione di una moderna tecnica di estrazione delle *feature*, chiamata istogrammi di gradiente orientati (HOG), applicata a immagini multispettrali VHR. L'algoritmo, già largamente utilizzato nell'ambito della *human detection*, ma totalmente innovativo in questo ambito, verrà analizzato dettagliatamente in ogni sua fase, evidenziando come a differenti configurazioni di parametri corrispondano variazioni nelle *performance*.

Sebbene alcune combinazioni di parametri abbiano mostrato una alta discrezionalità, gli esperimenti sono risultati largamente soddisfacenti, soprattutto in relazione alla carenza di materiale preesistente in tale ambito. Si tratta di un problema di classificazione interessante, in quanto, le regioni coinvolte sono ben diversificate, estendendosi da aree omogenee, strutture geometriche ben definite a varie zone di suolo con tessiture regolari;

la principale difficoltà riscontrata in questo processo di classificazione risiede, infatti, nella sovrabbondanza del numero di classi da distinguere.

I risultati sperimentali hanno dimostrato che la classificazione, tramite l'implementazione degli HOG, presenta forti limiti per quanto riguarda le classi caratterizzate da zone omogenee. Tuttavia, si sono registrati incrementi molto significativi ed altamente soddisfacenti per quanto riguarda classi con strutture geometriche ben definite.

# **Summary**

Here you should write the summary

# **Ringraziamenti**

I candidati ringraziano vivamente il Granduca di Toscana per i mezzi messi loro a disposizione, ed il signor Von Braun, assistente del prof. Albert Einstein, per le informazioni riservate che egli ha gentilmente fornito loro, e per le utili discussioni che hanno permesso ai candidati di evitare di riscoprire l'acqua calda.

# Indice

<b>Sommario</b>	III
<b>Summary</b>	V
<b>Ringraziamenti</b>	VI
<b>1 Introduzione</b>	1
<b>2 Classificazione di immagini telerilevate ad alta risoluzione</b>	3
2.1 Cenni sul Telerilevamento . . . . .	4
2.1.1 Sensori . . . . .	4
2.1.2 Tipologie di sensori . . . . .	5
2.1.3 Il ruolo della risoluzione . . . . .	6
2.1.4 Telerilevamento tramite sensori ottici multispettrali . . . . .	6
2.2 Classificazione di immagini telerilevate . . . . .	9
2.2.1 Premessa . . . . .	9
2.2.2 Spazi di rappresentazione . . . . .	9
2.2.3 La classificazione nello spazio delle <i>feature</i> . . . . .	10
2.3 Il ruolo dell'informazione spaziale . . . . .	13
2.3.1 Estrazione dei parametri di tessitura . . . . .	13

2.3.2	Tecniche Region-based . . . . .	14
2.3.3	Markov Random Fields . . . . .	14
<b>3</b>	<b>Istogrammi di gradienti orientati per classificazione di immagini</b>	<b>17</b>
3.1	Schema generale dell'algoritmo . . . . .	18
3.2	Pulitura dal rumore . . . . .	18
3.3	Calcolo dei Gradienti . . . . .	20
3.4	Costruzione degli istogrammi . . . . .	24
3.5	Normalizzazione dei blocchi . . . . .	25
3.5.1	Schemi di normalizzazione dei blocchi . . . . .	26
3.6	Costruzione del vettore delle feature . . . . .	26
<b>4</b>	<b>SVM - <i>Support Vector Machine</i></b>	<b>29</b>
4.1	SVM lineare per classificazione binaria . . . . .	30
4.1.1	Hard Margin SVM . . . . .	30
4.1.2	Soft Margin SVM . . . . .	34
4.2	SVM non lineare per classificazione binaria . . . . .	36
4.3	SVM multclasse . . . . .	40

# Elenco delle tabelle

2.1 Principali intervalli di lunghezza d'onda significativi per il telerilevamento passivo . . . . .	8
3.1 Operatori derivativi più usati per il calcolo del gradiente . . . . .	23

# Elenco delle figure

2.1	Schema a blocchi (concettuale) di un sistema di telerilevamento.	4
2.2	Metodi di scansione rispettivamente dei <i>line scanner</i> (a), <i>whiskbroom scanner</i> (b) e <i>pushbroom scanner</i> (c).	7
2.3	Composizione dello spettro delle onde elettromagnetiche	8
2.4	Esempi di rappresentazione in diversi spazi.	10
3.1	Schema a blocchi del metodo HOG	18
3.2	Gaussiana in 3D di varianza $\sigma_x \sigma_y$ e media nulla	20
3.3	Riduzione del rumore nell'immagine di test tramite filtraggio gaussiano a media nulla e varianza $\sigma = 2$ pixel	20
3.4	Calcolo del gradiente per righe e per colonne dell'immagine di test	22
3.5	Schema di approssimazione del gradiente di un'immagine	22
3.6	Schema esemplificativo della distribuzione spaziale di istogrammi, celle e blocchi	27
4.1	Dataset linearmente separabile	31
4.2	Margine tra i campioni di training di due classi linearmente separabili	32
4.3	Dataset non linearmente separabile	34
4.4	Esempio di soft margin	35
4.5	Confronto di strategia tra Hard Margin SVM e Soft Margin SVM	37
4.6	Situazione esemplificativa della funzione di trasformazione	38

# Capitolo 1

## Introduzione

Questo documento è costituito nel seguente modo: nel Capitolo 2 viene introdotto il problema del telerilevamento, con accenni alle tipologie di sensori utilizzabili, ai tipi di classificatori e allo stato dell'arte delle varie strategie per l'estrazione di informazioni aggiuntive per il miglioramento della classificazione; nel Capitolo 3, invece, è presente un'analisi dettagliata dell'algoritmo di *detection* introdotto da Dalal e Triggs [2] con particolare riferimento alle modifiche da noi apportate necessarie per l'uso di questo approccio al contesto del telerilevamento. Il Capitolo 4 propone la trattazione matematica della teoria alla base del classificatore SVM, successivamente utilizzato per la fase di implementazione e test. Nel Capitolo ?? si procede con una analisi puntuale e precisa dei risultati ottenuti con l'implementazione dell'algoritmo HOG e della SVM, valutandone la potenzialità, i punti di forza e debolezza e le situazioni nelle quali è vantaggioso o no utilizzarlo, basandoci sugli indici di accuratezza introdotti nel Capitolo ??.



## Capitolo 2

# Classificazione di immagini telerilevate ad alta risoluzione

Il telerilevamento (*remote sensing*) viene definito come la disciplina che permette di ricavare informazioni su oggetti posti a distanza da uno o più sensori, che ne permettono l'estrazione. Questa definizione generale comprende un vasto range di sotto-discipline attente ad estrarre informazioni, tra le quali l'extrapolazione di dati di interesse ambientale, l'identificazione di oggetti posti sul fondale marino o fino a tecniche di *l'human detection*. Nel nostro caso, ci focalizzeremo in particolare nell'ambito dell'Osservazione della Terra (*Earth Observation*, EO) in cui l'oggetto da analizzare sarà una porzione di suolo.

## 2.1 Cenni sul Telerilevamento

### 2.1.1 Sensori

Nel caso di telerilevamento al fine di *EO*, i sensori sono tipicamente montati su un aereo (sensore aviotrasportato) o su un satellite (sensore satellitare), i quali trasducono l'onda elettromagnetica incidente su di essi in una immagine digitale, al fine di essere elaborata da un calcolatore. I sensori hanno, infatti, lo scopo di acquisire parametri ambientali utili ad una successiva elaborazione delle informazioni, seguendo il procedimento rappresentato dallo schema a blocchi riportato in Figura 2.1.

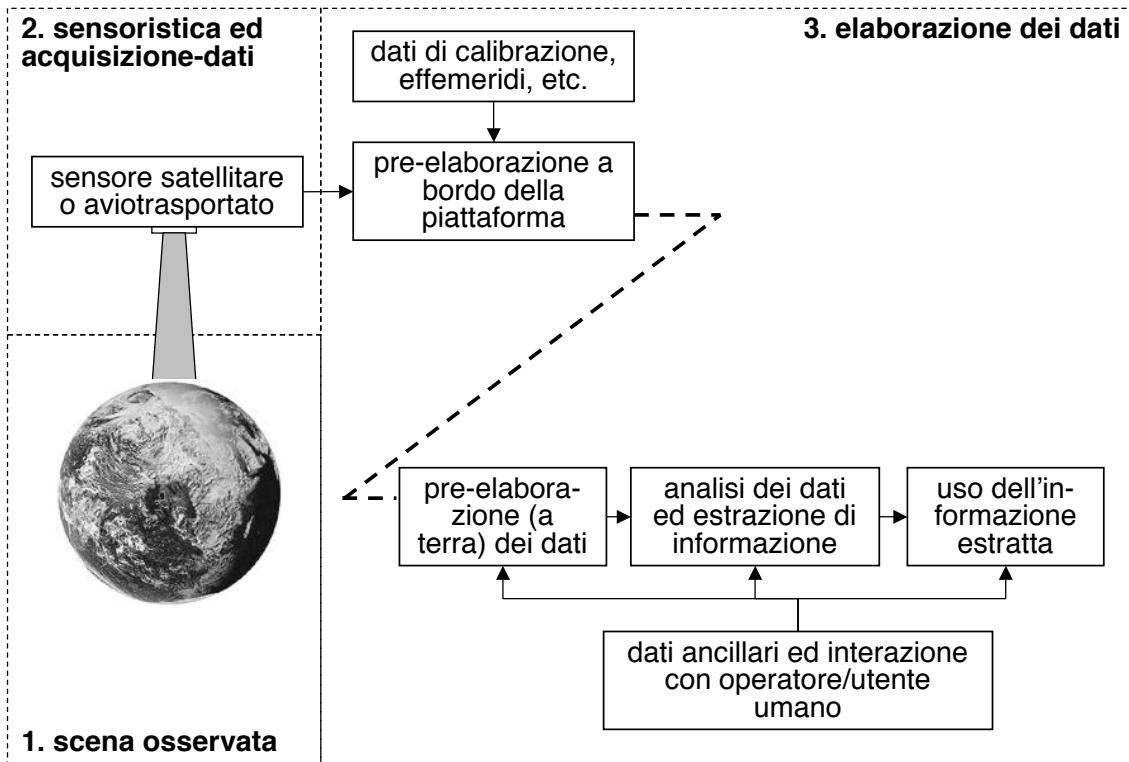


Figura 2.1. Schema a blocchi (concettuale) di un sistema di telerilevamento.

Il procedimento è il seguente:

- innanzitutto la radiazione emessa dall'oggetto in esame (ad esempio una porzione di

suolo) viene ricevuta in ingresso al fotorivelatore, per essere elaborata dal sistema;

- il sistema di elaborazione prevede in taluni casi, una fase iniziale di pre-elaborazione dei dati a bordo della piattaforma, basata su informazioni inerenti ai sensori (ad esempio dati di calibrazione del sensore) o alla piattaforma stessa (ad esempio effemeridi);
- successivamente, vengono effettuate operazioni di pre-elaborazione analoghe ma a terra, aventi fini quali un’ulteriore calibrazione dei dati o la rimozione di distorsioni geometriche o radiometriche dovute al movimento della piattaforma;
- infine il sistema procede all’elaborazione vera e propria dei dati telerilevati al fine di estrarre l’informazione che viene infine fornita all’utilizzatore.

### 2.1.2 Tipologie di sensori

Generalmente, i sensori per telerilevamento vengono divisi in due macro-famiglie, i sensori passivi e quelli attivi. La prima tipologia non trasmette alcun segnale ma, bensì, riceve unicamente la radiazione emessa dall’oggetto in esame, ovvero la radiazione elettromagnetica emessa dalla porzione di superficie, la quale può essere spontanea (tipicamente radiazione infrarossa) oppure la radiazione riflessa o diffusa proveniente dal sole (radiazione infrarossa e/o nello spettro del visibile). I sensori attivi, invece, trasmettono un’onda elettromagnetica nella direzione della superficie in esame e analizzano il segnale (simile ad un eco sonoro) ritrasmesso dalla porzione di suolo stessa. Tali onde elettromagnetiche sono tipicamente segnali laser, i quali usano sensori *Light Detection And Ranging* (LIDAR), oppure a microonde, analizzati tramite sensori *RAdio Detection And Ranging* (RADAR).

La nostra trattazione si baserà unicamente su segnali rilevati tramite sensori passivi, in particolare sensori ottici.

### 2.1.3 Il ruolo della risoluzione

Un parametro di vitale importanza per un sensore orientato all’ *EO* è la risoluzione, con la quale si intende una risoluzione spaziale, temporale e spettrale. La prima rappresenta il più piccolo dettaglio della superficie in esame che risulta distinguibile dopo essere stata estratta dal sensore, in particolare questo parametro viene espresso come la grandezza della superficie rappresentata da un singolo pixel. La risoluzione temporale, invece, è definita come la frequenza con cui il sensore osserva una stessa porzione di superficie. Essa è infatti definita come il tempo intercorso tra due passaggi del sensore sulla stessa area geografica, intervallo di tempo che può variare dal mese fino, addirittura, al singolo giorno. Infine, la risoluzione spettrale viene definita come il numero di bande (o canali) misurate per ciascun pixel e dalla larghezza di banda di ogni singolo canale. Ad esempio, il sensore iperspettrale AVIRIS campiona la radiazione incidente acquisendo 224 bande distinte, ognuna avente una larghezza pari a 9.3 nm, rendendolo un sensore ad alta risoluzione temporale, a differenza di sensori pancromatici che elaborano solamente l’intervallo di lunghezze d’onda della radiazione visibile.

### 2.1.4 Telerilevamento tramite sensori ottici multispettrali

La nostra analisi si focalizzerà su sensori passivi multispettrali costituiti da sistemi di elaborazione ottica, i quali indirizzano la radiazione incidente su uno o più foto-rivelatori, che trasducono una grandezza fisica (l’intensità della radiazione elettromagnetica) in una tensione elettrica. Queste tipologie di sensori vengono trasportati da una piattaforma che viaggia ad una quota  $h$  e fa percorrere al sensore una direzione di volo (direzione *in-track*), lungo la quale esso effettua uno scan lungo la direzione ortogonale (detta direzione *cross-track*). I sensori possono essere suddivisi, in base alla tecnica di scansione utilizzata, in tre famiglie:

- *Line Scanner*, sensori che montano un solo rivelatore ottico che scandisce l'intera riga per mezzo di uno specchio oscillante;
- *Whiskbroom Scanner*, i quali possiedono un array di rivelatori orientati secondo la direzione *in-track*, che scandiscono, in questo modo, più linee alla volta;
- *Pushbroom Scanner*, contenenti un numero elevato di foto-rivelatori, lungo la direzione *cross-track*, i quali permettono di scandire intere righe dell'immagine senza l'utilizzo di parti mobili.

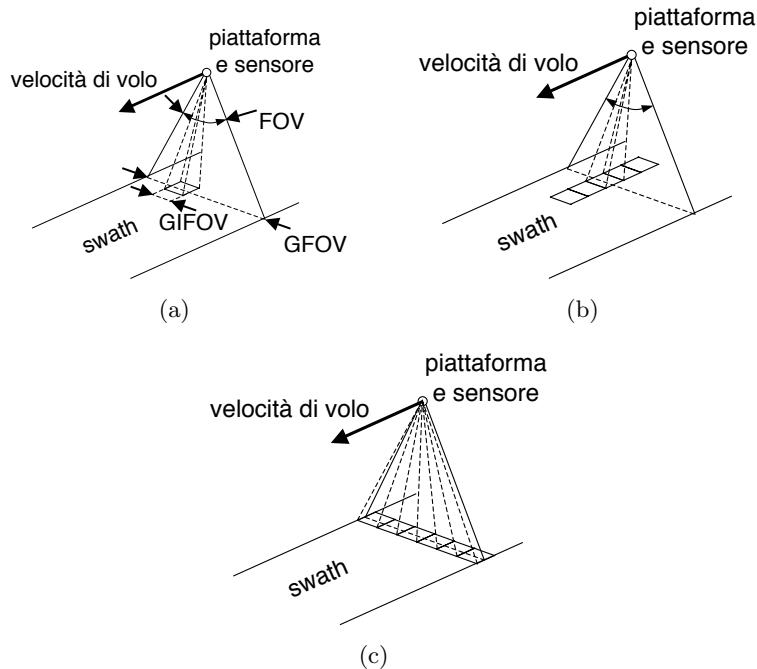


Figura 2.2. Metodi di scansione rispettivamente dei *line scanner* (a), *whiskbroom scanner* (b) e *pushbroom scanner* (c).

Parametri chiave dei sensori ottici sono il cosiddetto *Field Of View* (FOV), ovvero l'ampiezza espressa in radianti della zona di suolo osservata in direzione *cross-track* e la corrispondente larghezza a terra della striscia osservata, il *Ground Field Of View* (GFOV). In modo analogo, l'estensione angolare di ogni foto-rivelatore è detta *Istantaneous Field*

*Of View* (IFOV) mentre la sua proiezione viene detta *Ground Instantaneous Field Of View* (GIFOV). Fra queste grandezze esistono delle relazioni espresse dalle seguenti equazioni:

$$GFOV = 2h \tan \frac{FOV}{2} \quad (2.1)$$

$$GIFOV = 2h \tan \frac{IFOV}{2} \quad (2.2)$$

Tali sensori ottici passivi analizzano vari range di frequenze che tendenzialmente vanno dall'infrarosso termico (TIR, 8-9.5  $\mu\text{m}$ , 10-14  $\mu\text{m}$ ) allo spettro del visibile (VIS, 0.4-0.7  $\mu\text{m}$ ), rilevando la radianza spettrale, grandezza che permette di descrivere la distribuzione spaziale della radiazione elettromagnetica.

Tabella 2.1. Principali intervalli di lunghezza d'onda significativi per il telerilevamento passivo

Nome	Abbreviazione	Lunghezza d'onda [ $\mu\text{m}$ ]
Visible	VIS	0.4 – 0.7
Near InfraRed	NIR	0.7 – 1.1
Short Wave InfraRed	SWIR	1.1 – 1.35, 1.4 – 1.8, 2 – 2.5
Mid Wave InfraRed	MWIR	3 – 4, 4.5 – 5
Thermal Infrared	TIR	8 – 9.5, 10 – 14

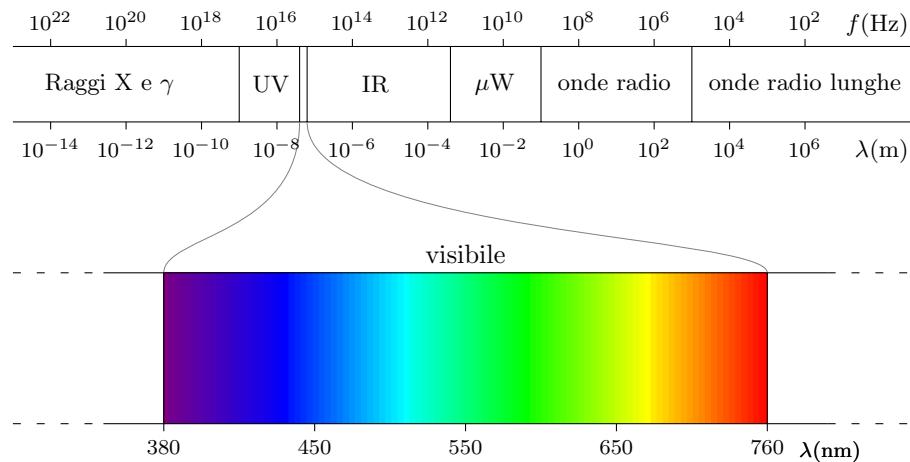


Figura 2.3. Composizione dello spettro delle onde elettromagnetiche

## 2.2 Classificazione di immagini telerilevate

### 2.2.1 Premessa

Si definisce classificazione di immagini telerilevate, il processo di assegnazione di una "etichetta" a ciascun pixel dell'immagine, in modo tale da renderlo appartenente ad una specifica classe, rappresentativa di una data copertura di suolo. La classificazione è il primo step del processo di estrappolazione di dati di carattere informativo da una immagine telerilevata, che vengono forniti, poi, alle successive fasi di riconoscimento (*matching*) ed interpretazione. Queste tre fasi compongono la disciplina del *Pattern Recognition*, il cui obiettivo è appunto sviluppare tecniche con cui implementare i tre processi. Oltre alla generazione di mappe tematiche tramite sistemi di telerilevamento, il *pattern recognition* abbraccia vari ambiti quali l'analisi di immagini biomedicali, orientate alla robotica (*Computer Vision*), o per la videosorveglianza.

### 2.2.2 Spazi di rappresentazione

Esistono tre principali metodologie di rappresentazione dei dati, la rappresentazione nello "spazio-immagine" (*Image Space*), la rappresentazione nello "spazio spettrale" (*Spectral Space*) e quella nello "spazio delle features" (*Features Space*). La prima e più immediata consiste nel visualizzare i dati canale per canale tramite terne RGB, in modo tale che ogni banda risulti una immagine a sé. La seconda tipologia consiste, invece, nel visualizzare per ciascun pixel, i livelli di grigio di ogni canale, per rappresentarli poi in un grafico. La rappresentazione nello "spazio delle feature" si basa sull'assegnazione di assi cartesiani distinti ad ogni banda, e, così facendo, ad ogni pixel viene assopciato un vettore n-dimensionale, con n il numero di canali. Quest'ultima rappresentazione non solamente evidenzia i livelli di grigio dei pixel ma anche la distribuzione statistica nelle varie bande, rendendola generalmente più vantaggiosa rispetto alle altre due in quanto a differenti distribuzioni nello

spazio delle feature corrispondono tipicamente a differenti coperture di suolo.

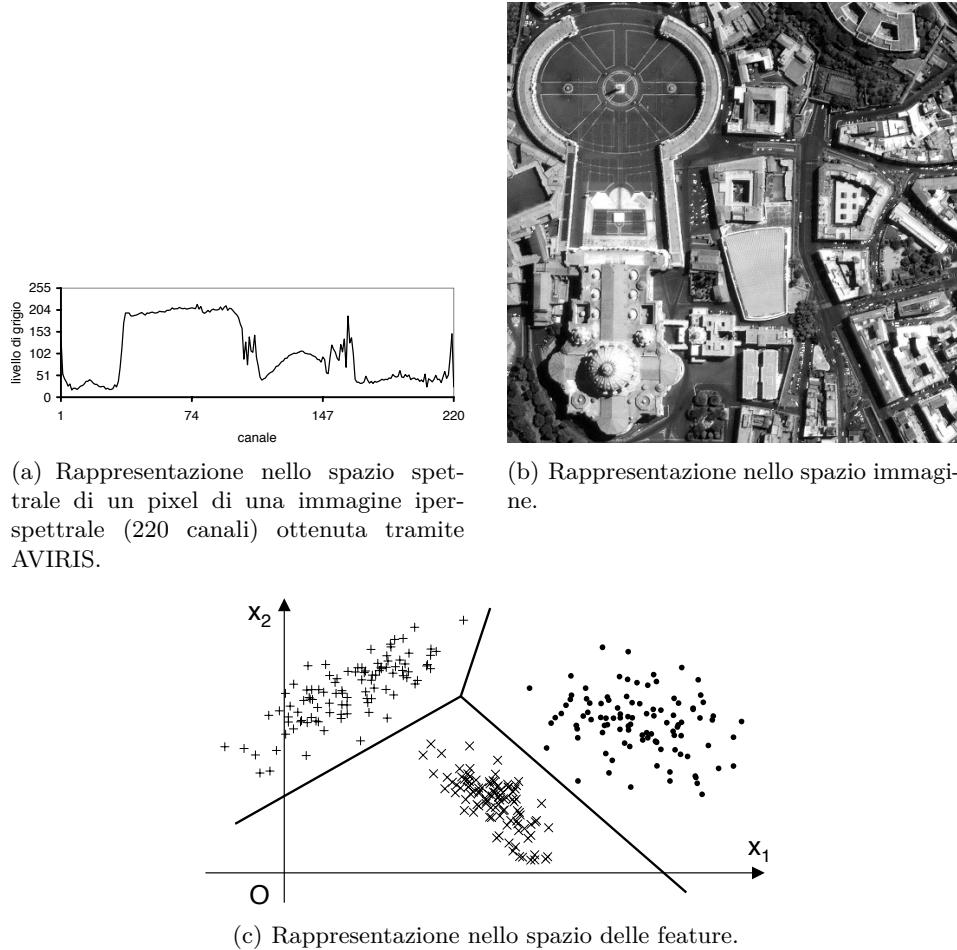


Figura 2.4. Esempi di rappresentazione in diversi spazi.

### 2.2.3 La classificazione nello spazio delle *feature*

A fini di classificazione, la rappresentazione nello spazio delle *feature* si rivela, solitamente, la più vantaggiosa, in quanto gli andamenti spettrali di pixel appartenenti a classi distinte possono essere maggiormente separabili. In quest'ottica, classificare significa quindi partizionare lo spazio delle "feature" in opportuni sottoinsiemi, ciascuno associato ad una data classe.

I classificatori possono essere divisi in due macro-famiglie, classificatori supervisionati (*supervised*) e non supervisionati (*unsupervised*). La prima tipologia prevede principalmente due fasi, una fase iniziale di addestramento (*training*) e una di verifica (*test*); nel primo step, il sistema ottimizza parametri del classificatore tramite un insieme di pixel preclassificati (*training set*), fino a raggiungere una adeguata accuratezza. Successivamente, il sistema viene testato in modo analogo alla fase di training, attraverso un insieme di pixel preclassificati ma differenti rispetto a quelli coinvolti nella fase di *training*. Nei classificatori non supervisionati, invece, non viene utilizzato alcun *training set*, solitamente perchè non sono note né facilmente identificabili le classi coinvolte nell'applicazione in esame.

Nei capitoli a seguire la nostra attenzione sarà focalizzata sul concetto di classificazione supervisionata, che si basa, in generale, su tre assunti:

- si ha una conoscenza a priori esaustiva su un sottoinsieme di pixel preclassificati (*training set*);
- le classi esistono in numero finito e sono note *ex ante*;
- ogni pixel è rappresentabile tramite un insieme di valori detti vettore delle *feature* (*feature vector*).

### Concetti chiave e definizioni

Data un'immagine telerilevata, a ciascun pixel  $(m, n) \in \mathbb{Z}$  può essere associato un vettore d-dimensionale delle *feature*  $\mathbf{x}(m, n) \in \mathbb{R}^d$ , le cui componenti (le d *feature*) possono

essere non solamente i livelli di grigio dei pixel  $(m, n)$  nelle varie bande, ma anche parametri aggiuntivi come ad esempio i cosiddetti parametri di tessitura (o *feature* di tessitura).

Le *feature* di tessitura vengono estratte al fine di analizzare le differenze nella distribuzione spaziale dei livelli di grigio dei pixel, permettendo di distinguere coperture di suolo differenti. In generale, se i pixel estratti da classi differenti sono situati in zone disgiunte dello spazio delle feature, la accuratezza di classificazione è tanto maggiore quanto più sono separate le regioni. Se infatti, in una stessa regione si trovano classi distinte, esse risulteranno sovrapposte e quindi distinguerle risulterà largamente più difficoltoso.

Dato un insieme  $\Omega = \{\omega_1, \dots, \omega_c\}$  costituito da  $C$  classi distinte, note a priori, si assume che ogni oggetto o entità da analizzare sia appartenente ad una ed una sola classe (nel nostro caso ad una data copertura del suolo). A ciascun pixel è associata, quindi, anche un'etichetta di classe  $y(m, n) \in \Omega$  e l'immagine  $y(m, n)$  (con  $m, n \in \mathbb{Z}$ ) è detta mappa di classificazione.

Il *training set* è l'insieme dei pixel di cui è nota a priori l'etichetta di classe. L'insieme di tali etichette rappresenta un'ulteriore immagine, detta *mappa di training*, che evidenzia alcuni pixel dell'immagine assegnati a ciascuna classe, distinguendoli dai "pixel di sfondo", cioè quei pixel per i quali l'etichetta di classe è incognita.

## 2.3 Il ruolo dell’informazione spaziale

Una ulteriore distinzione che permette di differenziare tra loro le tipologie di classificatori è quella inherente al ruolo dei pixel adiacenti nella analisi della copertura al suolo. Esistono, infatti, classificatori cosiddetti non contestuali, in cui la copertura viene analizzata senza tener conto dei pixel adiacenti, risparmiando così un alto costo computazionale, ma trascurando la forte correlazione tra pixel limitrofi. Infatti, la probabilità che una zona sia formata da pixel tutti appartenenti alla stessa classe è molto elevata; si pensi solamente a zone boschive o fluviali in cui tali regioni possono estendersi per chilometri.

Queste due tipologie di classificatori sono chiaramente differenti anche in base all’ambito in cui vengono applicate; mentre le tecniche di classificazione supervisionata non contestuale risultano molto efficaci e largamente consolidate per le immagini con una risoluzione piuttosto grossolana, mostrano forti limiti per le immagini *Very High Resolution*. Una maggiore risoluzione spaziale comporta, infatti, una maggiore eterogeneità e una corrispondente buona definizione delle strutture geometriche quali strade e edifici, caratteristiche che rendono necessario l’utilizzo di classificatori contestuali. A tal fin, un ruolo chiave viene giocato principalmente da tre approcci metodologici, l’estrazione di parametri di tessitura, metodi basati su regioni e oggetti e i *Markov Random Fields* (MRF).

### 2.3.1 Estrazione dei parametri di tessitura

L’estrazione di *texture* ha come obiettivo quello di rilevare, in una determinata regione dell’immagine, strutture ripetitive nella distribuzione spaziale dei pixel quali zone urbane o boschive. Ciò fornisce una sorgente di dati complementare per le applicazioni in cui l’informazione relativa unicamente allo spettro dell’immagine risulta non sufficiente ai fini della classificazione. Le principali tecniche di estrazione di parametri di tessitura si basano

sui semivariogrammi, che consistono in una statistica del secondo ordine delle intensità dei pixel, sulla morfologia matematica o addirittura su metodi che coinvolgono finestre mobili in cui vengono effettuate analisi non sul singolo pixel ma su una intera zona adiacente.

### 2.3.2 Tecniche Region-based

Gli approcci basati invece sulle regioni (*region-based methods*) si basano su tecniche che puntano a suddividere le immagini in segmenti o regioni omogenee. In generale, una buona tecnica di segmentazione possiede:

- pixel nella stessa categoria aventi livelli di grigio simili e formano una regione connessa;
- pixel adiacenti che sono in categorie differenti hanno valori differenti.

### 2.3.3 Markov Random Fields

I *Markov Random Fields*, che generalizzano il concetto di catena markoviana monodimensionale ad un sistema 2D, massimizzano l'accuratezza tramite la dipendenza che sussiste tra pixel adiacenti. Esso offre, infatti, una soluzione computazionalmente efficiente per restringere la zona di interesse dall'intera immagine (elaborazione globale) ad un intorno del pixel (elaborazione locale). In particolare, siano  $i$  e  $j$  due pixel dell'immagine, si ha allora che se la funzione di probabilità  $P(Y) > 0$  per ogni configurazione  $Y$  e se la seguente condizione è garantita per tutti i pixel  $i$  dell'immagine:

$$P(y_i|y_j, j \neq i) = P(y_i|y_j, j \sim i) \quad (2.3)$$

Ciò indica che la distribuzione di probabilità delle etichette di ciascun pixel  $i$ , condizionata ai valori di tutti gli altri pixel dell'immagine, può essere ristretta alla distribuzione delle etichette di  $i$  condizionato solamente alle etichette dei pixel adiacenti. Si può chiaramente osservare come tale definizione sia una generalizzazione dei processi markoviani

monodimensionali, in cui la probabilità di transizione da uno stato ad un altro dipende unicamente dallo stato precedente.



## Capitolo 3

# Iistogrammi di gradienti orientati per classificazione di immagini

In questo capitolo verrà presentato il metodo di estrazione di *feature* tramite istogrammi di gradienti orientati (*Histogram of Oriented Gradients*, HOG). Questo metodo, introdotto per la prima volta da Dalal e Triggs [2], si basa sulla valutazione di istogrammi calcolati in base a direzione e intensità dei gradienti dell’immagine in ingresso. L’idea di base è che la forma di un oggetto può essere rappresentata in modo esaustivo dalla distribuzione di modulo e fase dei gradienti locali.

### 3.1 Schema generale dell’algoritmo

In pratica, dopo aver calcolato il gradiente dell’immagine punto per punto, essa viene divisa in piccole regioni spaziali denominate "celle". Per ogni cella viene costruito un istogramma sulla base della direzione del gradiente dei suoi pixel. Per una migliore invarianza agli effetti dovuti all’illuminazione viene effettuata una normalizzazione degli istogrammi delle celle, ottenuta sulla base di una regione spaziale di dimensione maggiore denominata "blocco". I singoli istogrammi, combinati, danno origine ai vettori delle *feature* che vengono poi utilizzate da un classificatore, al fine di classificare l’immagine di partenza. Lo schema a blocchi del descrittore HOG è rappresentato in Figura 3.1.

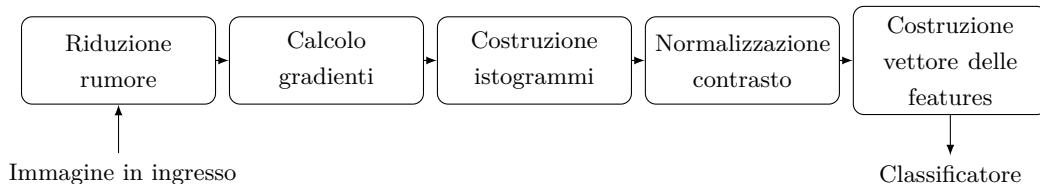


Figura 3.1. Schema a blocchi del metodo HOG

Di seguito verranno riportate in maniera dettagliata le singole fasi che lo costituiscono, evidenziando le scelte progettuali e analizzando come i diversi parametri varino la prestazione.

### 3.2 Pulitura dal rumore

La strumentazione del sensore passivo multispettrale introduce, in generale, disturbi nella misura della radianza, dovuta ad esempio ad una variazione di riflettanza, emittanza o

illuminazione solare indicati complessivamente con il termine di rumore. Questo può essere modellato come rumore additivo gaussiano<sup>1</sup>. Per limitarne gli effetti sull’immagine di ingresso, è opportuno applicare un filtraggio di pulizia dal rumore.

La riduzione del rumore (*noise cleaning*) può essere ottunuta tramite la convoluzione con una gaussiana bidimensionale (Figura 3.2) che, essendo notoriamente un passabasso, elimina o riduce le frequenze dove il contenuto spettrale di rumore è maggiore di quello delle immagini. Lo stesso tipo di filtraggio può essere applicato anche sulle immagini in uscita dall’algoritmo HOG.

La gaussiana di varianza  $\sigma_x \ \sigma_y$  è data da:

$$H_\sigma(x, y) = \frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot e^{-\frac{x^2}{(\sigma_x)^2}} \cdot \frac{1}{\sigma_y \cdot \sqrt{2\pi}} \cdot e^{-\frac{y^2}{(\sigma_y)^2}} \quad (3.1)$$

La sua versione discreta si ottiene campionandola su una finestra quadrata di dimensioni  $K \times K$  con  $K > 3 \cdot 2\sqrt{\sigma}$  per rendere trascurabili gli effetti del troncamento. La maschera così ottenuta viene moltiplicata per  $\frac{1}{c}$ , dove il fattore di normalizzazione  $c$  è scelto in modo tale che:

$$\sum_j \sum_k g_{jk} = 1 \quad (3.2)$$

dove  $g_{jk}$  è il generico elemento della gaussiana campionata.

---

<sup>1</sup>Come conseguenza del teorema centrale del limite, la distribuzione di probabilità di ogni pixel può essere modellata come gaussiana, visto l’elevato numero di contributi di variabili che concorrono alla definizione del pixel stesso.

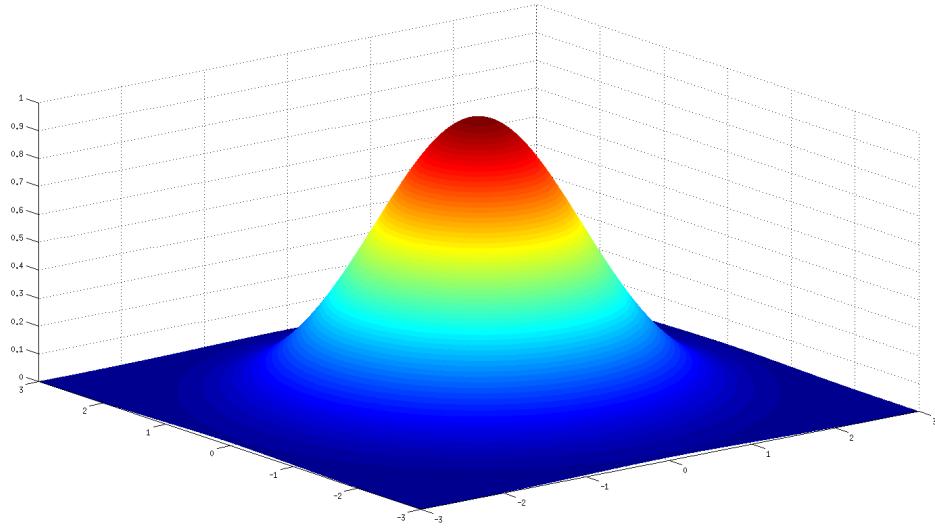
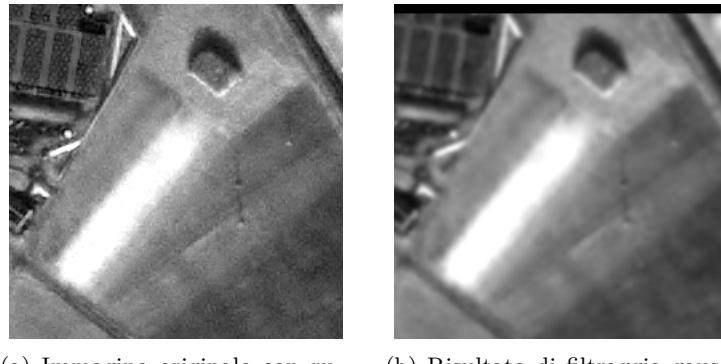


Figura 3.2. Gaussiana in 3D di varianza  $\sigma_x \ \sigma_y$  e media nulla



(a) Immagine originale con rumore spazialmente uniforme  
 (b) Risultato di filtraggio gaussiano

Figura 3.3. Riduzione del rumore nell’immagine di test tramite filtraggio gaussiano a media nulla e varianza  $\sigma = 2$  pixel

### 3.3 Calcolo dei Gradienti

Il gradiente di un’immagine misura la variazione direzionale di intensità della stessa. Matematicamente, il gradiente di una funzione a due variabili associa ad ogni punto dell’immagine un vettore 2D con componenti date dalle derivate calcolate sulle direzioni orizzontali

e verticali. Poiché la funzione intensità di un’immagine è conosciuta solo in punti discreti, si assume che le sue derivate siano calcolate su di una funzione intensità continua campionata nei punti dell’immagine.

Dal punto di vista matematico, detta  $F(x, y)$  una funzione continua e derivabile, il suo gradiente è dato da:

$$\nabla F = \frac{\partial F}{\partial x} \hat{x} + \frac{\partial F}{\partial y} \hat{y} \quad (3.3)$$

dove:

- $\frac{\partial F}{\partial x} = G_x$  è il gradiente calcolato lungo la direzione x;
- $\frac{\partial F}{\partial y} = G_y$  è il gradiente calcolato lungo la direzione y.

Nel caso di immagini bidimensionali, approssimazioni di queste funzioni derivative possono essere definite al variare del grado di accuratezza. Il seguente schema evidenzia uno dei metodi di approssimazione più comune:

La stima di  $G_x$  e  $G_y$  si ottiene dall’espressione della derivata monodimensionale limitata ad un piccolo intorno (spesso solo 3 punti). L’approccio più semplice utilizza una risposta all’impulso monodimensionale di tipo  $\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$ .

Dal momento che la derivata, soprattutto quando calcolata su un intervallo così breve, è molto sensibile al rumore, si usa mediare in direzione ortogonale prima di effettuare la differenza per stabilizzarla. Con l’operatore di Prewitt, ad esempio,  $G_y$  viene stimata mediante la differenza in orizzontale della media in verticale calcolata su tre punti. Nella Tabella 3.3 sono riassunti le varianti più comuni.

Questi operatori forniscono risultati accettabili solo per immagini poco rumorose, infatti è opportuno che l’immagine, come introdotto precedentemente, in ingresso sia filtrata

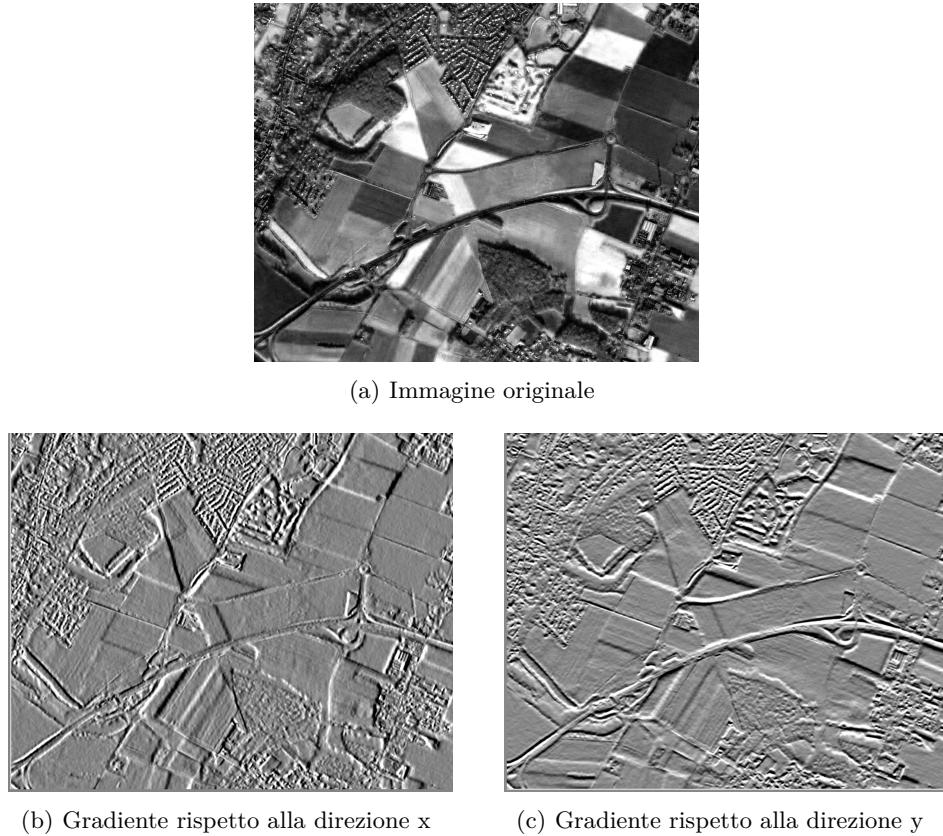


Figura 3.4. Calcolo del gradiente per righe e per colonne dell’immagine di test

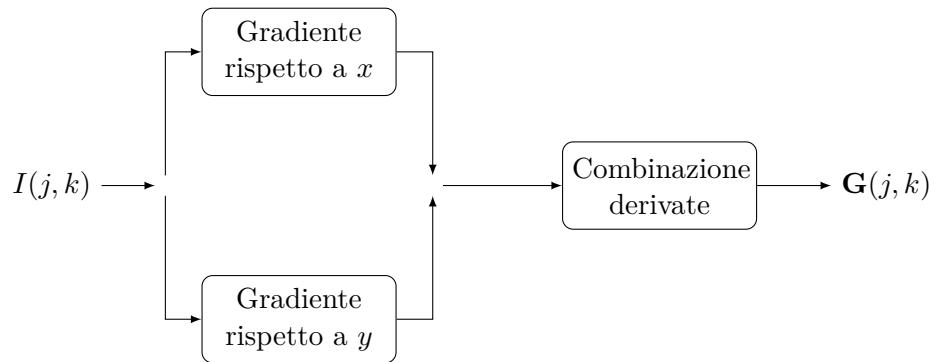


Figura 3.5. Schema di approssimazione del gradiente di un’immagine

al fine di limitare gli effetti del rumore.

Tabella 3.1. Operatori derivativi più usati per il calcolo del gradiente

	Gradiente per riga	Gradiente per colonna
Roberts	$\begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
Prewitt	$\frac{1}{3} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$
Sobel	$\frac{1}{4} \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$	$\frac{1}{4} \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$

Siano  $G_x$  e  $G_y$  le immagini gradiente generate da:

$$G_x(j, k) = Img(j, k) * h_x(j, k) \quad (3.4)$$

$$G_y(j, k) = Img(j, k) * h_y(j, k) \quad (3.5)$$

dove  $*$  rappresenta la convoluzione e dove  $h_x$  e  $h_y$  rappresentano rispettivamente la maschera per riga e per colonna scelta tra gli operatori descritti precedentemente. Modulo e fase del gradiente si ottengono, per ogni pixel dell'immagine, combinando  $G_x$  e  $G_y$  rispettivamente come:

$$G(j, k) = \sqrt{(G_x(j, k))^2 + (G_y(j, k))^2} \quad (3.6)$$

$$\theta(j, k) = atan \left( \frac{G_y(j, k)}{G_x(j, k)} \right) \quad (3.7)$$

Per alcune applicazioni, il segno del gradiente, e quindi il valore di  $\theta$  compreso tra  $[0, 2\pi]$  è rilevante per il problema di classificazione. Nella maggior parte delle applicazioni però, il

segno del gradiente fornisce informazioni secondarie e irrilevanti; dunque  $\theta(j, k)$  può essere calcolato nell'intervallo  $[0, \pi]$ .

### 3.4 Costruzione degli istogrammi

Il passo successivo al calcolo dei gradienti è quello di costruire gli istogrammi. A tale fine l'immagine di ingresso viene divisa in "celle" che possono essere di due diverse forme geometriche: rettangolari (R-HOG), di dimensioni  $N \times N$  pixel, o circolari (C-HOG), di raggio  $N$  pixel.

Per ogni cella viene creato un istogramma accumulando all'interno dei canali i voti dei gradienti di ciascun pixel della cella, pesati in base al modulo del gradiente. Gli *orientation bins* sono spaziati uniformemente nell'intervallo  $[0, 2\pi]$  (gradiente con segno) o  $[0, \pi]$  (gradiente senza segno).

Si considerino  $n_\theta$  *angle bins* tra  $[0, \pi]$  (o, come discusso in precedenza, potenzialmente, tra  $[0, 2\pi]$ ). I descrittori HOG racchiudono le statistiche locali dei gradienti (modulo e fase) in quanto ogni pixel dà il suo voto ad uno specifico *angle bin* il cui peso è proporzionale alla magnitudine del gradiente in quel determinato pixel.

Detto  $n_\theta$  il numero di *angle bins* e posto  $\phi_k = \frac{180 \cdot k}{n_\theta}$ , con  $k = 0 \dots n_\theta$ , e dette  $n_x$  e  $n_y$  il numero di celle rispettivamente sulla riga e sulla colonna dell'immagine, si costruisce una matrice tridimensionale di dimensione  $n_x \times n_y \times n_\theta$  data da:

$$V(i, j, k) = f(G(i, j))\delta(\phi_{k-1} < \theta(i, j) \leq \phi_k), \quad \text{con } k = 1 \dots n_\theta \quad (3.8)$$

dove  $\delta(x)$  restituisce 1 quando l'argomento è vero, 0 quando è falso e  $f$  è una funzione del modulo del gradiente (lineare, radice quadrata, quadrato o una forma saturata tra quelle riportate per rappresentare la presenza o assenza di contorni).

Un’istogramma con  $n_\theta$  canali (*orientations bins*) ad esempio è costruito nel modo seguente.

- i voti di tutti i gradienti della celle che hanno un angolo compreso nell’intervallo  $\left[0, \frac{180}{n_\theta}\right)$  sono accumulati nel primo canale;
- i voti di tutti i gradienti della celle che hanno un angolo compreso nell’intervallo  $\left[\frac{180}{n_\theta}, \frac{180}{n_\theta} \cdot 2\right)$  sono accumulati nel secondo canale;
- i voti di tutti i gradienti della celle che hanno un angolo compreso nell’intervallo  $\left[\frac{180}{n_\theta} \cdot 2, \frac{180}{n_\theta} \cdot 3\right)$  sono accumulati nel terzo canale;
- $\vdots$
- fino al canale  $n_\theta$  dove sono accumulati i gradienti delle celle che hanno angolo compreso tra  $\left[\frac{180}{n_\theta} \cdot k - 1, \frac{180}{n_\theta} \cdot k\right)$

Il voto dato da ciascun pixel è proporzionale alla magnitudine del gradiente di quel punto, poiché risulta importante associare ad ogni orientazione del gradiente in un dato intervallo un voto che tenga conto dell’importanza del gradiente in un determinato pixel. Infatti, il gradiente calcolato attorno a un bordo risulta molto più significativo di quello calcolato in una zona uniforme dell’immagine ed è essenziale per estrarre informazioni utili ad avere una descrizione dettagliata delle strutture geometriche presenti.

### 3.5 Normalizzazione dei blocchi

La magnitudine del gradiente utilizzata nei descrittori HOG è sensibile ai cambiamenti locali in ambienti luminosi. Per questo motivo è essenziale per il raggiungimento di elevate prestazioni eseguire una normalizzazione dell’istogramma, così da introdurre una migliore invarianza a diverse condizioni di luminosità, contrasto, ombre.

In questa fase ogni istogramma è dunque normalizzato separatamente in base ad un fattore di normalizzazione calcolato sulla base di raggruppamenti di celle circostanti, denominati "blocchi". Ognuno di questi blocchi è composto da  $M \times M$  celle.

### 3.5.1 Schemi di normalizzazione dei blocchi

Si possono valutare diversi schemi per calcolare il valore di normalizzazione.

Detto  $\mathbf{v}$  il vettore descrittore non ancora normalizzato e sia  $\|\mathbf{v}\|_k$  la sua norma  $k$ -esima, con  $k = 1, 2$ , si possono utilizzare i seguenti schemi come proposto da Dalal e Triggs [2]:

- L1-Norm :  $\mathbf{v} = \frac{\mathbf{v}}{\|\mathbf{v}\|_1 + \varepsilon}$
- L2-Norm :  $\mathbf{v} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2^2} + \varepsilon^2$

dove  $\varepsilon$  è una costante introdotta per evitare la divisione per zero e sufficientemente piccola da non alterare significativamente il risultato.

Un altro metodo per effettuare la normalizzazione è quello proposto da Torrione e Morton [7].

Detta  $N(c)$  l'insieme di celle comprese nel blocco di interesse, il valore di normalizzazione può essere calcolato come segue:

$$H(c, k) = \frac{H_1(c, k)}{\left( \sum_{c_i \in N(c)} \sqrt{\|H_1(c_i)\|_2^2 + \varepsilon^2} \right)} \quad (3.9)$$

dove  $H_1(c, k) = \sum_{(i,j) \in c} V(i, j, k)$  e  $H_1(c)$  rappresenta il vettore colonna  $[H_1(c, 1), \dots, H_1(c, n_\theta)]^T$ .

## 3.6 Costruzione del vettore delle feature

A questo punto si procede alla costruzione del descrittore vettoriale (*feature vector*), che verrà classificato mediante un classificatore. Il descrittore avrà le stesse dimensioni dell'immagine originale con un numero di bande pari a  $C + C \times B$ , dove  $C$  è il numero di

canali dell’immagine da classificare e  $B$  il numero di *bins* scelto nella fase di costruzione degli istogrammi.

Tenendo ben presente che per immagini multispettrali l’HOG viene calcolato separatamente per ogni canale, la costruzione dei vettori delle *feature* avviene nel modo seguente: per ogni pixel vengono concatenati l’immagine originale a  $C$  canali con i  $C$  istogrammi, costituiti da  $B$  canali, relativi a ciascuna cella e alla banda corrispondente.

Matematicamente, detto  $\mathbf{X}_i$  il vettore delle *feature* corrispondente all’  $i$ -esimo pixel dell’immagine:

$$\mathbf{X}_i = [\mathbf{X}_{RGB_i}, \mathbf{f}(\mathbf{X}_{RGB_i})] \quad (3.10)$$

con

$$\mathbf{f}(\cdot) = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k) \quad (3.11)$$

dove  $\mathbf{f}_k$  è la fusione HOG calcolata sulla singola banda  $k$ -esima e  $\mathbf{X}_{RGB}$  è l’immagine originale.

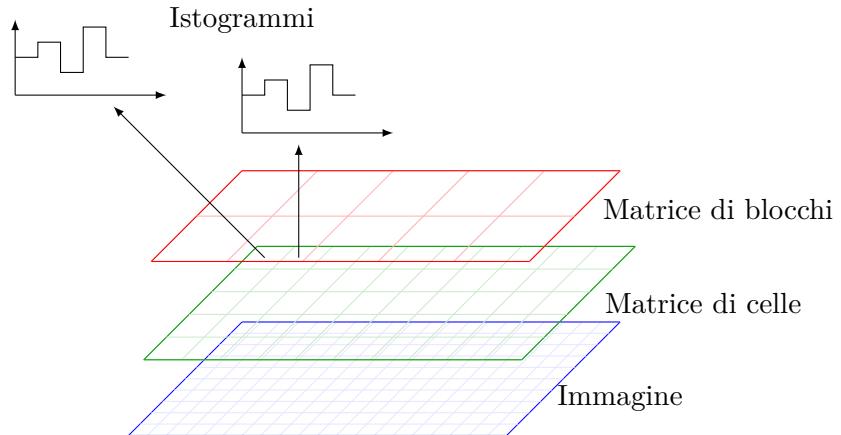


Figura 3.6. Schema esemplificativo della distribuzione spaziale di istogrammi, celle e blocchi



## Capitolo 4

# SVM - *Support Vector Machine*

Una SVM (*Support Vector Machine*) è un modello di apprendimento supervisionato associato ad algoritmi largamente utilizzati per il *data analysis* e il *pattern recognition* (tra cui anche il problema della classificazione). L'SVM è un modello abbastanza recente; anche se la sua formulazione risale agli anni '60, solo negli ultimi 15/20 anni si è registrato un incremento significativo nell'uso di algoritmi SVM per classificazione. La sua fortuna risiede nel fatto che può essere facilmente esteso in molti campi, senza stravolgere la semplicità che caratterizza questi classificatori.

Questo capitolo servirà per introdurre la teoria matematica su cui si basa un classificatore SVM e sarà impostato nel seguente modo: si inizierà descrivendo dettagliatamente il modello più semplice di SVM (SVM lineare per classificazione binaria), poi si proseguirà con la descrizione di come è possibile estendere questo modello per una classificazione non-lineare e, infine, verranno introdotte le possibili modifiche che permettono la classificazione multclasse.

## 4.1 SVM lineare per classificazione binaria

Dato un *training set* linearmente separabile  $(\mathbf{x}_i, y_i)$ ,  $i = 1, \dots, N$ , con  $\mathbf{x}_i \in \mathbb{R}^d$  e  $y_i \in \{-1, 1\}$ , l'obiettivo è addestrare un classificatore affinché

$$f(\mathbf{x}_i) : \begin{cases} > 0 & \text{per } y_i = +1 \\ < 0 & \text{per } y_i = -1 \end{cases} \quad (4.1)$$

ovvero  $y_i f(\mathbf{x}_i) > 0$  per una corretta classificazione. La funzione  $f$  è detta *regola di decisione* ed è costruita in modo tale che, preso un qualsiasi *campione incognito*  $\mathbf{u}$  da classificare, il valore di  $f$  valutato in  $\mathbf{u}$  restituisce una stima dell'etichetta  $y_u$  da associare. In equazioni:

$$f \text{ tale che } \begin{cases} \text{Se } f(\mathbf{u}) > 0 & \text{allora } y_u = +1 \\ \text{Se } f(\mathbf{u}) < 0 & \text{allora } y_u = -1 \end{cases} \quad (4.2)$$

### 4.1.1 Hard Margin SVM

Un insieme di elementi in  $\mathbb{R}^d$  è linearmente separabile se esiste almeno un iperpiano (che in generale avrà dimensione  $d - 1$ ) in grado di separare, nello spazio vettoriale dei campioni in ingresso, quelli che richiedono un'etichetta positiva da quelli che richiedono un'etichetta negativa.

Si prenda, per esempio, la situazione proposta in Figura 4.1: qui si può facilmente evincere che esiste almeno un iperpiano (in questo caso una retta) che divida lo spazio in due semispazi, ciascuno dei quali contiene campioni di una sola classe.

Il problema nasce dal fatto che possono esistere infiniti iperpiani e la scelta di quale usare per l'addestramento potrebbe avere ripercussioni notevoli per la fase di classificazione. L'SVM risolve questo problema cercando l'iperpiano che massimizza il margine tra i due insiemi di elementi.

In termini più matematici, un iperpiano in  $\mathbb{R}^d$  ha forma:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (4.3)$$

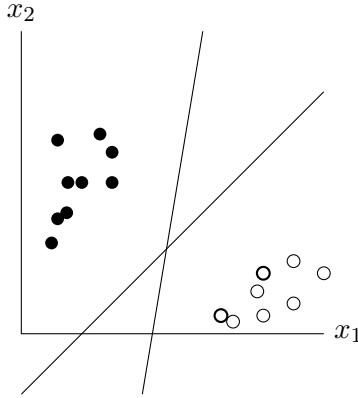


Figura 4.1. Dataset linearmente separabile

dove  $\mathbf{w} \in \mathbb{R}^d$  è la normale all'iperpiano e  $b/\|\mathbf{w}\|$  è la distanza perpendicolare dall'iperpiano all'origine.

Alla luce di ciò, si può riscrivere la *regola di decisione* presentata nell'equazione (4.2) nel seguente modo:

$$\begin{cases} \text{Se } \mathbf{w} \cdot \mathbf{u} + b > 0 \text{ allora } y_u = +1 \\ \text{Se } \mathbf{w} \cdot \mathbf{u} + b < 0 \text{ allora } y_u = -1 \end{cases} \quad (4.4)$$

Dato che  $\mathbf{w} \cdot \mathbf{x} + b = 0$  e  $c(\mathbf{w} \cdot \mathbf{x} + b) = 0$  definiscono la stessa regola di decisione (per  $c > 0$ ), si ha libertà di scegliere la normalizzazione di  $\mathbf{w}$ . In questo caso, si può scegliere il fattore di normalizzazione in modo tale che  $\mathbf{w} \cdot \mathbf{x}_i + b \geq 1$  e  $\mathbf{w} \cdot \mathbf{x}_i + b \leq -1$ , per gli elementi della prima classe e della seconda rispettivamente. Per convenienza matematica, dato che  $y_i \in \{-1, 1\}$ , le due identità precedenti possono essere riscritte nel seguente modo:

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i = 1, \dots, N \quad (4.5)$$

o, analogamente,

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad \forall i = 1, \dots, N \quad (4.6)$$

con l'uguaglianza valida per gli elementi sul bordo (i punti rossi in Figura 4.2).

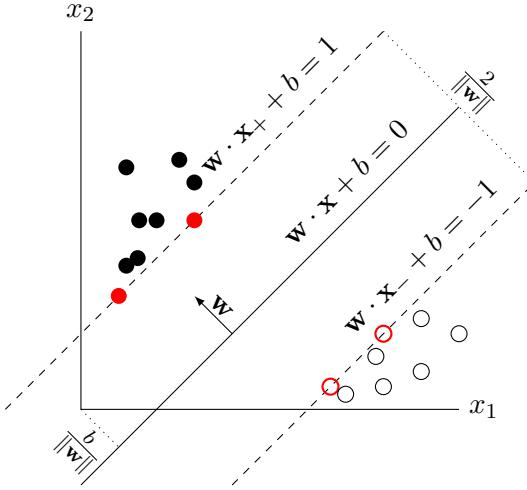


Figura 4.2. Margine tra i campioni di training di due classi linearmente separabili

Geometricamente, sotto queste ipotesi, si dimostra che il margine risulta essere:

$$\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_+ - \mathbf{x}_-) = \frac{2}{\|\mathbf{w}\|} \quad (4.7)$$

dove  $\mathbf{x}_+$  e  $\mathbf{x}_-$  sono i campioni delle due classi (rispettivamente) più vicini al iperpiano separatore (vedi Figura 4.2 ed equazione (4.6)).

L'obiettivo ora è quindi massimizzare questo valore, che equivale a minimizzare  $\|\mathbf{w}\|$ , il quale, a sua volta, per convenienza matematica, può essere espresso nel seguente termine:

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{vincolato a } y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 = 0 \quad \forall i = 1, \dots, N \end{aligned} \quad (4.8)$$

Essendo questo un chiaro esempio di calcolo di estremi vincolati, per semplificare i calcoli, si possono usare i *Moltiplicatori di Lagrange*, che permettono di lavorare su un problema duale, ovvero l'ottimizzazione (minimizzazione rispetto a  $\mathbf{w}$  e a  $b$ ) della seguente funzione:

$$\min_{\mathbf{w}, b} L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] \quad (4.9)$$

dove si sottintende, da questo punto, che la sommatoria sia per ogni  $i = 1, \dots, N$ .

Procedendo con il calcolo del gradiente di  $L$  si ottengono i seguenti risultati:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (4.10)$$

$$\frac{\partial L}{\partial b} = - \sum_i \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i y_i = 0 \quad (4.11)$$

L'equazione (4.10) suggerisce che il vettore  $\mathbf{w}$  non sia altro che una combinazione lineare di alcuni vettori di training (per alcuni di loro  $\alpha_i$  sarà pari a 0), mentre l'equazione (4.11) sarà utile più avanti.

A questo punto si introduce il cosiddetto *Problema lagrangiano duale*: invece di *minimizzare* rispetto a  $\mathbf{w}$  e  $b$ , si può *massimizzare* rispetto ad  $\alpha$  con vincoli le relazioni ottenute precedentemente per  $\mathbf{w}$  e  $b$  (equazioni (4.10) e (4.11))<sup>1</sup>.

Dato che l'equazione (4.10) restituisce una definizione per  $\mathbf{w}$ , possiamo sostituire questo risultato nella lagrangiana  $L$  (eq.ne (4.9)), ottenendo:

$$L(\mathbf{w}, b) = \frac{1}{2} (\sum_i \alpha_i y_i \mathbf{x}_i) \cdot (\sum_j \alpha_j y_j \mathbf{x}_j) + \\ - \sum_i [\alpha_i y_i \mathbf{x}_i \cdot (\sum_j \alpha_j y_j \mathbf{x}_j)] - \sum_i \alpha_i y_i b + \sum_i \alpha_i \quad (4.12)$$

A questo punto, cambiando l'ordine delle sommatorie nel secondo membro e notando che il penultimo addendo è sempre nullo (vedi equazione (4.11)), si ottiene la versione finale della lagrangiana duale:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (4.13)$$

L'equazione appena riportata è un problema di programmazione quadratica (*quadratic programming*, QP) che assicura l'esistenza e l'unicità di una e una sola soluzione.

---

<sup>1</sup>Il teorema di Kunh-Tucker assicura che la soluzione di questo problema è la stessa del problema originario.

Per completare il discorso sulla SVM lineare, si consideri nuovamente la *regola di decisione* introdotta nell’equazione (4.4). Avendo ora una definizione formale per  $\mathbf{w}$ , questa può essere riscritta nel seguente modo:

$$\begin{cases} \text{Se } \sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{u} + b > 0 \text{ allora } y_u = +1 \\ \text{Se } \sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{u} + b < 0 \text{ allora } y_u = -1 \end{cases} \quad (4.14)$$

Il motivo dell’uso di questo espediente matematico risiede nel fatto che adesso sia l’ottimizzazione della lagrangiana (4.13) sia la *regola di decisione* (4.14) dipendono esclusivamente da un prodotto scalare tra due vettori,  $\mathbf{x}_i \cdot \mathbf{x}_j$  e  $\mathbf{x}_i \cdot \mathbf{u}$  rispettivamente, e questo semplificherà notevolmente la trattazione della SVM non lineare.

#### 4.1.2 Soft Margin SVM

Prima di introdurre l’estensione della SVM che permetta la classificazione non-lineare, è interessante discutere di come sia possibile usare una SVM lineare anche per situazioni in cui il training set non sia linearmente separabile.

Si prenda, come esempio, la situazione proposta di seguito (Figura 4.3).

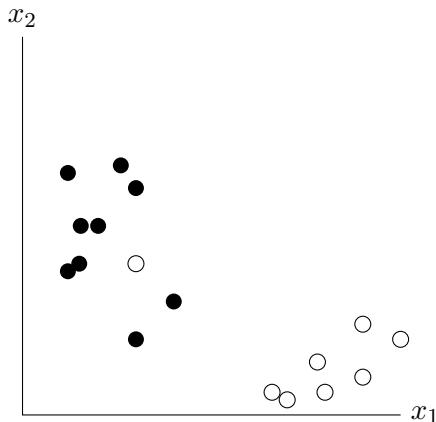


Figura 4.3. Dataset non linearmente separabile

Si nota chiaramente come, in questo caso, non esista alcun iperpiano separatore. Nonostante ciò, modificando leggermente il modello di SVM introdotto fino a questo punto, si

può dimostrare che è ancora possibile utilizzare una SVM lineare.

La chiave di questo nuovo modello sta nell'introduzione di una *variabile slack*, in modo tale da rilassare quei vincoli rigidi che non permetterebbero l'uso di una SVM lineare. Precedentemente, i margini erano definiti dei vincoli:

$$y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 \quad y_i \in \{-1,1\} \quad (4.15)$$

Nella soluzione proposta, invece, sono definiti da:

$$y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i \quad \xi_i \geq 0, y_i \in \{-1,1\} \quad (4.16)$$

La situazione che si è andata a definire, graficamente, appare in Figura 4.4.

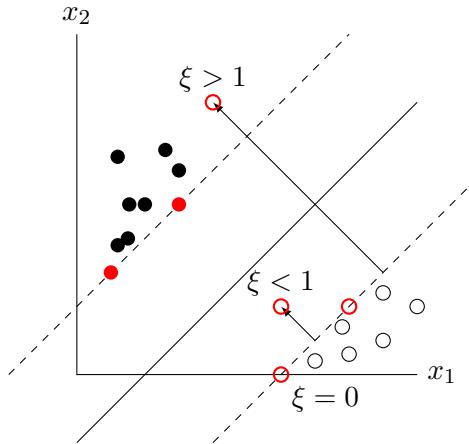


Figura 4.4. Esempio di soft margin

Il nuovo vincolo permette al margine funzionale di essere minore di 1. L'errore che si commette è, però,  $C\xi_i$ , sia per i punti che ricadrebbero nella parte corretta dell'iperpiano separatore ( $0 < \xi_i \leq 1$ ), sia per quelli che sarebbero nel lato sbagliato ( $\xi > 1$ ). Si sono, quindi, "rilassati" i vincoli in modo tale da classificare dati non-separabili, con una penalità linearmente proporzionale all'entità dell'errore commesso sul training set.

Il nuovo problema di ottimizzazione è il seguente:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \quad (4.17)$$

$$\text{vincolato a} \quad y_i (\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i \quad (4.18)$$

$$\xi_i \geq 0, \quad \forall i = 1, \dots, N \quad (4.19)$$

La costante  $C$  (con  $C \geq 0$ ) gestisce il compromesso fra la minimizzazione dei due contributi alla funzione obiettivo: se  $C = 0$  gli errori sul training set non sono penalizzati; se  $C$  è molto grande, il termine legato al margine è minoritario e gli errori sul training set sono molto penalizzati.

Allo stesso modo del caso lineare, possono essere usati i moltiplicatori di Lagrange; la funzione da ottimizzare quindi è la seguente:

$$L = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i + \sum_i \alpha_i [1 - \xi_i - y_i (\mathbf{x}_i \cdot \mathbf{w} + b)] - \sum_i \beta_i \xi_i \quad (4.20)$$

dove  $\beta_i$  sono i moltiplicatori di Lagrange necessari per vincolare la positività della *variabile slack*.

È interessante notare come l'applicazione di una soft-margin SVM possa avere prestazioni migliori della hard-margin SVM anche per dataset linearmente separabili. Per chiarificare il motivo, un esempio è riportato nella figura successiva: il training set permetterebbe l'uso di una hard-margin SVM, ma il risultato avrebbe un margine piuttosto limitato, confrontato con quello che si otterrebbe in caso di soft-margin SVM.

## 4.2 SVM non lineare per classificazione binaria

Nonostante l'introduzione della variante Soft-Margin della SVM possa il qualche caso tornare utile, può capitare che anche questa strategia sia difficilmente applicabile o produca

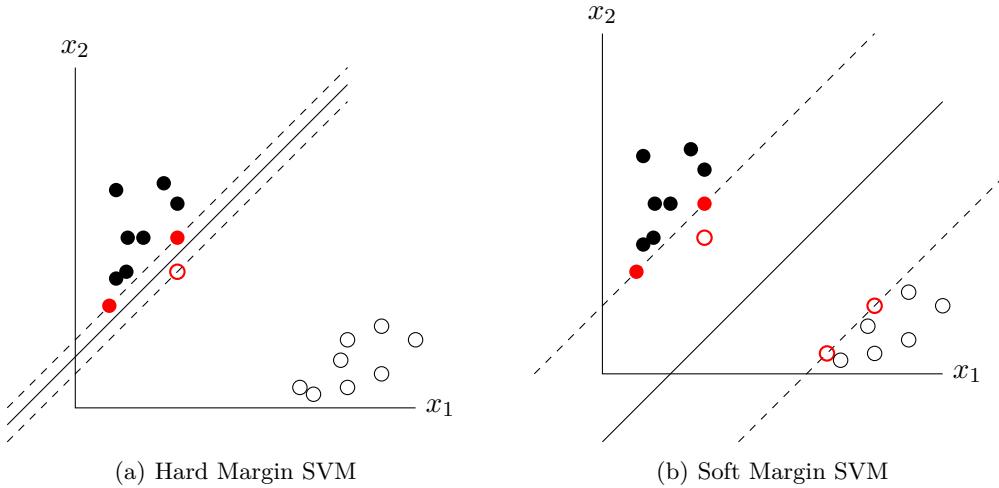


Figura 4.5. Confronto di strategia tra Hard Margin SVM e Soft Margin SVM

risultati non soddisfacenti (si rimanda al Capitolo X dove si discuterà come avere una stima delle prestazioni di un classificatore).

L'intento della classificazione SVM non-lineare è quello di trasformare lo spazio dei vettori di training non linearmente separabili  $\mathbb{R}^d$  in uno spazio  $\mathcal{H}$  (la cui dimensionalità sarà maggiore di  $d$  o anche, potenzialmente, infinita) in cui i campioni siano disposti in modo tale da permettere l'utilizzo di una SVM lineare. Questo passo si giustifica tramite il *Teorema di Cover sulla separabilità*, il quale afferma che un problema di classificazione complesso, formulato attraverso una trasformazione non-lineare dei dati in uno spazio ad alta dimensionalità, ha maggiore probabilità di essere linearmente separabile che in uno spazio a bassa dimensionalità.

Si ricerca quindi una funzione di trasformazione:

$$\Phi : \mathbb{R}^d \rightarrow \mathcal{H} \quad (4.21)$$

Graficamente, accade una situazione del genere (Figura 4.6).

Alla luce di ciò, la lagrangiana (4.13), per la fase di addestramento, e la *regola di decisione*

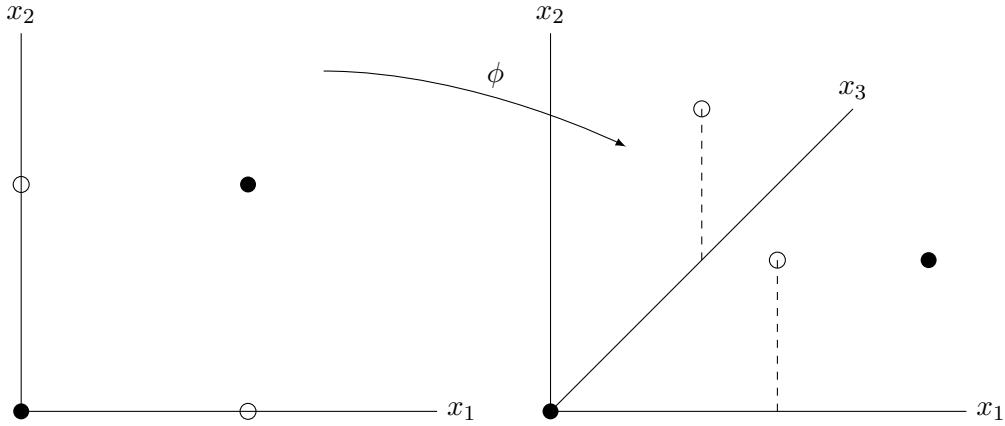


Figura 4.6. Situazione esemplificativa della funzione di trasformazione

(4.14), per la fase di classificazione, possono essere riscritte come segue:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \quad (4.22)$$

$$f = \sum_i \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{u}) + b \quad (4.23)$$

Come già accennato in precedenza, operativamente, entrambe le due fasi sono caratterizzate solo dal prodotto scalare dei vettori trasformati; questo suggerisce che non sia necessaria la conoscenza di  $\Phi$ , in quanto è sufficiente definire un *kernel*:

$$K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \quad \text{tale che} \quad K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \quad (4.24)$$

Esistono condizioni necessari e sufficienti affinché una data funzione  $K$  sia un *kernel*, ovvero affinchè esistano uno spazio  $\mathcal{H}$  e una funzione di trasformazione  $\Phi$ , tali che in  $\mathcal{H}$  sia definito il prodotto scalare.

Le condizioni di Mercer sono condizioni necessarie e sufficienti per definire per quali famiglie di *kernel*  $K$  esiste la coppia  $\{\mathcal{H}, \Phi\}$  con le proprietà presentate precedentemente.

**Teorema 1** Sia  $K(\mathbf{x}, \mathbf{y})$  un kernel continuo che può essere espanso nella serie

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \Phi(\mathbf{x})_i \cdot \Phi(\mathbf{y})_i \quad (4.25)$$

Affinché tale espansione sia valida e per la sua convergenza assoluta, è necessario e sufficiente che la condizione

$$\int \int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0 \quad (4.26)$$

sia vera per ogni  $g(\cdot)$  che soddisfano

$$\int g^2(\mathbf{x}) d\mathbf{x} < \infty \quad (4.27)$$

Questo espediente matematico è detto **kernel trick** in quanto, dato un kernel che soddisfi tali condizioni, un classificatore SVM risulta identificato dal kernel stesso, senza alcuna necessità di definire esplicitamente né  $\Phi$  né  $\mathcal{H}$ .

Per completezza, vengono riportati ora i problemi di ottimizzazione della lagrangiana e la regola di decisione usando il *kernel trick*:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (4.28)$$

$$f = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{u}) + b \quad (4.29)$$

Alcuni dei *kernel* più famosi e usati sono i seguenti:

- **Lineare:**  $K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$
- **Polinomiale:**  $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^d$  con  $d > 0$
- **Radiale Gaussiano (RBF):**  $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$  il quale introduce una dimensionalità dello spazio delle *feature* trasformate infinita.

### 4.3 SVM multiclassse

Un problema con più classi si affronta mediante SVM esprimendolo come combinazione di problemi binari. Assumendo di operare con  $C$  classi, tale operazione si effettua tipicamente in due modi possibili:

- l'approccio *one-against-one* (OAO) calcola mediante SVM binaria una *regola di decisione*  $f_{ij}$  per ogni coppia di classi, per un totale di  $C(C-1)/2$  funzioni discriminanti, e classifica un campione incognito  $\mathbf{u} \in \mathbb{R}^d$  mediante "votazione" (se  $f_{ij} > 0$  si dà un voto alla classe  $i$ -esima, altrimenti alla  $j$ -esima; si assegna infine  $\mathbf{u}$  alla classe che ha ricevuto più voti);
- l'approccio *one-against-all* (OAA) calcola mediante SVM binaria una *regola di decisione*  $f_i$  per ogni decisione del tipo "classe  $i$  contro classe non- $i$ ", per un totale di  $C$  funzioni discriminanti e assegna  $\mathbf{u} \in \mathbb{R}^d$  alla classe  $i$ -esima se  $f_i(\mathbf{u}) \geq f_j(\mathbf{u})$  per ogni  $j = 1, \dots, C$  con  $i \neq j$ .

# Bibliografia

- [1] CHEN, Q., AND GONG, P. Automatic variogram parameter extraction for textural classification of the panchromatic ikonos imagery. *IEEE Transactions on Geoscience and Remote Sensing* 42, 5 (May 2004), 1106–1115.
- [2] DALAL, N., AND TRIGGS, B. Histograms of oriented gradientes for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2005).
- [3] GHAMINI, P., MURA, M. D., AND BENEDIKTSSON, J. A. A survey on spectral-spatial classification techniques based on attribute profiles. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (May 2015), 2335–2353.
- [4] KUO, B.-C., AND LANDGREBE, D. A. Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing* 42, 5 (May 2004), 1096–1105.
- [5] MOSER, G., SERPICO, S., AND BENEDIKTSSON, J. A. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *IEEE* 101, 3 (March 2013), 631–651.
- [6] PESARESI, M., AND BENEDIKTSSON, J. A. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience*

## BIBLIOGRAFIA

---

and *Remote Sensing* 39, 2 (February 2001), 309–320.

- [7] TORRIONE, P. A., MORTON, K. D., SAKAGUCHI, R., AND COLLINS, L. M. Histograms of oriented gradientes for landmine detection in ground-penetrating radar data. *IEEE Transactions on Geoscience and Remote Sensing* 52, 3 (March 2014), 1539–1550.