POLITECNICO DI TORINO

Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea Magistrale

Design di un trading system azionario di tipo trend reversal basato su modelli di Machine Learning

Relatori

Prof. Luca CAGLIERO

Dott. Giuseppe ATTANASIO

Laureando

Federico CAREGLIO

253288

Anno Accademico 2019 - 2020

Ringraziamenti

Vorrei ringraziare i miei relatori, il professor L. Cagliero e il dottorando G. Attanasio, per avermi guidato nella stesura di questo lavoro sperimentale con gentilezza e disponibilità.

Un pensiero speciale va alla mia famiglia, le mie certezze, che con grande amore e sacrifici sono stati sempre presenti al mio fianco.

Ringrazio i miei migliori amici, i miei 'bro', per essere sempre stati presenti in questi anni e avermi dimostrato quanto valga la frase "chi trova un amico trova un tesoro".

Ringrazio la mia ragazza con cui ho condiviso avversità e successi e che è sempre stata al mio fianco sopportandomi e supportandomi nelle scelte.

Infine un grazie va anche alle splendide persone che ho conosciuto in questi anni e che mi sono state vicino.

Indice

Elenco delle tabelle 6

Elenco delle figure 8

Elenco degli algoritmi 10

Acronimi 11

1 Introduzione 14

2 Introduzione ai trading systems 18
2.1 I mercati finanziari 19
2.2 Strategie di trading 21
2.3 Approcci basati su trend recognition 23

3 Introduzione al Data Mining 24
3.1 Support Vector Machines 28
3.2 K Nearest Neighbors 31
3.3 Classificatori Bayesiani 31
3.4 Random Forest Classification 33
3.5 Multilayer Perceptron 34
3.6 Majority Voting 36

4 Studio della letteratura 37

5 Metodologia presentata 42
5.1 Raccolta dati e calcolo degli indicatori 46
5.2 Preparazione dei dati 58
5.3 Riconoscimento automatico di un trend in corso 59

5.3.1 Consecutive 60
5.3.2 SMA 61
5.3.3 MACD 64
5.3.4 Filtro sul Volume Scambiato per Azione 64
5.4 Addestramento di un classificatore 65
5.5 Previsione di un'inversione di trend 66
5.6 Gestione del trade 72

6 Esperimenti 77
6.1 Configurazione degli algoritmi 77
6.2 Risultati 79
6.2.1 Analisi della significatività statistica 80
6.2.2 Confronto tra le strategie di identificazione del trend reversal trigger 82
6.2.3 Confronto tra le combinazioni di w e n per la strategia volume consecutive 87
6.2.4 Confronto tra i classificatori e le features per la configurazione volume consecutive n3 w5 92

7 Conclusioni e Lavori Futuri 106

Bibliografia 109

Elenco delle tabelle

4.1 Tabella riassuntiva della Letteratura 40

5.1 Elenco dei dataset 46
5.2 Struttura di un dataset $s_i y_j$ 46
5.3 Elenco dei descrittori utilizzati 57
5.4 Dataset utilizzato come input del modulo di trading 72

6.1 Confronto tra le strategie sul *prt* medio calcolato su 7 anni. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata 83
6.2 Confronto tra le strategie sul *rmia* medio calcolato su 7 anni. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata 85
6.3 Tabella riassuntiva dei valori medi relativi alle strategie su 7 anni e delle relative deviazioni standard (indicate tra parentesi) 86
6.4 Confronto tra le combinazioni di n e w sul *prt* medio calcolato su 7 anni per la strategia volume consecutive. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata . . 88
6.5 Confronto tra le combinazioni di n e w sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive 90
6.6 Tabella riassuntiva dei valori medi assunti dalle combinazioni di n e w della strategia volume consecutive su 7 anni e delle relative deviazioni standard (indicate tra parentesi) 91

6.7 Confronto tra i classificatori sul *prt* medio calcolato su 7 anni per la strategia volume consecutive n3 w5. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata . . 93
6.8 Confronto tra la features sul *prt* medio calcolato su 7 anni per la strategia volume consecutive n3 w5 95
6.9 Confronto tra i classificatori sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive n3 w5 96
6.10 Confronto tra le features sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive n3 w5 97
6.11 Confronto tra le combinazioni di features e classificatori sul *prt* medio calcolato su 7 anni per la strategia volume consecutive n3 w5 101
6.12 Confronto tra le combinazioni di features e classificatori sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive n3 w5 102
6.13 Tabella riassuntiva dei valori medi assunti dai classificatori della configurazione volume consecutive n3 w5 su 7 anni e delle relative deviazioni standard (indicate tra parentesi) . . 104
6.14 Tabella riassuntiva dei valori medi assunti dalle features della configurazione volume consecutive n3 w5 su 7 anni e delle relative deviazioni standard (indicate tra parentesi) 105

Elenco delle figure

3.1 Processo di Knowledge Discovery in Databases [8] 24
3.2 Matrice di Confusione per un problema binario. [8] 27
3.3 Possibili hyperplanes per un problema lineare [8] 29
3.4 Hyperplane con i margini minimi e massimi [8] 29
3.5 Raffigurazione di un modello non lineare e del suo corrispettivo lineare [8] 30
3.6 Tre diversi k nearest neighbors (con k = 1, 2 e 3) [8] 32
3.7 Struttura di un albero di decisione [9] 33
3.8 Schema di un percettrone [10] 34
3.9 Schema di una rete neurale feed forward a più layer [8] . . . 35

5.1 Architettura del sistema di trading proposto 45
5.2 Incroci di due medie mobili a breve e lungo periodo [42] . . . 62
5.3 Strategia consecutive 69
5.4 Strategie basate su medie mobili 70
5.5 Filtro sul volume 71
5.6 Andamento del prezzo di un'azione s 73
5.7 Processo di trading dato un giorno d_t e un'azione $s(d_t)$. . . . 74

6.1 Andamento del *prt* medio delle strategie su 7 anni 83
6.2 Andamento del *rmia* medio delle strategie su 7 anni 84
6.3 Andamento del *prt* medio delle combinazioni di n e w su 7 anni per la strategia volume consecutive 87
6.4 Andamento del *rmia* medio delle combinazioni di n e w su 7 anni per la strategia volume consecutive 89
6.5 Andamento del *prt* medio dei classificatori su 7 anni per configurazione volume consecutive n3 w5 92
6.6 Andamento del *prt* medio delle features su 7 anni per la configurazione volume consecutive n3 w5 94

6.7 Andamento del *rmia* medio dei i classificatori su 7 anni per la configurazione volume consecutive n3 w5 96
6.8 Andamento del *rmia* medio delle features su 7 anni per la configurazione volume consecutive n3 w5 97
6.9 Andamento del *prt* medio generale dei classificatori su 7 anni 99
6.10 Andamento del *rmia* medio generale dei classificatori su 7 anni 99
6.11 Andamento del *prt* medio generale delle features su 7 anni . 100
6.12 Andamento del *rmia* medio generale delle features su 7 anni 100
6.13 Andamento del *prt* su 7 anni delle configurazioni volume consecutive n3 w5 SVC OSC+VOL e volume consecutive n3 w5 MNB OSC+VOL 103
6.14 Andamento del *rmia* su 7 anni delle configurazioni volume consecutive n3 w5 SVC OSC+VOL e volume consecutive n3 w5 MNB OSC+VOL 103

Elenco degli Algoritmi

1 Preparazione dei dati. Riceve in ingresso il dataframe descrivente una specifica azione in un determinato anno con tutti i descrittori citati nella sezione precedente. 59
2 Strategia Consecutive. Utilizzata per individuare un trend consecutivo e quindi individuare un trend reversal trigger. . 61
3 Strategia SMA. Utilizzata per individuare un incrocio di due medie mobili e conseguentemente un possibile trend reversal. 63
4 Filtro sul volume utilizzato in combinazione con ciascuna delle altre strategie. 65
5 Comportamento del classificatore 66
6 Majority Voting e creazione etichette composte 68
7 Gestione del trading automatica. 76

Acronimi

ANN

Artificial Neural Network

GNB

Gaussian Naïve Bayes

KNN

K Nearest Neighbor

MACD

Moving Average Convergence Divergence

ML

Machine Learning

MLP

Multilayer Perceptron

MNB

Multinomial Naïve Bayes

MV

Majority Voting

NN

Neural Network

RF

Random Forest

SMA

Simple Moving Average

SVM

Support Vector Machine

Capitolo 1

Introduzione

I mercati finanziari consentono la compravendita (trading) di strumenti finanziari. Esistono diversi tipi di mercati (bond, forex, derivati, ...), ma il più noto è probabilmente quello azionario in cui si scambiano quote di società quotate.

Il trading fondamentalmente può essere classificato in: (i) discrezionale, se basato interamente sulle competenze, abilità ed esperienza del trader ad analizzare i grafici e gli indicatori di mercato senza l'utilizzo di supporti automatici; (ii) quantitativo, se basato invece sull'utilizzo di un software che è in grado di analizzare i dati di mercato e di generare automaticamente i segnali di acquisto e di vendita, aprendo in automatico le posizioni per conto dell'investitore. Quest'ultima tipologia di trading si basa su algoritmi deterministici che operano in base a specifiche regole.

Grazie alla disponibilità di grandi moli di dati storici relativi ai mercati finanziari, una promettente direzione di ricerca è l'utilizzo di tecniche di analisi dei dati per definire le regole con cui i sistemi di trading investono sul mercato azionario in modo automatico. I processi predittivi si basano sull'analisi di dati storici mediante algoritmi di Machine Learning. Essi includono una vasta scelta di algoritmi di classificazione e regressione, finalizzati a predire il valore di una variabile target.

Nell'ambito dell'analisi di dati finanziari, l'addestramento dei modelli predittivi può basarsi su (i) l' analisi tecnica, ovvero lo studio dei mercati attraverso indicatori, oscillatori statistici e grafici che modellano comportamenti ripetitivi nell'andamento dei prezzi storici, (ii) l'analisi fondamentale, ovvero lo studio dei mercati attraverso gli aspetti economico-finanziari e (iii) il news trading, ovvero l'insieme di opinioni di esperti attraverso le notizie in

rete. A seconda dell'orizzonte con cui le posizioni vengono aperte, il trading si può dividere ulteriormente in intraday se le posizioni rimangono aperte per massimo un giorno, mentre si definisce multiday se le posizioni rimangono aperte per un periodo più lungo.

Le strategie più comuni basate sull'analisi dei dati mirano a identificare un trend nelle serie storiche dei dati. Un trend indica una tendenza dei prezzi a seguire la stessa direzione nel mercato. Si parla di uptrend se il prezzo cresce, di downtrend se il prezzo cala. Le strategie basate sul trend possono essere: (i) trend following, che si basa sull'idea di cavalcare l'onda del trend scommettendo quindi sulla sua continuità, oppure (ii) trend reversal, che scommette sulla fine di un trend e di conseguenza sulla sua inversione.

L'obiettivo di questa tesi è proporre un sistema di trading quantitativo che investe sul mercato azionario mediante una strategia multiday di tipo trend reversal basata sull'addestramento di algoritmi di Machine Learning. Il contributo principale del lavoro di tesi è l'estensione di una versione preliminare del sistema secondo varie direzioni che includono (i) l'analisi di diverse strategie per il riconoscimento del trend reversal, (ii) l'esplorazione più approfondita dell'impatto delle tecniche di classificazione e della loro configurazione, (iii) la valutazione delle performance del sistema rispetto a un ampio set di dati storici, (iv) l'analisi della profittabilità del sistema in termini di equity prodotte e (v) l'analisi statistica della significatività dei miglioramenti di prestazioni ottenuti rispetto alla versione precedente.

Il trading system quantitativo include i seguenti passi:

1. **Raccolta dati e calcolo degli indicatori**: si raccolgono dati relativi alle azioni dell'indice americano *Standard & Poor 500*, su base giornaliera, secondo diversi tipi di descrittori basati sul prezzo, analisi tecnica e news sentiment.

2. **Preparazione dei dati**: prima di essere passati al classificatore i dati vengono (i) normalizzati, ovvero trasformati in modo tale che assumano valori compresi tra 0 e 1 al fine di renderli omogenei e confrontabili, (ii) filtrati, in modo da tenere solo i descrittori relativi alla strategia scelta, e infine (iii) scalati, al fine di avere, per ciascun descrittore, il valore ad esso relativo per il giorno corrente fino a N giorni precedenti (con N parametro del sistema) .

3. **Riconoscimento automatico di un trend in corso**: si ricerca un trigger che identifichi una possibile inversione di trend attraverso l'uso di diverse strategie quali:

   (i) l'utilizzo di una sliding window, ovvero una finestra di W giorni nei quali la direzione del prezzo dell'azione deve essere concorde. Se la direzione trovata è 'B' (buy), ovvero W giorni in cui il prezzo dell'azione sale, si scommette su un downtrend, se invece la direzione trovata è 'S' (sell) si scommette su un uptrend.

   (ii) l'utilizzo dell'incrocio di medie mobili quali SMA e MACD. Se il risultato è negativo si scommette su un downtrend, se è positivo si scommette su un uptrend.

   (iii) la combinazione di ciascuna delle precedenti strategie con un filtro sul volume scambiato per azione. Questo filtro non è altro che la differenza tra il volume scambiato nel giorno corrente con la media del volume scambiato nei W giorni precedenti. Per poter procedere il risultato deve essere concorde con il reversal ipotizzato dalla strategia applicata, per cui se è stato ipotizzato un downtrend il risultato del filtro deve essere negativo, se invece è stato ipotizzato un uptrend il risultato deve essere positivo.

   Nel caso un trigger venga individuato, si memorizza la direzione del reversal ipotizzato in una variabile target, contenente quindi 'S' nel caso si ipotizzi un downtrend, 'B' nel caso contrario.

4. **Addestramento di un classificatore**: il classificatore viene addestrato al fine di predire la direzione (buy 'B', hold 'H', sell 'S') del prezzo di un'azione fino a N giorni in avanti rispetto al corrente. Per esempio, immaginando $N = 5$, il classificatore può predire una finestra di etichette del tipo [B, B, S, H, B].

5. **Previsione di un'inversione di trend**: in questo step si verifica se ci sia una corrispondenza tra la direzione del target ipotizzato e le etichette predette del classificatore. Questo avviene tramite l'uso del majority voting, una tecnica che restituisce, come risultato, la direzione presente in maggioranza tra le etichette predette (B nel caso dell'esempio precedente). Se il risultato del majority voting è concorde con il target ipotizzato precedentemente (le due etichette devono essere uguali) si procede con il modulo di trading.

6. **Gestione del trade**: questo modulo racchiude tutta la logica di gestione del trade come (i) chiudere le posizioni già aperte se un target non è più presente o se le perdite hanno superato una soglia di sicurezza definita come *stop loss*, (ii) aprire nuove posizioni del tipo indicato dal target quindi long-selling in caso di uptrend, short-selling in caso di downtrend e (iii) aggiornare il budget di mercato.

E' stata condotta una campagna sperimentale su un campione di 7 anni, dal 2011 al 2017, di azioni dell'indice azionario americano Standard&Poor 500. L'obiettivo è quello di dimostrare l'efficacia dell'approccio basato su Machine Learning rispetto ad un approccio tradizionale basato su analisi tecnica e definire le configurazioni più appropriate del sistema analizzando l'impatto di vari fattori sui risultati della simulazione trading.

I risultati hanno dimostrato che la strategia basata sul Machine Learning impiegata nell'individuazione di un reversal tramite l'utilizzo di un filtro sul volume scambiato per azione, applicato a una finestra di W giorni consecutivi nei quali la direzione del prezzo dell'azione è concorde, sia la più performante in termini di profitto relativo e profitto medio per operazione.

Capitolo 2

Introduzione ai trading systems

Un sistema di trading rappresenta un insieme di regole predefinite, implementate da algoritmi sviluppati da informatici e matematici, che permettono, in maniera completamente automatica, di definire strategie di entrata, di uscita e di gestione del denaro, di eseguire e monitorare le negoziazioni e di determinare quando e quali strumenti di mercato dovrebbero essere scambiati al fine di generare una strategia che garantisca profitto all'investitore .

Il vantaggio principale nell'operare attraverso un sistema di trading è rappresentato dall'eliminazione, per la maggior parte, della componente emotiva dell'utente che può influire negativamente sulle strategie di mercato . L'emotività infatti può condizionare lo stato d'animo del trader attraverso stati di stress, paura ed euforia che non gli permetterebbero una totale lucidità di pensiero cosa che invece un sistema automatico garantisce. Questi sistemi infatti operano scelte sul mercato senza essere soggetti ad alcun tipo di emozione, sentimento o pregiudizio. Altri vantaggi risiedono nella possibilità (i) di testare le strategie di mercato su dati storici passati al fine di valutarne l'efficacia (*backtesting*), (ii) di poter gestire le negoziazioni in maniera rapida, veloce e di monitorarne l'andamento e (iii) di poter aprire contemporaneamente più posizioni anche in mercati diversi.

Gli svantaggi principali invece risiedono (i) nella possibilità che i sistemi risultino inaffidabili, (ii) nel fatto che la maggior parte di questi sistemi decida come operare sul mercato basandosi su regole empiriche non validate su un

set affidabile di dati, (iii) nella necessità di un costante monitoraggio del sistema, (iv) nella possibilità che, nonostante il sistema sia stato impostato correttamente questo non rispetti le aspettative e (v) nella possibilità che il sistema risulti troppo ottimizzato perdendo generalità ($overfitting$) .

2.1 I mercati finanziari

Con il termine $mercato\ finanziario$ si intende un luogo virtuale all'interno del quale avviene la compravendita (trading) di strumenti finanziari (azioni, quote, obbligazioni, ...), indirizzati al breve, medio e lungo periodo [1]. Quando si parla di mercato bisogna distinguere innanzitutto tra :

- **Mercato primario**: dove vengono acquistati titoli appena emessi. In questa prima fase, un titolo finanziario viene creato e progressivamente inizierà ad assumere importanza fino ad entrare nel mercato secondario.

- **Mercato secondario**: dove vengono acquistati e venduti, spesso tramite l'aiuto di intermediari (security brokers), titoli già sottoscritti e in circolazione. I mercati finanziari fanno parte di questa categoria. Questo tipo di mercato offre la possibilità sia ai compratori che agli acquirenti di avere sempre a disposizione le informazioni sui propri investimenti e di trasformare molto velocemente i propri strumenti finanziari in denaro liquido.

I mercati si differenziano soprattutto sulla base della tipologia di strumenti finanziari scambiati [2] . Si possono distinguere i seguenti tipi:

- **Money Markets** (mercati monetari): sono mercati in cui avviene la negoziazione di titoli con scadenze a breve termine, di massimo un anno. Sono molto sicuri e caratterizzati da un ritorno d'investimento relativamente basso. A causa della natura a breve termine di questi fondi, la loro fluttuazione di prezzo sul mercato è relativamente piccola.

- **Capital Markets** (mercati di capitali): sono mercati in cui avviene la negoziazione di titoli con scadenze superiori a un anno. Solitamente i maggiori fornitori di questi titoli sono società e governi. Differentemente dal precedente, la fluttuazione di prezzo di questi titoli può variare di molto durante il mercato.

- **Over-the-Counter Markets**: è un tipo di mercato che non ha luogo fisico ed è per questo decentralizzato. Il trading in questo tipo di mercato è condotto elettronicamente (via telefono, computer, ..) e nel quale i due agenti di una transazione negoziano senza un tramite, come potrebbe essere un broker. Solitamente in un mercato OTC avviene lo scambio di titoli non quotati nei maggiori indici di mercato e per cui appartenenti per la maggior parte a piccole società.

- **Bond Markets**: è un mercato in cui avviene la vendita di bond. Un bond, o obbligazione, è un titolo su cui un agente investitore presta del denaro ad un tasso di interesse prestabilito per un determinato periodo. I bond sono di solito emessi da società, comuni, governi e stati come finanziamento.

- **Derivatives Markets**: è un mercato di negoziazione di derivati. Un derivato è un titolo secondario il cui valore è appunto derivato dal relativo titolo primario.

- **Forex Markets**: è un mercato in cui si negoziano le valute. E' un mercato liquido poiché al posto di titoli si negoziano direttamente i soldi. Come per il mercato OTC anche questo è decentralizzato.

- **Mortages Markets**: mercato che negozia i mutui come strumento finanziario. Differisce dagli altri mercati dei capitali principalmente perché i mutuatari sono privati e non stati o istituzioni. I prestiti per i mutui vengono rilasciati con scadenze e importi variabili a seconda delle necessità dei richiedenti.

- **Stock Markets**: mercato in cui avviene la compravendita di azioni di società quotate in borsa. Le aziende mettono a disposizione le proprie azioni per essere acquistate dagli investitori i quali, possedendo una quota delle azioni, diventano in un certo senso, proprietari di una parte dell'azienda garantendosi dei diritti sulla stessa. Gli investitori possono ottenere un profitto se il prezzo delle azioni aumenta nel tempo o se l'azienda paga i dividendi degli azionisti. Principalmente le azioni servono quindi come strumento per finanziare le aziende.

2.2 Strategie di trading

Con l'avanzare di Internet il trading online ha sempre più preso piede offrendo agli utenti la possibilità di fare compravendita di strumenti finanziari tramite il web. Ciò comporta diversi vantaggi come i minori costi di commissioni per gli utenti e la possibilità di essere sempre aggiornati con gli ultimi valori di mercato grazie a grafici e indicatori. Il trading online viene effettuato tramite piattaforme di trading messe a disposizioni da società chiamate broker che consentono di operare sui mercati finanziari dopo l'apertura di un conto di trading. Queste piattaforme solitamente offrono l'aggiornamento dei dati in tempo reale e forniscono informazioni derivanti dall'analisi tecnica e fondamentale per operare le proprie scelte. Tra queste possiamo trovare *Trade.com* [3] (sicura e affidabile), *Plus500* [4] (per professionisti), *eToro* [5] e altre.

Il trading generalmente si può dividere in due tipologie di investimento:

- **Trading Discrezionale**: si basa unicamente sulle competenze, abilità ed esperienza del trader ad analizzare i grafici e gli indicatori di mercato senza il supporto di un software automatico che determina le scelte di investimento.

- **Trading Quantitativo**: si basa invece sull'utilizzo di sistemi automatici e algoritmi in grado di analizzare quantitativamente i dati di mercato generando segnali di acquisto e di vendita che il trader può scegliere di seguire. Molto spesso questi sistemi sono anche in grado di aprire e gestire in automatico le posizioni di mercato per conto degli investitori.

I dati di mercato sui quali trader e sistemi fanno affidamento si possono dividere in diverse categorie [6]:

1. **Analisi Tecnica**: ovvero lo studio dei mercati attraverso indicatori, oscillatori statistici e grafici che modellano comportamenti ripetitivi nell'andamento dei prezzi storici di un'azione.

2. **Analisi Fondamentale**: ovvero lo studio di tutti gli aspetti economici che possono influenzare l'andamento del mercato come le performance di una azienda, il livello di disoccupazione, le condizioni di un dato settore e molto altro.

3. **News Sentiment**: ovvero l'insieme di opinioni di esperti e analisti rilasciate su piattaforme online quali siti di informazione, blog e social networks.

Il trading inoltre si può dividere ulteriormente in base agli orizzonti temporali per cui un'operazione sul mercato viene effettuata [1]. Si divide in:

- **Scalper**: rappresenta la strategia in cui le posizioni vengono aperte e chiuse in un intervallo di tempo molto breve dell'ordine di pochi minuti o addirittura di secondi. Lo scalper quindi non cerca di sfruttare un trend di mercato bensì mira ad ottenere tanti piccoli profitti cercando di limitare le perdite il più possibile.

- **Intraday**: rappresenta la strategia di trading più diffusa in cui le posizioni vengono aperte e chiuse nell'arco della stessa giornata sfruttando a proprio vantaggio i movimenti dei prezzi giornalieri. Questo porta il trader ad avere un buon controllo sulle perdite ma nello stesso tempo a non massimizzare i profitti.

- **Multiday**: seguendo questa strategia le posizioni vengono aperte e mantenute tali per un periodo che può variare da diversi giorni ad addirittura mesi. La strategia si basa sul concetto che 'la storia si ripete' e tendenze di mercato verificatesi in passato possono ripetersi in futuro. Non è una strategia semplice in quanto risulta molto difficile per un multiday trader riuscire a prevedere come sarà il mercato molto più avanti nel tempo. E' infatti per questo motivo che i multiday trader cercano di identificare e seguire i trend in corso, ignorando le oscillazioni a brevi periodi. Seguendo questa tecnica i profitti possono essere molto alti così come le perdite.

Più la strategia dura e più alti potenzialmente sono i rendimenti, così come le perdite. Al contrario però il monitoraggio della posizione e di tutto ciò che la concerne diventa più difficoltoso.

2.3 Approcci basati su trend recognition

Tra le varie strategie adottate per il trading si distinguono quelle basate sull'identificazione di un trend. Quest'ultimo indica una tendenza dei prezzi a seguire la stessa direzione nel mercato. Si parla di uptrend se il prezzo cresce, di downtrend se il prezzo cala. Il trader che investe su questa strategia mira di più ad analizzare il prezzo di una azione nel giorno corrente, per sapere se è in un trend o meno, piuttosto che a predire il prezzo che avrà la stessa in futuro. In queste strategie risultano particolarmente utili gli studi di indicatori tecnici come le medie mobili. Si distinguono le strategie:

- **Trend following**: ovvero individuare un trend continuo, che sia rialzista (up) o ribassista (down), e scommettere sulla sua continuità.
- **Trend reversal**: ovvero riuscire ad identificare la fine di un trend e di conseguenza scommettere sulla sua inversione.

L'obiettivo di questa tesi è di proporre un sistema di trading quantitativo che investe sul mercato azionario mediante una strategia multiday di tipo trend reversal basata sull'addestramento di algoritmi di machine learning con dati a granularità giornaliera.

Capitolo 3

Introduzione al Data Mining

I dati sono una grande fonte di conoscenza e sapere. Più gli anni passano e più dati si accumulano. Tuttavia ricavare delle informazioni utili è risultato essere una delle sfide più impegnative. Il *Data Mining* è un processo che unisce i metodi tradizionali di analisi con dei complessi algoritmi per estrarre informazioni da grandi quantità di dati al fine di identificare dei modelli che li descrivano [7]. E' anche utilizzato per predire l'esito di una osservazione futura.

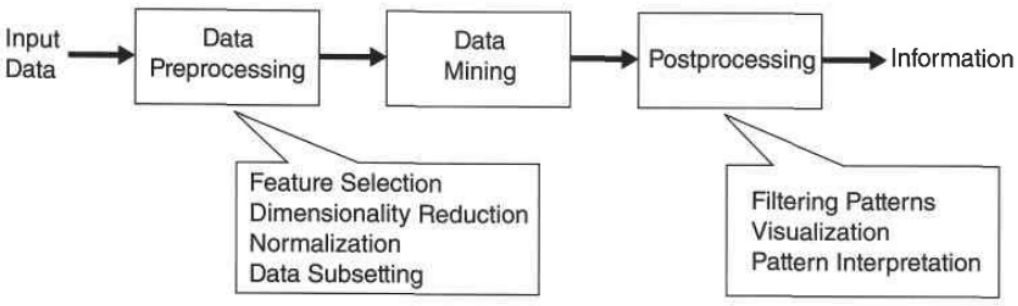

Figura 3.1: Processo di Knowledge Discovery in Databases [8]

Il data mining è una parte di un processo chiamato *knowledge discovery of databases* che porta i dati ad essere analizzati e si compone di diverse fasi, visibili in figura 3.1:

- *Raccolta dei dati grezzi*, provenienti da diverse fonti e salvati su piattaforme centralizzate o distribuite.

- *Data Preprocessing*, ovvero la trasformazione dei dati grezzi in un formato che permetta di essere studiato dall'analisi seguente. In questa fase trovano spazio la pulizia dei dati, l'aggregazione di dati provenienti da diverse fonti, la rimozione di dati duplicati ecc. E' la parte che di solito richiede più lavoro e tempo.

- *Applicazione di algoritmi supervisionati e non*, al fine di ricavare informazioni.

- *Postprocessing*, ovvero un ultimo filtraggio per garantire che solamente i dati funzionali vengano mantenuti.

Il data mining si basa sull'analisi di grandi quantità di dati (*Big Data*) così estese in termini di volume, velocità e varietà da non poter essere analizzate attraverso l'uso di metodi convenzionali.

Quando si parla di Big Data non è possibile definire una dimensione di riferimento, dato che la quantità di dati prodotta aumenta esponenzialmente di anno in anno, così come la sua velocità. Nel 2001, l'analista Doug Laney ha definito il modello di crescita dei Big Data con 3 significative parole quali volume, velocità e varietà a cui, con il passare del tempo, se ne sono aggiunte altre due, ovvero veridicità e valore. Queste cosiddette 5V racchiudono i seguenti significati:

- **Volume**: la quantità di dati cresce in maniera esponenziale richiedendo capacità di elaborazione sempre più elevate.

- **Velocità**: i dati vengono prodotti con sempre maggior velocità richiedendo quindi abilità nel processarli in tempo reale per operare tempestivamente nelle scelte.

- **Varietà**: poter elaborare differenti tipologie di dato.

- **Veridicità**: i dati devono essere quanto più possibile affidabili .

- **Valore**: i dati devono avere valore per poter estrarre informazioni utili.

Il data mining, attraverso l'analisi dei Big data, deve invece affrontare delle sfide che estendono le problematiche precedenti, quali:

- **Scalabilità**: all'aumentare della quantità di dati gli algoritmi devono riuscire ad adattarsi efficacemente.
- **Alta Dimensionalità**: i sistemi devono saper gestire una moltitudine di dati composti da tanti attributi.
- **Dati Complessi**: i sistemi devono saper gestire dati eterogenei e dati invece più complessi contenenti diverse caratteristiche.
- **Distribuzione**: i dati possono essere sparsi e quindi i sistemi devono poter affrontare sfide di questo tipo.

Con il termine *machine learning* si indica quella strategia tramite cui dei modelli vengono addestrati ad imparare e a comportarsi di conseguenza, cambiando il loro comportamento sulla base di quello passato.

"Si impara quando si cambia il proprio comportamento in un modo che lo svolgimento del proprio compito sia migliore in futuro" [7].

I compiti del data mining si possono dividere in due macro categorie:

- I compiti **predittivi** devono predire il valore di un attributo partendo dallo studio e dall'analisi dei valori di molti altri.
- I compiti **descrittivi** si occupano di derivare modelli che spiegano e analizzano le relazioni tra i dati.

Tra i modelli predittivi trovano spazio i modelli di classificazione e quelli di regressione. Con il termine *Classificazione* si intende il compito di delineare una funzione, o modello, che mappa un dato attributo in ingresso ad una specifica classe in uscita. Un modello di questo tipo può essere usato per predire l'etichetta di classe di un particolare dato. Le tecniche di classificazione infatti sono le più indicate per scopi predittivi [8].

Ci sono diversi classificatori quali alberi di decisione, regole di classificazione e associazione, support vector machines, reti neurali e altri. In seguito verranno brevemente spiegati solo i classificatori utilizzati in questa tesi.

In generale ogni tecnica utilizza un algoritmo di apprendimento per identificare un modello che riesca a definire efficacemente una relazione tra gli attributi in ingresso e le loro etichette di classe cosicché, da un dato sconosciuto in ingresso, si possa riuscire ad attribuirgli un'accurata etichetta. Affinché ciò accada, i classificatori vengono addestrati tramite l'uso di un *training set* di dati, ciascuno dei quali ha già un'etichetta di classe. In seguito si fa uso del *test set* contenente invece dati con etichetta sconosciuta. Sarà compito dei classificatori capire, per ciascuno di essi, quale sia la corretta etichetta di classe da assegnare.

L'efficacia di un classificatore si valuta con il conteggio di quanti record sono stati predetti correttamente rispetto agli errori. Questi conteggi vengono inseriti in una matrice di confusione che riassume le valutazioni. La

		Predicted Class	
		+	−
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	−	f_{-+} (FP)	f_{--} (TN)

Figura 3.2: Matrice di Confusione per un problema binario. [8]

figura 3.2 mostra una matrice di confusione per un problema binario. Essa si compone di quattro tipologie di esiti quali:

- *True Positive*: sia l'etichetta predetta che quella di test sono positive.
- *False Positive*: l'etichetta predetta è positiva mentre quella di test è negativa.
- *False Negative*: l'etichetta predetta è negativa mentre quella di test è positiva.
- *True Negative*: sia l'etichetta predetta che quella di test sono negative.

Per valutare un classificatore di solito si usano metriche quali:

$$Accuratezza = \frac{\text{Numero di predizioni corrette}}{\text{Numero totale di predizioni}} \tag{3.1}$$

e di conseguenza anche:

$$Tasso\ di\ Errore = \frac{\text{Numero di predizioni errate}}{\text{Numero totale di predizioni}} \tag{3.2}$$

Non sempre queste metriche sono sufficienti. Talvolta serve sapere, tra tutti gli esiti predetti positivi, quanti effettivamente lo siano. Questa condizione viene espressa attraverso la precisione, definita come:

$$Precisione = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{3.3}$$

Altre volte invece occorre misurare quanti esiti positivi siano stati predetti rispetto al totale degli attuali positivi. Viene così definito il richiamo, espresso come:

$$Richiamo = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{3.4}$$

Infine, quando occorre avere un equilibrio tra le metriche precedenti si può ricorrere ad un'altra misura quale:

$$F1\ Score = 2 * \frac{\text{Precisione * Richiamo}}{\text{Precisione + Richiamo}} \tag{3.5}$$

Nel seguito di questa sezione vengono spiegati brevemente le tipologie di algoritmi di classificazione utilizzate all'interno di queto lavoro di tesi.

3.1 Support Vector Machines

Molti modelli di machine learning sono lineari ovvero possono solo rappresentare linearmente le classi delle istanze di un problema, il che li rende troppo semplici per molte applicazioni. Le *Support Vector Machines* usano invece modelli lineari per rappresentare problemi non lineari [8]. Partiamo però dai modelli semplici.

Prendiamo in esempio la figura 3.3. Questa mostra un insieme di dati, divisi in cerchi e quadrati, disposti all'interno di un grafico. Il data set è

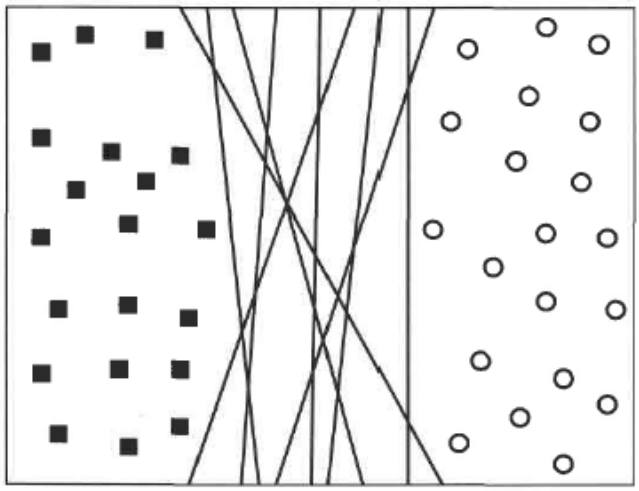

Figura 3.3: Possibili hyperplanes per un problema lineare [8]

linearmente separabile, ovvero è possibile trovare un *hyperplane* che divida esattamente i due insiemi in modo tale che ognuno contenga dati dello stesso tipo, ovvero solo quadrati o solo cerchi. Si può intuire che molte linee possono essere tracciate per soddisfare la precedente richiesta ma non è detto che tutte siano efficaci allo stesso modo. Per cui quale si sceglie? In figura

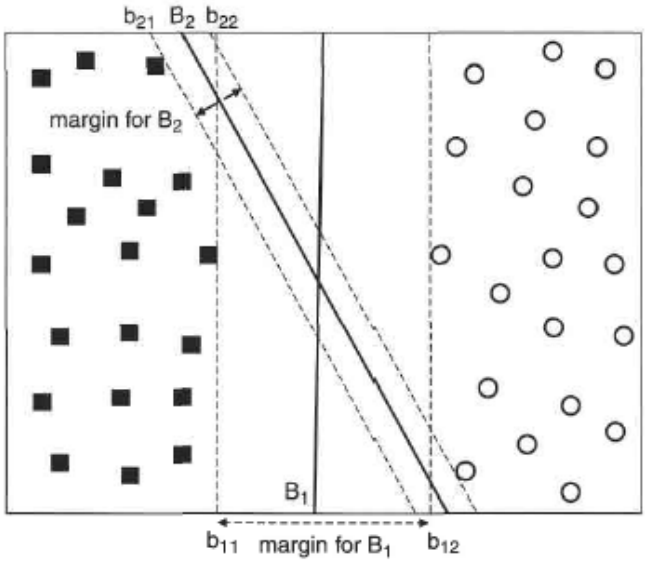

Figura 3.4: Hyperplane con i margini minimi e massimi [8]

3.4 si vede la definizione di due diversi hyperplanes. Nonostante entrambi separino correttamente i due insiemi, la differenza risiede nel fatto che i margini degli hyperplanes, definiti come la distanza tra l'hyperplane stesso e i più vicini punti ad esso, chiamati *support vectors*, per entrambi i data set, sono minimizzati in un caso e massimizzati nell'altro. Modelli con margini piccoli sono più flessibili e possono adattarsi a molti training set mentre modelli con margini grandi hanno errori di generalizzazione più bassi per cui si tende a prediligere gli ultimi, definiti dunque *maximum margin hyperplanes*.

Un modello lineare SVM non è altro che un classificatore che utilizza, in uno spazio lineare, un maximum margin hyperplane per dividere i data set. Un SVM però è anche in grado di gestire problemi non lineari, ovvero quando banalmente non basta una linea, come nei casi precedenti, per poter dividere con precisione i due insiemi. Questo viene fatto tramite l'utilizzo di una funzione di trasformazione che mappa in un nuovo spazio, questa volta lineare, un modello non lineare. Si tratta dunque di elevare a una dimensione superiore un dato modello per poterlo rendere linearmente separabile.

La figura 3.5 mostra come un modello non lineare (sinistra), tramite l'utilizzo di una funzione di trasformazione, viene rimappato in uno spazio nuovo, di dimensione superiore, linearmente separabile.

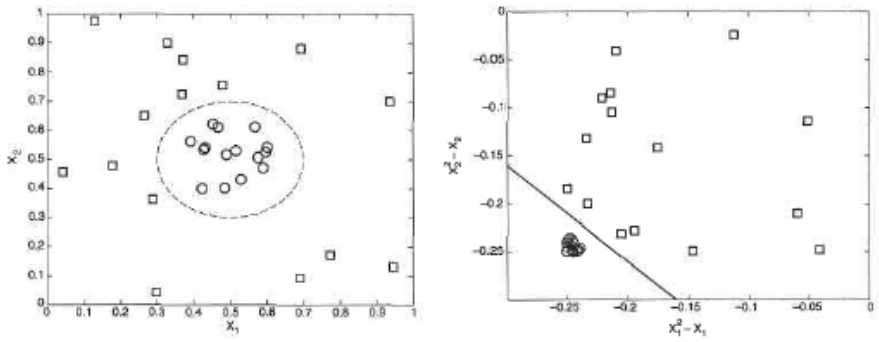

Figura 3.5: Raffigurazione di un modello non lineare e del suo corrispettivo lineare [8]

3.2 K Nearest Neighbors

E' un modello di classificazione in cui l'etichetta di classe di un'istanza di test viene assegnata in base alle classi di istanze di training più simili ad essa [8]. Un modello *Nearest Neighbor* rappresenta ogni istanza del problema come un punto in uno spazio d-dimensionale dove d rappresenta il numero degli attributi. Data un'istanza di test, si determina la sua etichetta di classe confrontandola con le istanze a lei più vicine (k appunto nel caso di *k nearest neighbor*). In che modo viene calcolata la distanza? Se il numero di attributi delle istanze è pari a uno allora la distanza è semplicemente la differenza tra i valori dei due attributi. Se invece le istanze possiedono più di un attributo allora bisogna utilizzare diverse metriche. La più comune è la distanza Euclidea che però richiede la normalizzazione dei valori degli attributi stessi affinché siano confrontabili. La distanza è facilmente calcolabile quando gli attributi hanno valori numerici. In caso contrario è necessario stabilire un grado di importanza tra gli attributi al fine di riuscire a distinguerli tra loro. Solitamente, nel caso più semplice, si assegna alla distanza il valore *uno* se gli attributi sono uguali, *zero* altrimenti. Tuttavia nel caso si voglia mantenere una distinzione più sofisticata e rendere simili attributi diversi si può fare uso di *pesi* nella definizione dei valori degli attributi che riflettono dunque il concetto di similarità. Nel caso il numero di vicini di una particolare istanza sia più di uno, l'etichetta di classe viene scelta in base alla maggioranza delle etichette dei vicini stessi. Bisogna prestare particolare attenzione nello scegliere il valore di k. Valori bassi possono rendere il classificatore suscettibile all'overfitting, ovvero quando un classificatore risulta troppo adatto ai dati di training perdendo generalità, a causa dei rumori presenti nel training data set. Valori alti invece possono portare il classificatore a predire erroneamente un'etichetta dovuto al fatto che sono stati contati nel vicinato dei punti che in realtà sono molto distanti dall'istanza di test.

La figura 3.6 mostra un k nearest neighbor con rispettivamente 1, 2 e 3 vicini.

3.3 Classificatori Bayesiani

In molti casi l'etichetta di classe di un'istanza di test non può essere determinata con certezza anche se i suoi attributi sono identici a quelli di altre istanze

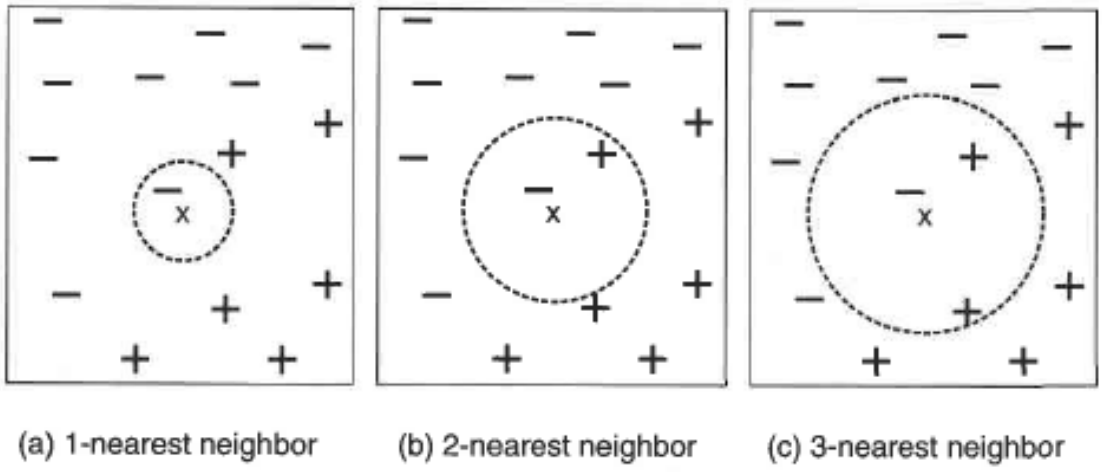

Figura 3.6: Tre diversi k nearest neighbors (con k = 1, 2 e 3) [8]

di training, rendendo quindi la correlazione tra gli attributi e le etichette di classe non deterministica. Un classificatore Bayesiano rappresenta un modello che descrive in maniera probabilistica la relazione tra attributi e etichette tramite l'utilizzo del teorema di Bayes il quale serve per calcolare la probabilità condizionata di un evento Y dato un evento noto X [8] e si esprime come:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{3.6}$$

Nel caso di un problema di classificazione, X rappresenta l'insieme di attributi mentre Y rappresenta la variabile etichetta di classe.

Un classificatore *Naïve Bayes*, data l'etichetta di classe y, stima la probabilità condizionata della classe assumendo che gli attributi siano indipendenti tra loro.

Questi classificatori sono robusti rispetto al rumore, ai valori mancanti e agli attributi irrilevanti ma uno dei maggiori svantaggi risiede nella possibilità che ci sia una dipendenza tra gli attributi, cosa che può drasticamente ridurre l'efficacia del modello.

In questa tesi sono stati utilizzati due tipi di classificatori bayesiani quali il *Multinomial Naïve Bayes* e il *Gaussian Naïve Bayes* i quali utilizzano rispettivamente una distribuzione di probabilità multinomiale e gaussiana.

3.4 Random Forest Classification

Un *albero di decisione* è una tecnica di classificazione composta da tre tipi di nodi.

- **Nodo radice**, che ha zero ingressi e zero o più uscite.
- **Nodi intermedi**, ciascuno dei quali ha esattamente un ingresso e due o più uscite.
- **Nodi foglia**, ciascuno dei quali ha esattamente un ingresso e zero uscite.

Ogni foglia rappresenta un'etichetta di classe mentre gli altri nodi contengono delle condizioni con cui gli attributi vengono confrontati e separati. Classificare un'istanza si riduce dunque ad applicare, partendo dalla radice, le condizioni di test di ciascun nodo all'istanza stessa e a seguire il ramo composto dai risultati dei confronti. Il nodo foglia che verrà raggiunto alla fine determinerà l'etichetta dell'istanza di partenza. In figura 3.7 si può notare una rappresentazione di un albero di decisione.

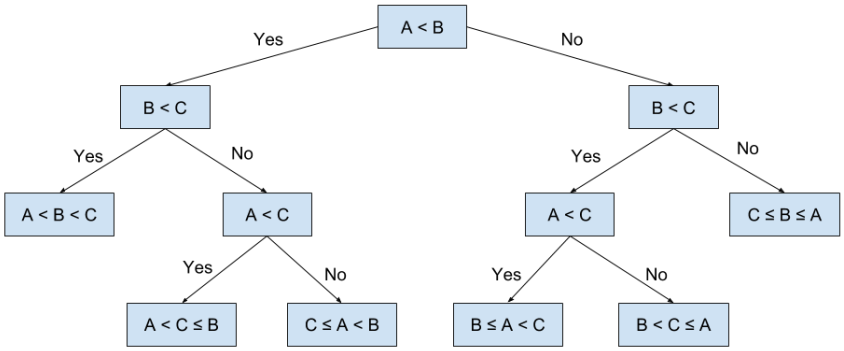

Figura 3.7: Struttura di un albero di decisione [9]

Il *Random Forest* è un tipo di ensemble method che combina le predizioni di diversi alberi di decisione ciascuno dei quali è generato da un vettore costituito da dati del training set presi randomicamente secondo una specifica distribuzione di probabilità [7]. L'etichetta finale viene determinata utilizzando il majority voting.

3.5 Multilayer Perceptron

Lo studio di una rete neurale artificiale è stato ispirato dal sistema nervoso umano e come esso è composta da una rete di connessioni tra nodi che simulano il comportamento dei neuroni [8]. Il modello più semplice è quello del percettrone, o *perceptron*, visualizzato in figura 3.8. Esso consiste in

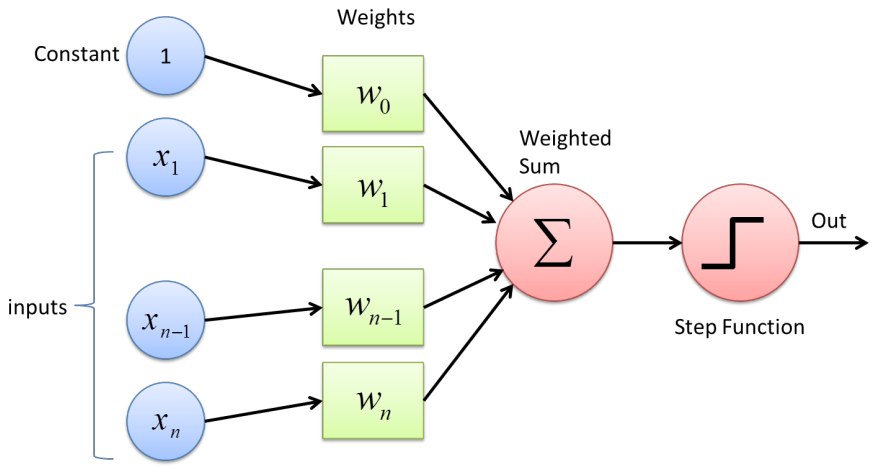

Figura 3.8: Schema di un percettrone [10]

due tipologie di nodi: di ingresso, rappresentanti gli attributi, e di uscita, rappresentante l'output del sistema. Ogni nodo di ingresso è connesso tramite un link pesato al nodo di uscita dove il peso della connessione serve a determinare la forza della stessa. La configurazione dei pesi deve essere gestita in modo da ottimizzare le connessioni al fine di riuscire ad ottenere una buona correlazione tra l'input e l'output del modello. Un perceptron determina il suo valore di output sommando tutti i valori in input considerando il peso di ciascuna connessione, poi togliendo un fattore *bias* (correzione) dalla somma e infine analizzando il segno del risultato attraverso una apposita funzione di attivazione. Brevemente si può descrivere la formula come:

$$y = sign(\sum_{i=1}^{N} w_i x_i - t) \tag{3.7}$$

dove x_i rappresenta l'attributo d'ingresso i-esimo, w_i rappresenta il peso della connessione i-esima e t il fattore bias da sottrarre alla somma. Nella figura 3.8 il fattore bias t da sottrarre è rappresentato da $w_0 x_0$ con $x_0 = 1$ e $w_0 =$ -t.

Durante la fase di addestramento, i pesi vengono ripetutamente aggiustati fino a quando l' output del sistema diventa consistente con gli output dei dati di training, ovvero con i risultati aspettati. Si parla così di tasso di apprendimento o *learning rate*.

Molte volte un modello con un singolo neurone non basta. Si parla così di *Multilayer Artificial Neural Network*, ovvero una rete che contiene, oltre a un input e un output layer, uno o più layer intermedi chiamati layer nascosti. Ci sono due tipi di multilayer NN ovvero **feed forward**, figura 3.9, dove ogni nodo in un layer è collegato solamente ai nodi del layer successivo, e **recurrent**, dove ogni nodo di un layer può essere collegato a nodi del suo stesso layer o anche a nodi dei layer precedenti, oltre che ovviamente ai layer successivi. Le Multilayer neural networks permettono di risolvere problemi più complessi.

Lo scopo dell'algoritmo di apprendimento di una ANN è di determina-

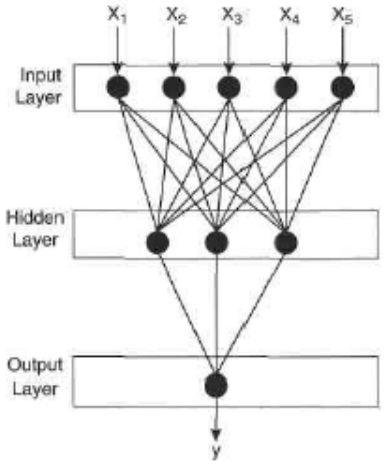

Figura 3.9: Schema di una rete neurale feed forward a più layer [8]

re i valori dei pesi delle connessioni in modo da minimizzare l'errore. Una scelta è quella di basarsi sul metodo *gradient descent* che riesce a risolvere questo problema di ottimizzazione. Per i nodi intermedi invece sorge un problema poiché è difficile comprendere il tasso d'errore senza sapere con esattezza come dovrebbe essere l'output corretto, dato che l'output di un layer intermedio non coincide con l'output desiderato. Si applica così una tecnica chiamata *backpropagation* che confronta il valore di uscita del sistema

con quello desiderato. Sulla base della differenza calcolata, ovvero l'errore, l'algoritmo, tramite una retroazione, modifica i pesi della rete cercando di far convergere progressivamente i risultati in uscita con quelli desiderati.

3.6 Majority Voting

Nel Machine Learning le tecniche basate sul voting sono tecniche ensemble di classificazione [11] che combinano un set di valori in ingresso per produrre uno specifico valore in uscita secondo regole ben definite. Immaginando di avere un set di classificatori, ognuno dei quali produce in output un determinato valore, si fa uso del voting per combinare insieme gli output dei classificatori, secondo regole specifiche, e capire quale di questi dovrà essere utilizzato come output finale del sistema. La tecnica più comune è quella del majority voting che riceve in input un set di valori e restituisce come output il valore presente in maggioranza all'interno del set stesso.

Capitolo 4

Studio della letteratura

I sistemi di trading sono sempre stati i mezzi con cui, attraverso lo studio di grafici e indicatori matematici, poter operare nei mercati finanziari. Con l'avvento della digitalizzazione è stato possibile introdurre l'utilizzo di algoritmi all'interno dell'analisi, del monitoraggio e della predizione dei mercati stessi. Si sono quindi sviluppati i primi sistemi quantitativi dove le conoscenze di figure esperte come analisti tecnici e statisti vengono unite alla programmazione al fine di creare dei modelli automatici di analisi e predizione. Nella loro costruzione, molte tecniche di Machine Learning e Data Mining sono state studiate, implementate e analizzate ai fini di predire prezzi e direzioni di mercato durante gli anni.

Tra le tecniche utilizzate in letteratura si è fatto un grande uso di algoritmi di classificazione, utilizzati per predire la direzione o il prezzo delle azioni nei mercati e talvolta anche per identificare trend in corso. Tra queste si possono trovare modelli basati su classificatori Bayesiani [12], Deep Learning [13], Support Vector Machines [14], K Nearest Neighbor [15, 16], Random Forest [17, 18] e Reti Neurali [19, 20]

Al fine di migliorare le performance dei singoli classificatori, molti studi in letteratura implementano modelli che combinano gli algoritmi visti in precedenza. In [21] gli autori espongono un modello di predizione degli indici di mercato basato su una combinazione di SVR, ANN e RF. In [22] invece gli autori utilizzano il majority voting e il bagging combinato con alberi di decisione, regressori logistici e reti neurali al fine di predire gli utili di mercato mentre in [23] è stata utilizzata una combinazione di SVM e

KNN per predire gli indici del mercato azionario cinese. Infine in [24] gli autori hanno implementato un sistema di supporto alle decisioni di trading basato su una combinazione di algoritmi quali KNN, RFC, MNB, MLP e SVC.

L'utilizzo di queste tecniche non è stato limitato esclusivamente al mercato azionario ma è possibile trovarne delle applicazioni ad altri tipi di mercato come il mercato Forex. In [25] gli autori, utilizzando una combinazione di un classificatore gaussiano multivariante con il Bayesian Voting hanno cercato di identificare un trend in corso mentre in [26] gli autori si sono concentrati nel riuscire a predire l'inversione di un trend tramite l'utilizzo di un algoritmo genetico.

Nell'implementazione dei sistemi di trading quantitativi si è soliti utilizzare nell'analisi le informazioni provenienti dagli indicatori tecnici e dalle serie temporali dei prezzi delle azioni. In letteratura però sono stati presentati dei modelli di predizione che integrano anche il contributo del news sentiment. Sono esempi i lavori [27, 28] in cui si fa uso di una Rete Neurale in un caso, e tecniche di Deep Learning nell'altro, al fine di predire rispettivamente il prezzo e la direzione delle azioni sul mercato.

Modelli quantitativi di trading basati sull'identificazione di un trend reversal sono già stati proposti in letteratura. Per esempio in [29] gli autori basandosi sul volume di ricerche fatte su Google hanno mostrato come investire a breve termine su azioni il cui volume di ricerca si ingigantisce in poco tempo, migliori la redditività della strategia short-term reversal.

Altre strategie sono basate su (i) modello delle candele giapponesi [30, 31, 32], un tipo di grafico utilizzato per l'analisi del prezzo di un'azione, (ii) il prezzo di un'azione combinato al proprio momentum [33, 34] e (iii) pattern recognition [35].

Sono basati invece sull'uso di reti neurali i lavori [36, 37]. In [36] gli autori propongono un metodo per addestrare una rete neurale tramite la strategia particle swarm optimization per riuscire a massimizzare i guadagni e minimizzare i rischi nelle transizioni mentre in [37] gli autori propongono un metodo di individuazione di un trend reversal basato sull'analisi dei grafici di mercato.

Recentemente gli autori di [38] hanno costruito un sistema di predizione della direzione di un'azione basandosi sul trend reversal e sul deep-learning.

Hanno infatti implementato un LSTM notando un'accuratezza maggiore rispetto ai singoli classificatori come SVM e MLP.

Anche per questa strategia, alcuni studi in letteratura propongono dei modelli che combinano diversi classificatori al fine di migliorare le performance. In [39] gli autori hanno creato un sistema combinando algoritmi di classificazione quali SVM e KNN mentre in [40] gli autori propongo un metodo a due step basato sulla combinazione di un Bayesian Factor Graph con ANN, SVM e HMM.

Infine in [41] gli autori hanno costruito un sistema di trading quantitativo per l'identificazione di un trend reversal basandosi su (i) serie storiche dei prezzi, (ii) indicatori tecnici, (iii) news sentiment e utilizzando un set di classificatori comprendente MLP, MNB, RFC, SVC e KNN.

Il sistema presentato in questo lavoro si propone di valutare diverse strategie di trading basate sul riconoscimento automatico di un inversione di trend attraverso l'uso di tecniche di Machine Learning quali SVC, KNN, RFC, GNB, MNB e MLP, già utilizzate precedentemente in letteratura [12], [14], [15], [16], [17], [18], [20], sfruttando informazioni riguardanti il prezzo delle azioni, indicatori tecnici, come in [39], [40], [24] e news sentiment, come in [27], [28].

Questo sistema può essere visto come un'estensione del lavoro svolto da Cagliero L., Baralis E., Attanasio G. *et al.* in [41] con la differenza che qui vengono valutate (i) diverse strategie di riconoscimento di un trend in corso e (ii) una nuova strategia di identificazione di un inversione di trend basata sul metodo ensemble del majority voting. Inoltre parte del lavoro svolto è servito come sperimentazione volta a comprendere l'impatto dei principali parametri del sistema e la corretta configurazione degli algoritmi in esso integrati.

L'obiettivo di questo lavoro è di dimostrare l'efficacia dell'approccio basato sul Machine Learning rispetto ad un approccio tradizionale basato su analisi tecnica e definire le configurazioni più appropriate del sistema analizzando l'impatto di vari fattori sui risultati della simulazione trading.

Tabella 4.1: Tabella riassuntiva della Letteratura

Articolo	**Mercato**	**Tipologia Dati**	**Tecniche Usate**	**Obiettivo di Predizione**
[30]	Stock	Prezzo, Indicatori	Candlestick Charts	Reversal
[31]	Stock	Prezzo, Indicatori	Candlestick Charts	Market Timing
[32]	Stock	Prezzo	Candlestick Charts, Fuzzy Logic	Reversal
[33]	Stock	Prezzo, Momentum	Multivariate Regression	Reversal
[34]	Stock	Prezzo, Momentum	Mean Reversion	Reversal
[29]	Stock	Google Searches	Search Volume Index	Reversal
[35]	Stock	Prezzo, Indicatori	Pattern Recognition	Reversal
[36]	Stock	Prezzo, Indicatori	NN, PSO	Reversal
[37]	Stock	Prezzo, Indicatori	ANN, Grafici	Reversal
[41]	Stock	Prezzo, News Indicatori	SVC, KNN, MLP, MNB, RFC	Reversal
[39]	Stock	Prezzo, Indicatori	Ensemble SVM - KNN	Reversal
[40]	Stock	Prezzo, Indicatori	Ensemble DBFG - ANN, SVM, HMM	Reversal
[24]	Stock	Prezzo, News, Indicatori	Ensemble KNN, MLP, RFC, SVC, MNB	Supporto Stock Trading
[26]	Forex	Prezzo, DC	Genetic Algorithm	Reversal
[38]	Stock	Prezzo, Indicatori	LSTM	Reversal
[25]	Forex	Prezzo, Indicatori	Ensemble MGC, Bayesian Voting	Identificazione Trend

Articolo	**Mercato**	**Tipologia Dati**	**Tecniche Usate**	**Obiettivo di Predizione**
[12]	Stock	Prezzo	Bayesian Classifier	Direzione Stock
[13]	Stock	Prezzo	Deep Learning	Prezzo Stock
[14]	Stock	Prezzo, Indicatori	SVM	Direzione Stock
[15]	Stock	Prezzo, Indicatori	KNN	Identificazione Trend
[16]	Stock	Prezzo	KNN	Prezzo Stock
[23]	Stock	Prezzo, Indicatori	SVM, KNN	Indici di Mercato
[17]	Stock	Prezzo, Indicatori	RFC	Identificazione Trend
[18]	Stock	Prezzo, Indicatori	RFC	Direzione Stock
[19]	Stock	Prezzo, Indicatori	SVM, NB, DT, MLP, LSTM	Identificazione Trend
[20]	Stock	Prezzo, Indicatori	MLP	Indici di Mercato
[27]	Stock	Prezzo, News	NN	Prezzo Stock
[28]	Stock	Prezzo, News Indicatori	LSTM	Direzione Stock
[21]	Stock	Prezzo, Indicatori	Ensemble SVR - ANN, RF, SVR	Indici di Mercato
[22]	Stock	Prezzo	Ensemble DT NN, LR	Utile di Investimento

Capitolo 5

Metodologia presentata

In questa Sezione viene descritto un sistema di trading quantitativo che investe sul mercato azionario mediante una strategia multiday di tipo trend reversal. Il modello è stato testato con dati veri rappresentanti i valori delle azioni dell'indice americano Standard & Poor 500. Il sistema proposto si basa sull'utilizzo di tecniche di machine learning quali algoritmi di classificazione per predire la direzione di un'azione nei giorni successivi. I classificatori utilizzati sono quelli descritti nella Sezione 3. L'obiettivo del sistema è quello di valutare l'efficacia delle strategie presentate nell'identificazione di un reversal e cercare quindi di capire quale sia la più performante ai fini della strategia di trading applicata.

Come mostrato dalla figura 5.1 la struttura di questa tesi si compone di diversi blocchi, quali:

1. **Raccolta dati e calcolo degli indicatori**: si raccolgono dati relativi alle azioni dell'indice americano *Standard & Poor 500*, su base giornaliera, secondo diversi tipi di descrittori basati sul prezzo, analisi tecnica e news sentiment.

2. **Preparazione dei dati**: prima di essere passati al classificatore i dati vengono (i) normalizzati, ovvero trasformati in modo tale che assumano valori compresi tra 0 e 1 al fine di renderli omogenei e confrontabili, (ii) filtrati, in modo da tenere solo i descrittori relativi alla strategia scelta, e infine (iii) scalati, al fine di avere, per ciascun descrittore, il valore ad esso relativo per il giorno corrente fino a N giorni precedenti (con N parametro del sistema) .

3. **Riconoscimento automatico di un trend in corso**: si ricerca un trigger che identifichi una possibile inversione di trend attraverso l'uso di diverse strategie quali:

   (i) l'utilizzo di una sliding window, ovvero una finestra di W giorni nei quali la direzione del prezzo dell'azione deve essere concorde. Se la direzione trovata è 'B' (buy), ovvero W giorni in cui il prezzo dell'azione sale, si scommette su un downtrend, se invece la direzione trovata è 'S' (sell) si scommette su un uptrend.

   (ii) l'utilizzo dell'incrocio di medie mobili quali SMA e MACD. Se il risultato è negativo si scommette su un downtrend, se è positivo si scommette su un uptrend.

   (iii) la combinazione di ciascuna delle precedenti strategie con un filtro sul volume scambiato per azione. Questo filtro non è altro che la differenza tra il volume scambiato nel giorno corrente con la media del volume scambiato nei W giorni precedenti. Per poter procedere il risultato deve essere concorde con il reversal ipotizzato dalla strategia applicata, per cui se è stato ipotizzato un downtrend il risultato del filtro deve essere negativo, se invece è stato ipotizzato un uptrend il risultato deve essere positivo.

   Nel caso un trigger venga individuato, si memorizza la direzione del reversal ipotizzato in una variabile target, contenente quindi 'S' nel caso si ipotizzi un downtrend, 'B' nel caso contrario.

4. **Addestramento di un classificatore**: il classificatore viene addestrato al fine di predire la direzione (buy 'B', hold 'H', sell 'S') del prezzo di un'azione fino a N giorni in avanti rispetto al corrente. Per esempio, immaginando $N = 5$, il classificatore può predire una finestra di etichette del tipo [B, B, S, H, B].

5. **Previsione di un'inversione di trend**: in questo step si verifica se ci sia una corrispondenza tra la direzione del target ipotizzato e le etichette predette del classificatore. Questo avviene tramite l'uso del majority voting, una tecnica che restituisce, come risultato, la direzione presente in maggioranza tra le etichette predette (B nel caso dell'esempio precedente). Se il risultato del majority voting è concorde con il target ipotizzato precedentemente (le due etichette devono essere uguali) si procede con il modulo di trading.

6. **Gestione del trade**: questo modulo racchiude tutta la logica di gestione del trade come (i) chiudere le posizioni già aperte se un target non è più presente o se le perdite hanno superato una soglia di sicurezza definita come *stop loss*, (ii) aprire nuove posizioni del tipo indicato dal target quindi long-selling in caso di uptrend, short-selling in caso di downtrend e (iii) aggiornare il budget di mercato.

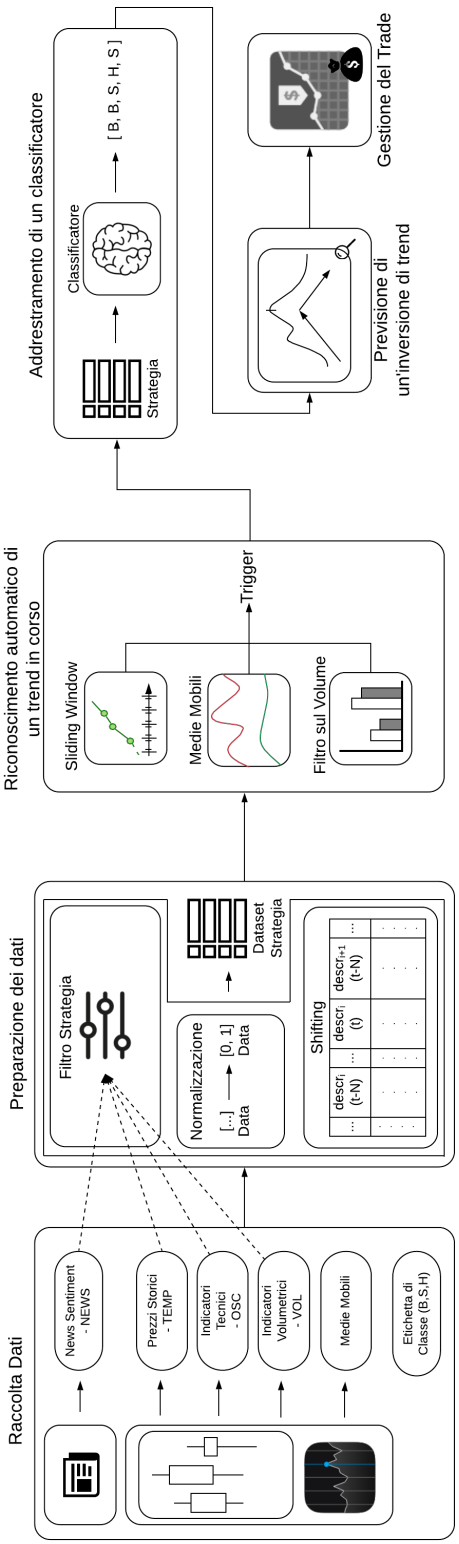

Figura 5.1: Architettura del sistema di trading proposto

5.1 Raccolta dati e calcolo degli indicatori

Questo modulo prevede la raccolta dei dati relativi alle azioni dell'indice americano S&P500 su base giornaliera, l'estrazione da esso di un insieme di caratteristiche (features) specifiche e infine il loro inserimento in un dataset relazionale dove ad ogni attributo corrisponde un determinato descrittore.

Viene innanzitutto creato un dataset per ogni coppia stock-anno come mostrato nella tabella 5.1 dove s rappresenta un'azione e y rappresenta l'anno.

Elenco Dataset
s_1y_1
...
s_1y_N
...
...
...
s_My_1
...
s_My_N

Tabella 5.1: Elenco dei dataset

All'interno di ciascun dataset s_iy_j, ogni riga, rappresentante un giorno di mercato, descrive i valori assunti da ciascun descrittore come mostrato nella tabella 5.2 dove *desc* rappresenta un descrittore e *day* rappresenta un giorno di mercato dell'anno y_j.

	$desc_1$	$desc_2$	...	$desc_G$
day_1				
day_2				
...				
...				
day_L				

Tabella 5.2: Struttura di un dataset s_iy_j

I dati inoltre vengono puliti precedentemente per evitare un appesantimento durante la fase di training. Vengono infatti rimosse dai dataset quelle

azioni per cui, per almeno un anno, presentano un valore mancante.

Per ciascuna azione sono disponibili i valori storici per ogni giorno di mercato, esclusi quindi il sabato e la domenica, dal 2007 al 2017, di una serie di descrittori [6] che possono essere suddivisi nelle seguenti categorie, o features:

- **Medie Mobili**: le medie mobili sono uno dei più utili indicatori tecnici per generare segnali di mercato e per definire la direzione di un trend dello strumento analizzato [42]. Una media mobile rappresenta una media calcolata su una determinata quantità di dati all'interno di una finestra temporale. E' definita mobile perchè prende in considerazione gli ultimi dati della finestra disponibili in ordine temporale. Quando nuovi dati verranno prodotti, la finestra si sposterà eliminando dal calcolo i dati più vecchi.

  Le medie mobili presenti in questi dataset sono di tre diverse tipologie.

  - *Media Mobile Semplice*: detta anche Simple Moving Average, o SMA, ed è la media mobile più utilizzata in virtù della sua semplicità. Per il suo calcolo vengono presi i dati di uno specifico periodo e ne viene fatta semplicemente la media. E' calcolata come

    $$\mathrm{SMA_N(t)} = \frac{1}{N}\sum_{i=1}^{N} x_{t-i} \tag{5.1}$$

    dove N è il numero dei periodi, t è il giorno corrente in cui la media viene calcolata e x è la caratteristica su cui la media viene calcolata, per esempio il prezzo di chiusura. Talvolta si preferisce includere il giorno stesso in cui la media viene calcolata all'interno del periodo N cosicché la formula precedente diventa:

    $$\mathrm{SMA_N(t)} = \frac{1}{N}\sum_{i=0}^{N-1} x_{t-i} \tag{5.2}$$

    La scelta su quale delle due versioni utilizzare è a discrezione dell'utilizzatore. Questo discorso può essere esteso per la maggior parte dei descrittori spiegati in questa Sezione.
  - *Media Mobile Esponenziale*: detta anche Exponential Moving Average, o EMA, rappresenta un indicatore tecnico molto complesso.

Questa media è costruita utilizzando una ponderazione esponenziale decrescente ovvero l'utilizzo dei dati più lontani nel tempo sarà sempre presente nel calcolo ma con un peso minore. In sostanza si tende dare una maggiore importanza ai valori più recenti assegnando ad essi un peso maggiore nel calcolo. Viene calcolata come:

$$\text{EMA}_\text{N}(\text{t}) = \left(x_t \cdot \left(\frac{\text{Smooth}}{1+N}\right)\right) + \text{EMA}_\text{N}(\text{t-1}) \cdot \left(1 - \left(\frac{\text{Smooth}}{1+N}\right)\right) \tag{5.3}$$

dove, esattamente come per la SMA, N è il numero dei periodi, t è il giorno corrente, x è la caratteristica mentre *Smooth* è un parametro che indica il peso con cui i giorni più recenti vengono tenuti in considerazione all'interno del calcolo. Maggiore è il fattore Smooth, maggiore è l'influenza che i giorni più recenti hanno nel calcolo della EMA. Solitamente si assegna allo Smooth il valore 2.

– *Convergenza e Divergenza di Medie Mobili*: detta anche Moving Average Convergence/Divergence, o MACD, è un oscillatore complesso. Viene calcolata prima di tutto la differenza tra due medie mobili esponenziali di periodi diversi (12 e 26) chiamata MACD line. Successivamente rappresenta la differenza tra la MACD line e una media mobile esponenziale delle stesse precedenti come un istogramma che può essere quindi analizzato dai trader.

$$\text{MACD line} = \text{EMA}_{12} \text{ - } \text{EMA}_{26} \tag{5.4}$$

$$\text{Signal line} = \text{EMA}_{\text{MACD,9}} \tag{5.5}$$

$$\text{Istogramma} = \text{MACD - Signal line} \tag{5.6}$$

Per questo lavoro, nel calcolare questi indicatori, sono stati considerati i prezzi di chiusura di giornata delle azioni mentre per l'orizzonte temporale sono state utilizzate differenze di medie mobili di periodi differenti.

- **Oscillatori Tecnici e Indicatori di Volatilità**: Gli oscillatori tecnici sono degli indicatori, definiti sul prezzo, sul volume o su una combinazione di entrambi, il cui valore oscilla all'interno di un range, di solito 0 - 100, e servono ad identificare condizioni di ipervenduto o ipercomprato

per anticipare un possibile cambio del trend. Gli indicatori di volatilità invece servono a misurare quanto velocemente i prezzi cambino nel mercato. In seguito questa categoria verrà denotata con OSC.

Nei dataset sono presenti diversi tipi di indicatori tecnici e oscillatori:

– *Aroon Oscillator*: usato per determinare se un obiettivo è in trend e in caso positivo quanto forte questo sia. E' basato su due indicatori, rappresentanti due linee con valori compresi tra 0 e 100, chiamati Aroon Up e Aroon Down che misurano la spinta all'acquisto, il primo, e la spinta alla vendita, il secondo, in un determinato periodo. Quando la differenza tra i due è positiva significa che un uptrend è presente, viceversa quando la differenza è negativa significa che un downtrend è presente. Fissato N come il numero di periodi di osservazione (solitamente 14), i due indicatori si calcolano come:

$$\text{Aroon Up} = 100 \cdot \frac{N - \#\ \textit{periodi dall'ultimo massimo}}{N} \tag{5.7}$$

$$\text{Aroon Down} = 100 \cdot \frac{N - \#\ \textit{periodi dall'ultimo minimo}}{N} \tag{5.8}$$

Infine l'oscillatore si calcola come:

$$\text{AO}_N = \text{Aroon Up - Aroon Down} \tag{5.9}$$

– *Average True Range Percent*: l'average true range è un indicatore che serve per misurare la volatilità di uno strumento finanziario. Si definisce il *True Range* per un determinato giorno t come il massimo tra (i) la differenza tra il massimo e minimo del giorno t, (ii) la differenza in valore assoluto tra il massimo di t e il prezzo di chiusura di $t-1$, (iii) la differenza in valore assoluto tra il minimo di t e il prezzo di chiusura di $t-1$.

$$TR_t = \text{Max}[(H_t - L_t), \text{Abs}(H - pc_{t-1}), \text{Abs}(L - pc_{t-1})] \tag{5.10}$$

L'average true range semplicemente è una media dei TR calcolata su un periodo N (solitamente 14):

$$ATR_N = \frac{1}{N} \sum_{t=1}^{N} TR_t \tag{5.11}$$

mentre l'average true range percent non è altro che l'espressione dell'ATR in percentuale rispetto al prezzo di chiusura:

$$ATRP_N = \frac{ATR_N}{pc} \cdot 100 \tag{5.12}$$

– Differenza tra *DI+* e *DI-*: questi due indicatori, chiamati Positive Directional Indicator (DI+) e Negative Directional Indicator (DI-) misurano rispettivamente la presenza di un uptrend e di un downtrend. La loro differenza quindi rappresenta un valore che, se negativo identifica una tendenza nei prezzi a calare, mentre se positivo identifica una tendenza nei prezzi a salire.
L'indicatore DI+ si calcola come:

$$\mathrm{DM}_+(t) = max(t) - max(t-1) \tag{5.13}$$

$$\text{Smooth } \mathrm{DM}_+(t) = \left(\sum_{i=1}^{N} \mathrm{DM}_+(t-i)\right) - \left(\frac{\sum_{i=1}^{N} \mathrm{DM}_+(t-i)}{N}\right) + \mathrm{DM}_+(t) \tag{5.14}$$

$$\mathrm{DI}_+(t) = \left(\frac{\text{Smooth } \mathrm{DM}_+(t)}{\mathrm{ATR}_N}\right) \cdot 100 \tag{5.15}$$

dove $max(t)$ indica il prezzo massimo raggiunto nel giorno t, N è il numero di periodi su cui l'indicatore viene calcolato (solitamente 14) e ATR_N è l'average true range calcolato su N periodi. Similmente l'indicatore DI- si calcola come:

$$\mathrm{DM}_-(t) = min(t-1) - min(t) \tag{5.16}$$

$$\text{Smooth } \mathrm{DM}_-(t) = \left(\sum_{i=1}^{N} \mathrm{DM}_-(t-i)\right) - \left(\frac{\sum_{i=1}^{N} \mathrm{DM}_-(t-i)}{N}\right) + \mathrm{DM}_-(t) \tag{5.17}$$

$$\mathrm{DI}_-(t) = \left(\frac{\text{Smooth } \mathrm{DM}_-(t)}{\mathrm{ATR}_N}\right) \cdot 100 \tag{5.18}$$

dove $min(t)$ indica il prezzo minimo raggiunto nel giorno t.

– *Average Directional Index*: è oscillatore che misura l'intensità di un trend. Tanto più il suo valore, compreso tra 0 e 100, è alto tanto più forte è il trend. Si calcola come:

$$\mathrm{DX}(t) = \left| \frac{\mathrm{DI}_+(t) - \mathrm{DI}_-(t)}{\mathrm{DI}_+(t) + \mathrm{DI}_-(t)} \right| \cdot 100 \tag{5.19}$$

$$\mathrm{ADX}_N(t) = \frac{(\mathrm{ADX}_N(t-1) \cdot N - 1) + \mathrm{DX}(t)}{N} \tag{5.20}$$

dove N è il numero di periodi su cui l'indicatore viene calcolato (solitamente 14) e DI_+ e DI_- sono gli indicatori di positive directional index e negative directional index calcolati in precedenza.

– *Percentage Price Oscillator*: è un oscillatore di momentum basato sulla differenza tra due medie mobili esponenziali, una più lenta e una più veloce. Il risultato viene poi diviso per la più lenta.

$$PPO = \frac{\mathrm{EMA}_{12} - \mathrm{EMA}_{26}}{\mathrm{EMA}_{26}} \cdot 100 \tag{5.21}$$

Come si può evincere è praticamente uguale alla MACD line tranne per il fatto che in questo caso viene misurata la differenza percentuale tra le due medie mentre con il MACD viene misurata la differenza assoluta. Esattamente come per il MACD, il PPO dà informazioni sul momentum, sulla direzione di un trend e sulla sua durata.

– *Relative Strength Index*: indicatore che segnala fasi di ipercomprato o ipervenduto individuando così possibili inversioni di trend. Confronta la grandezza dei guadagni recenti con la grandezza delle perdite recenti restituendo un valore da 0 a 100.

$$RSI_N = 100 - \frac{100}{1 + \frac{\mathrm{profits}_N}{\mathrm{losses}_N}} \tag{5.22}$$

dove $\mathrm{profits}_N$ e losses_N indicano rispettivamente il numero di guadagni e di perdite negli ultimi N giorni (solitamente 14).

– *Money Flow Index*: utilizza sia il prezzo che il volume per misurare la pressione di acquisto e di vendita. E' solitamente utilizzato per identificare inversioni di prezzo. Può essere considerato una sorta di

RSI ma che prende anche in considerazione il volume. E' calcolato come:

$$\text{(TP) Typical Price}(t) = \frac{max(t) + min(t) + close(t)}{3} \tag{5.23}$$

$$\text{(MF) Money Flow}(t) = \text{TP}(t) \cdot \text{volume}(t) \tag{5.24}$$

$$\text{Positive Flow}(t) = \begin{cases} \text{MF}(t) & \text{if TP}(t) > \text{TP}(t-1) \\ 0 & \text{altrimenti} \end{cases} \tag{5.25}$$

$$\text{Negative Flow}(t) = \begin{cases} \text{MF}(t) & \text{if TP}(t) < \text{TP}(t-1) \\ 0 & \text{altrimenti} \end{cases} \tag{5.26}$$

$$\text{Money Flow Ratio} = \frac{\sum_{i=1}^{N} \text{Positive Flow}(t-i)}{\sum_{i=1}^{N} \text{Negative Flow}(t-i)} \tag{5.27}$$

$$\text{MFI}_N = 100 - \frac{100}{1 + \text{Money Flow Ratio}} \tag{5.28}$$

dove $max(t)$, $min(t)$ e $close(t)$ rappresentano il prezzo massimo, minimo e di chiusura raggiunti nel giorno t mentre N rappresenta la lunghezza del periodo (solitamente 14).

– *True Strength Index*: indicatore utilizzato per individuare le situazioni di ipervenduto o ipercomprato indicando potenziali inversioni di trend. Misura il momentum a breve termine sfruttando i valori delle medie mobili esponenziali. Si calcola come:

$$\text{TSI}(cp_t, N, M) = \frac{\text{EMA}(\text{EMA}(d, N), M)}{\text{EMA}(\text{EMA}(|d|, N), M)} \cdot 100 \tag{5.29}$$

dove cp è il prezzo di chiusura del giorno t, N è la lunghezza del primo periodo (solitamente 13), M è la lunghezza del secondo periodo (solitamente 25) e d è la differenza tra il prezzo di chiusura del giorno t con quello del giorno precedente.

– *Stochastic Oscillator*: indicatore utilizzato per determinare la fine di un trend e la sua possibile inversione. Confronta il prezzo di chiusura di un asset con il range massimo - minimo su uno specifico

numero di periodi. E' composto da due linee, una più veloce (%K) e una più lenta (%D) calcolate come:

$$\%K = \frac{cp - min_N}{max_N - min_N} \cdot 100 \tag{5.30}$$

$$\%D = \mathrm{SMA}_3(\%K) \tag{5.31}$$

dove *cp* indica il prezzo di chiusura corrente, min_N e max_N rappresentano il minimo più basso e il massimo più alto raggiunti nel periodo lungo N (solitamente 14).

– *Chande Momentum Oscillator*: serve per catturare il momento di un asset. L'indicatore oscilla tra -100 e 100 con un livello di ipercomprato oltre 50 e ipervenduto inferiore a -50. Quando il valore è positivo è possibile interpretarlo come un segnale di acquisto, viceversa quando il valore è negativo lo si può interpretare come segnale di vendita. Si calcola come:

$$Up(t) = \begin{cases} cp(t) - cp(t-1) & \text{if } cp(t) > cp(t-1) \\ 0 & \text{altrimenti} \end{cases} \tag{5.32}$$

$$Down(t) = \begin{cases} cp(t-1) - cp(t) & \text{if } cp(t) < cp(t-1) \\ 0 & \text{altrimenti} \end{cases} \tag{5.33}$$

$$Su = \sum_{i=1}^{N} Up(t-i) \tag{5.34}$$

$$Sd = \sum_{i=1}^{N} Down(t-i) \tag{5.35}$$

$$\mathrm{CPO}_N = \frac{Su - Sd}{Su + Sd} \cdot 100 \tag{5.36}$$

dove $cp(t)$ è il prezzo di chiusura del giorno t e N è la lunghezza del periodo (solitamente 14).

– *Percentage Volume Oscillator*: un indicatore di momentum molto simile al Percentage Price Oscillator con la differenza che le medie mobili esponenziali sono calcolate rispetto al volume scambiato e non rispetto al prezzo di chiusura.

- **Indici di Volume**: Indicati in seguito con VOL, si compongono di:

- *Force Index*: combina il prezzo, l'estensione e il volume per misurare la forza di un trend identificando eventuali inversioni o correzioni. E' calcolato come:

$$\mathrm{FI}_N = volume(t) \cdot [\mathrm{MA}_N(t) - \mathrm{MA}_N(t-1)] \tag{5.37}$$

  dove t indica il giorno per cui la MA è calcolata mentre N è il periodo.

- *On Balance Volume*: misura la pressione di acquisto e di vendita aggiungendo volume nei giorni rialzisti e sottraendone nei giorni ribassisti. Il calcolo è molto semplice e dipende dal prezzo di chiusura dell'asset nel giorno t:

$$OBV(t) = \begin{cases} OBV(t-1) + \text{volume}(t) & \text{if } cp(t) > cp(t-1) \\ OBV(t-1) - \text{volume}(t) & \text{if } cp(t) < cp(t-1) \\ OBV(t-1) & \text{if } cp(t) \approx cp(t-1) \end{cases} \tag{5.38}$$

- *Accumulation / Distribution Line*: è un indicatore di tipo momentum creato essenzialmente per misurare la domanda e l'offerta, cercando di capire se gli investitori stanno accumulando (comprando) o distribuendo (vendendo) asset.

- **News Sentiment**: questa categoria di features, in seguito chiamata NEWS, rappresenta le opinioni della community finanziaria. Ciò che viene rappresentato sono dei descrittori estratti dagli articoli finanziari che riguardano le azioni dei dataset. Il calcolo di questi descrittori si fonda sul conteggio delle occorrenze dei termini nelle news che sono presenti in un particolare dizionario basato sul sentiment. Per sentiment si intende il tipo di umore e il livello di aspettative degli investitori sulla possibile futura evoluzione di un mercato. I descrittori sono:

  - Numero totale di articoli con notizie pertinenti nel giorno considerato.
  - Numero totale di articoli con notizie pertinenti rilasciati nel giorno specificato e contenente parole con sentiment, positivo o negativo.

  - Numero totale di articoli con notizie pertinenti rilasciati nel giorno specificato che contengono almeno una parola con sentiment positivo e non contengono parole negative.
  - Numero totale di articoli con notizie pertinenti rilasciati nel giorno specificato che contengono almeno una parola con sentiment negativo e non contengono parole positive.
  - Numero totale di articoli con notizie pertinenti rilasciati nel giorno specificato che contengono qualsiasi parola con sentiment positivo e nessuna parola di sentiment negativo.
  - Numero totale di articoli con notizie pertinenti rilasciati nel giorno specificato che contengono qualsiasi parola con sentiment negativo e nessuna parola di sentiment positivo.
  - Numero totale di articoli con notizie pertinenti rilasciati nel giorno specificato espresso con il numero di parole contenute in essi.
  - Numero totale di parole con sentiment positivo contenute in qualsiasi articolo con notizie pertinenti.
  - Numero totale di parole con sentiment negativo contenute in qualsiasi articolo con notizie pertinenti.
  - Numero totale di parole con sentiment positivo contenute in articoli con notizie pertinenti rilasciati nel giorno specificato con una forte correlazione con la direzione dell'azione del giorno successivo nei giorni passati.
  - Numero totale di parole con sentiment negativo contenute in articoli con notizie pertinenti rilasciati nel giorno specificato con una forte correlazione con la direzione dell'azione del giorno successivo nei giorni passati.

- **Variazione percentuale del prezzo di chiusura**: questa categoria, denominata in seguito TEMP, si compone della differenza in percentuale tra il prezzo di chiusura dell'azione specificata per un dato giorno con il prezzo del giorno precedente, fino a 5 giorni precedenti alla data considerata. Indicando con *cp* il prezzo di chiusura dell'azione, questa categoria di descrittori è calcolata come:

$$\text{TEMP}(t) = \frac{cp_t - cp_{t-1}}{cp_{t-1}} \cdot 100, \text{ con } t \in [0, 4] \tag{5.39}$$

  dove 0 indica il giorno corrente.

- **Storico dei prezzi**: questa categoria include le serie storiche dei prezzi, raccolte su base giornaliera, come il prezzo di chiusura, il prezzo di apertura, il volume scambiato, il prezzo minimo e massimo raggiunto.
- **Etichetta di Classe**: definisce il segnale di mercato generato dal confronto tra il prezzo di chiusura di un determinato giorno con quello del giorno successivo. Indicando il prezzo di chiusura come *cp*, la formula che identifica il segnale di mercato per una data azione rispetto al giorno successivo si rappresenta come:

$$class(s,t) = \begin{cases} buy & \text{if } \frac{cp_{t+1} - cp_t}{cp_t} \cdot 100 \geq 1 \\ sell & \text{if } \frac{cp_{t+1} - cp_t}{cp_t} \cdot 100 \leq -1 \\ hold & \text{altrimenti} \end{cases} \tag{5.40}$$

  Siccome per poter determinare l'etichetta di una data azione in un determinato giorno è necessario conoscere le informazioni della stessa azione nel giorno seguente, è possibile calcolare le etichette solo per giorni antecedenti a quello corrente. Lo scopo dei classificatori è infatti quello di predire il segnale di mercato del giorno corrente rispetto al giorno successivo.

Feature	**Descrittore**
MA	$SMA_{5\text{-}20}$
	$SMA_{8\text{-}15}$
	$SMA_{20\text{-}50}$
	$EMA_{5\text{-}20}$
	$EMA_{8\text{-}15}$
	$EMA_{20\text{-}50}$
	$MACD_{12\text{-}26}$
OSC	AO_{14}
	ADX_{14}
	WD_{14}
	$PPO_{12\text{-}26}$
	RSI_{14}
	MFI_{14}
	TSI
	SO_{14}
	CMO_{14}
	$ATRP_{14}$
	PVO_{14}
VOL	FI_{13}
	FI_{50}
	ADL
	OBV
TEMP	...
	Guardare Descrizione
	...
NEWS	...
	Guardare Descrizione
	...
PRICE	Prezzo di Chiusura
	Prezzo di Apertura
	Volume Scambiato
	Massimo Prezzo Raggiunto
	Minimo Prezzo Raggiunto
CLASS	Etichetta

Tabella 5.3: Elenco dei descrittori utilizzati

5.2 Preparazione dei dati

I dati vengono prima di tutto normalizzati (algoritmo 1, riga 5) al fine di renderli omogenei e confrontabili. La normalizzazione usata è quella *min - max* che ridimensiona i valori in un range da 0 a 1.

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{5.41}$$

Successivamente, a seconda della strategia scelta, il dataset viene filtrato al fine di mantenere solo i descrittori scelti (algoritmo 1, riga 6). Le strategie utilizzate in questa tesi sono:

- TEMP
- VOL
- OSC+TEMP
- OSC+VOL
- NEWS+TEMP
- NEWS+VOL
- OSC+TEMP+NEWS
- OSC+VOL+NEWS
- ALL

Infine ogni record del dataset viene scalato in modo che ogni descrittore contenga non solo il valore relativo al giorno corrente ma contenga anche i valori degli ultimi N giorni, dove N è un parametro del sistema (algoritmo 1, riga 7). In sostanza per un singolo descrittore *desc*, per l'azione s e per il giorno t il record contiene i valori *desc(s, t)*, *desc(s, t-1)*, ..., *desc(s, t-N)*.

Algoritmo 1 Preparazione dei dati. Riceve in ingresso il dataframe descrivente una specifica azione in un determinato anno con tutti i descrittori citati nella sezione precedente.

1: **procedure** PREPROCESS($df, N, strategy$)
2: ▷ df è il dataframe
3: ▷ N è il numero di giorni precedenti al corrente di cui si vuole fare lo shift. Come verrà spiegato anche in seguito è anche il numero dei giorni successivi al corrente per cui il classificatore predirà le etichette
4: ▷ strategy è una delle strategie citate a inizio sezione
5: $df\_norm \leftarrow normalization(df)$
6: $df\_filter \leftarrow filter(df\_norm, strategy)$
7: $df\_scaled \leftarrow shift(df\_filter, N)$
8: **return** df_scaled
9: **end procedure**

5.3 Riconoscimento automatico di un trend in corso

In questa sezione vengono descritte diverse strategie per l'individuazione di un trend reversal trigger, ovvero di un giorno di mercato particolare in cui si può scommettere sull'inversione di un trend. Le strategie sono:

- *Consecutive*: si ricerca un trend per mezzo di una sliding window, ovvero una finestra temporale di un determinato numero di giorni consecutivi in cui la direzione di un'azione non muta.
- *SMA*: si ricerca l'inversione di un trend osservando i valori assunti dalla simple moving average.
- *MACD*: si ricerca l'inversione di un trend osservando i valori assunti dalla convergence/divergence moving average.
- *Volume filter*: si applica un filtro sul volume scambiato per ciascuna delle strategie precedenti.

Se queste strategie riscontrano una possibile inversione del trend si procede addestrando i classificatori con i dati restituiti dal blocco precedente per predire le etichette dei giorni successivi.

5.3.1 Consecutive

Il primo passo di questa strategia è individuare un trend per mezzo di una sliding window, ovvero una finestra di W giorni in cui la direzione di un'azione è continua e concorde. Indicando il prezzo di chiusura come cp la direzione di un'azione s, per il giorno d_t può essere definita come:

$$dir(s, d_t) = \begin{cases} up & \text{if } \frac{cp_t - cp_{t-1}}{cp_{t-1}} \cdot 100 \geq 1 \\ down & \text{if } \frac{cp_t - cp_{t-1}}{cp_{t-1}} \cdot 100 \leq -1 \\ stable & \text{altrimenti} \end{cases} \tag{5.42}$$

Come si può evincere da questa formula, viene definita *up* una variazione positiva al di sopra dell'1% indicante quindi un aumento del prezzo dell'azione s al giorno d_t rispetto al giorno d_{t-1}. Analogamente viene definita *down* una variazione negativa al di sotto del -1% indicante quindi un ribasso del prezzo dell'azione s al giorno d_t rispetto al giorno precedente. Nel caso invece la variazione di prezzo sia compresa tra -1% e 1% si registra una situazione stazionaria senza una significativa differenza.

L'individuazione di un trend, indicato in formula come trigger, può essere calcolato come:

$$trigger_{5.3}(d_t) = \begin{cases} uptrend & \text{if } dir(s, d_t), \ldots, dir(s, d_{t-W}) = up \\ downtrend & \text{if } dir(s, d_t), \ldots, dir(s, d_{t-W}) = down \end{cases} \tag{5.43}$$

Non sono ammessi segnali *stable* nell'individuazione di un trend (algoritmo 2, riga 13).

Una volta che un trend è stato individuato si definisce il target come il tipo di reversal su cui scommettere. Se è stato individuato un uptrend si scommette sul downtrend generando come target il segnale *sell*, mentre se è stato individuato un downtrend si scommette sull'uptrend generando il segnale *buy* (algoritmo 2, righe 14-19).

$$target_{5.3} = \begin{cases} S\ (sell) & \text{if } trigger_{5.3}(d_t) == uptrend \\ B\ (buy) & \text{if } trigger_{5.3}(d_t) == downtrend \end{cases} \tag{5.44}$$

Brevemente, si mostra nell'algoritmo 2 il funzionamento di questo processo. Come si può evincere da esso, la ricerca di un trend non parte dall'inizio dell'anno ma dopo un periodo indicato come *min_train_days* (algoritmo 2, riga 9) che rappresenta il numero minimo di giorni scelto come training set per il classificatore.

Algoritmo 2 Strategia Consecutive. Utilizzata per individuare un trend consecutivo e quindi individuare un trend reversal trigger.

procedure CONSECUTIVE($df\_scaled, W, min\_train\_days, idx\_end, labels, N$)
▷ df_scaled è il dataframe calcolato nella Sezione 5.2
▷ W è la grandezza della sliding window
▷ min_train_days è il numero minimo di giorni dell'anno che servono da training set per il classificatore. I giorni seguenti fungono invece da test set.
▷ idx_end rappresenta l'ultimo giorno di mercato dell'anno
▷ $labels$ è il vettore di etichette di ciascun giorno dell'anno
▷ N è la dimensione della finestra di giorni di predizione.
▷ Initialization
$idx\_start \leftarrow min\_train\_days + W$
▷ Execution
while $idx\_start < idx\_end$ **do**
$window \leftarrow labels[idx\_start - W : idx\_start]$
if $window == uptrend$ **or** $window == downtrend$ **then**
if uptrend **then**
$target_{5.3} \leftarrow S$ ▷ sell
end if
if downtrend **then**
$target_{5.3} \leftarrow B$ ▷ buy
end if
. . . . . ▷ Sezione 5.4 e Sezione 5.5
$idx\_start \leftarrow idx\_start + N$ ▷ Se si è trovato un trigger, l'intera finestra va saltata
else
$idx\_start \leftarrow idx\_start + 1$
end if
end while
end procedure

5.3.2 SMA

Questa strategia prevede l'utilizzo di un valore calcolato sulla simple moving average per capire se si può scommettere su un reversal. Si fa uso di un indicatore, appartenente alla categoria *MA* descritta nella Sezione 5.1,

chiamato *SMA5-20* che indica la differenza tra il valore di una SMA calcolata su un periodo di 5 giorni con una SMA calcolata su un periodo di 20 giorni. Questo indicatore è composto da un numero che può essere inferiore o superiore a 0. Quando il valore è inferiore a 0 vuol dire che i valori rappresentati dalla SMA breve sono inferiori a quelli della SMA lunga, analogamente quando il valore dell'indicatore è positivo vuol dire che la SMA breve è 'sopra' a quella lunga.

$$\text{SMA5} - 20(d_t) = \begin{cases} > 0 & \text{if SMA5}(d_t) > \text{SMA20}(d_t) \\ < 0 & \text{if SMA5}(d_t) < \text{SMA20}(d_t) \end{cases} \tag{5.45}$$

Si utilizzano una media breve e una più lunga per un motivo specifico. Una media breve sarà molto più vicina ai prezzi correnti e avrà come conseguenza una maggiore tempestività ma sarà soggetta ad un maggior numero di falsi segnali. Al contrario una media lunga darà origine a una linea più smussata che fornirà suggerimenti più affidabili, ma più in ritardo rispetto all'altra.

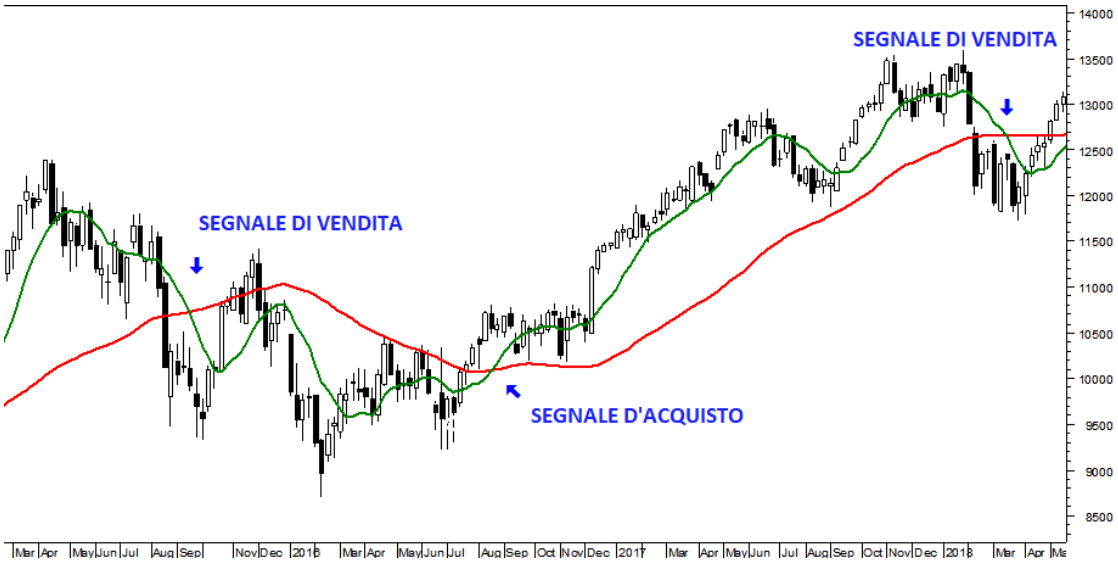

Figura 5.2: Incroci di due medie mobili a breve e lungo periodo [42]

Nell'utilizzo di due medie, quando la più breve taglia all'insù quella lunga verrà generato un segnale d'acquisto, al contrario un segnale di vendita verrà generato. Si prenda in esempio la figura 5.2 dove la linea verde rappresenta una media mobile breve mentre la linea rossa rappresenta una media mobile lunga.

Algoritmo 3 Strategia SMA. Utilizzata per individuare un incrocio di due medie mobili e conseguentemente un possibile trend reversal.

1: **procedure** SMA($df\_scaled, min\_train\_days, idx\_end, labels, N$)
2: ▷ df_scaled, min_train_days, idx_end, $labels$, N sono gli stessi dell'algoritmo 2
3: $idx\_start \leftarrow min\_train\_days$
4: **while** $idx\_start < idx\_end$ **do**
5: **if** $trigger_{5.3}(idx\_start) == buy$ **or** $trigger_{5.3}(idx\_start) == sell$ **then**
6: $target_{5.3} \leftarrow$
7: ..... ▷ Sezione 5.4 e Sezione 5.5
8: $idx\_start \leftarrow idx\_start + N$ ▷ Se si è trovato un trigger, l'intera finestra va saltata
9: **else**
10: $idx\_start \leftarrow idx\_start + 1$
11: **end if**
12: **end while**
13: **end procedure**

Per questa strategia si monitorano progressivamente i giorni a coppie per cercare il trigger rappresentato dal taglio delle due medie che si verifica quando nel giorno d_t il valore dell'indicatore SMA ha segno opposto rispetto al giorno precedente. Indicando d_t come il giorno corrente, il trigger (algoritmo 3, riga 5) e il target (algoritmo 3, riga 6) si calcolano rispettivamente come:

$$trigger_{5.3}(d_t) = \begin{cases} sell & \text{if SMA5} - 20(d_{t-1}) > 0 \ \ \& \ \ \text{SMA5} - 20(d_t) < 0 \\ buy & \text{if SMA5} - 20(d_{t-1}) > 0 \ \ \& \ \ \text{SMA5} - 20(d_t) < 0 \end{cases} \tag{5.46}$$

$$target_{5.3} = \begin{cases} S \ (sell) & \text{if } trigger_{5.3}(d_t) == sell \\ B \ (buy) & \text{if } trigger_{5.3}(d_t) == buy \end{cases} \tag{5.47}$$

Brevemente, si mostra nell'algoritmo 3 il funzionamento di questo processo. Differentemente dalla strategia *consecutive* (algoritmo 2) qui non si fa uso di un parametro W.

5.3.3 MACD

Analogamente alla strategia basata su SMA, anche questa fa uso di un incrocio tra medie mobili, più precisamente la MACD, già spiegata nella Sezione 5.1. Esattamente come nella precedente quando il valore è inferiore a 0 vuol dire che i valori rappresentati dalla MACD breve sono inferiori a quelli della MACD lunga, analogamente quando il valore dell'indicatore è positivo vuol dire che la MACD breve è 'sopra' a quella lunga.

In questo caso il trigger diventa:

$$trigger_{5.3}(d_t) = \begin{cases} sell & \text{if MACD12} - 26(d_{t-1}) > 0 \ \ \& \ \ \text{MACD12} - 26(d_t) < 0 \\ buy & \text{if MACD12} - 26(d_{t-1}) > 0 \ \ \& \ \ \text{MACD12} - 26(d_t) < 0 \end{cases} \tag{5.48}$$

mentre il target rimane pressoché lo stesso.

5.3.4 Filtro sul Volume Scambiato per Azione

Il filtro sul volume (volume filter) viene applicato a ciascuna delle strategie precedenti. In pratica, ogni volta che viene individuato un trigger si effettua un controllo sul volume scambiato calcolato, usando la sezione degli identificatori dedicata allo storico dei prezzi descritta nella Sezione 5.1, come la differenza tra il volume scambiato dell'azione specificata nel giorno in cui il trigger è stato identificato (in un sistema reale si tratterebbe del giorno corrente) con la media del volume scambiato della stessa negli ultimi Y giorni, considerando anche quello corrente. Y assume valori diversi a seconda della strategia utilizzata (algoritmo 4, righe 3-9).

- per la strategia consecutive, Y corrisponde alla grandezza della sliding window, per cui Y == W.
- per le strategie basate su medie mobili, Y corrisponde al periodo della media mobile più breve. Per cui per la strategia SMA, Y corrisponde a 5, mentre per la strategia MACD, Y vale 12.

Si procede con l'addestramento del classificatore se e solo se il risultato ottenuto da questa differenza è concorde con il target ipotizzato. Per fare un esempio, se il target è B (buy), per poter continuare, la differenza di volume deve essere positiva, viceversa se il target è S (short) allora per poter continuare, la differenza di volume deve essere negativa (algoritmo 4, righe

Algoritmo 4 Filtro sul volume utilizzato in combinazione con ciascuna delle altre strategie.

```
1: procedure STRATEGIA(..., Y)
2:     ▷ parametri variano in base alla strategia scelta
3:     if strategy == consecutive then
4:         Y ← W
5:     else if strategy == SMA then
6:         Y ← 5                          ▷ 5 è il periodo della SMA più breve
7:     else if strategy == MACD then
8:         Y ← 12                         ▷ 12 è il periodo della MACD più breve
9:     end if
10:    while idx_start < idx_end do
11:        ......
12:        target_5.3 ←
13:        if (target_5.3 == B and Δvolume(Y) > 0) or (target_5.3 == S
   and Δvolume(Y) < 0) then
14:            ......                     ▷ Sezione 5.4 e Sezione 5.5
15:            idx_start ← idx_start + N  ▷ Se si è trovato un trigger,
   l'intera finestra va saltata
16:        else
17:            idx_start ← idx_start + 1
18:        end if
19:    end while
20: end procedure
```

13).

$$\Delta volume(Y) = volume(s, d_t) - \frac{\sum_{i=0}^{Y-1} volume(s, d_{t-i})}{Y} \tag{5.49}$$

5.4 Addestramento di un classificatore

Una volta innescato il trigger, il classificatore scelto viene addestrato con i dati filtrati nella sezione 5.2.

Il classificatore predice fino a N etichette dove N, chiamato anche orizzonte di predizione, indica il numero di giorni successivi al corrente per cui il

classificatore deve predire la direzione. Indicando l'etichetta di classe, di un giorno d_t, come $label(d_t)$, il classificatore predirà $label(d_{t+1}), \ldots, label(d_{t+N})$ (algoritmo 5, righe 5-8). E' importante notare come l'etichetta $label(d_{t+1})$ viene prodotta per il giorno d_t poiché indica la direzione che il prezzo assumerà nel giorno d_{t+1} rispetto al giorno d_t. Per cui, se per esempio $label(d_{t+1}) == B$ vuol dire che tra il giorno d_t e il giorno d_{t+1} il classificatore ha predetto un aumento del prezzo. Brevemente il comportamento del classificatore lo si può rappresentare come:

Algoritmo 5 Comportamento del classificatore

```
1:  procedure STRATEGIA(. . . . . )
2:      . . . . . .                                  ▷ Sezioni precedenti
3:      n ← 1
4:      labels ← []
5:      while n ≤ N do
6:          labels[n − 1] ← classifier.predict()
7:          n ← n + 1
8:      end while
9:      . . . . . .                                  ▷ Sezione 5.5
10: end procedure
```

5.5 Previsione di un'inversione di trend

Una volta che le N etichette sono state predette si controlla che la predizione di un trend reversal, eseguita nella sezione 5.3, coincida con i valori ottenuti dalla sezione precedente. Questo viene verificato tramite il *majority voting*, una tecnica che, dato in input un dataset con valori appartenenti ad insiemi diversi, restituisce come output il valore dell'insieme che si trova in maggioranza all'interno del dataset di partenza (algoritmo 6, riga 7).

In questo caso, il valore di un'etichetta di un'azione s in un giorno d_t viene calcolata come:

$$val(label(s, d_t)) = \begin{cases} +1 & \text{if } label(s, d_t) == B \\ 0 & \text{if } label(s, d_t) == H \\ -1 & \text{if } label(s, d_t) == S \end{cases} \tag{5.50}$$

Il majority voting non fa altro che sommare i valori delle N etichette restituendo il risultato.

$$mv(labels) = \sum_{i=1}^{N} val(label(s, d_i)) \tag{5.51}$$

dove *labels* è l'insieme di $label(s, d_{t+1}), \ldots, label(s, d_{t+N})$. Una volta ottenuto il valore, il target si calcola come:

$$target_{5.5} = \begin{cases} B & \text{if } target_{5.3} == B \ \ \& \ \ mv(labels) > 0 \\ S & \text{if } target_{5.3} == S \ \ \& \ \ mv(labels) < 0 \\ N & \text{altrimenti} \end{cases} \tag{5.52}$$

quindi per procedere con il trading il majority voting deve confermare il target definito nella Sezione 5.3 (algoritmo 6, riga 10).

Per spiegare meglio il tutto con un esempio supponiamo che una delle strategie spiegate in precedenza restituisca come target B, ovvero si prevede un'inversione del trend, nei giorni successivi, da short a long. Il classificatore, addestrato opportunamente, restituisce un vettore di N=5 etichette definito come *labels = [B, B, H, S, H]*. A questo punto il majority voting, tramite il comportamento descritto precedentemente, restituisce +1 indicando quindi che c'è una corrispondenza tra le etichette predette, o meglio la direzione delle stesse, con il target definito dalla strategia. Se per esempio il classificatore avesse predetto *labels = [B, B, S, S, S]* non ci sarebbe stata corrispondenza tra il target B definito dalla strategia e la direzione assunta dalle etichette predette, maggioranza di S.

Per finire, i dati vengono convertiti nel formato con cui la Sezione 5.6 opera (algoritmo 6, righe 11-13, 15-17 e 22). Questo consiste nel creare, per ciascun giorno di mercato, un'etichetta composta X_Y dove X corrisponde all'etichetta predetta mentre Y corrisponde al target. X può assumere i valori $[N, B, S, H]$ mentre Y può assumere i valori $[N, B, S]$. Possono essere prodotte tre tipi di etichette composte:

- N_N: indica che, per un dato giorno, non si è identificato né un'etichetta né un target e quindi non bisogna operare. Banalmente, se un'etichetta così viene prodotta è perché non è stato innescato il trigger della Sezione 5.3.

Algoritmo 6 Majority Voting e creazione etichette composte

```
1: procedure STRATEGIA(.....)
2:     .....                                                    ▷ Sezione 5.3
3:     while idx_start < idx_end do
4:         if trigger_5.3 ← ... then                            ▷ Sezione 5.3
5:             target_5.3 ← ...                                 ▷ Sezione 5.3
6:             labels ← ...                                     ▷ Sezione 5.4
7:             mv_value ← mv(labels)
8:             n ← 0
9:             composite_labels ← []
10:            if target_5.5 == B or target_5.5 == S then
11:                while n < N_horizon do
12:                    composite_labels[n] ← labels[n]_target_5.5
13:                end while
14:            else
15:                while n < N_horizon do
16:                    composite_labels[n] ← labels[n]_N
17:                end while
18:            end if
19:            ...                                              ▷ Sezione 5.3
20:        else
21:            ...                                              ▷ Sezione 5.3
22:            composite_labels ← N_N
23:        end if
24:    end while
25: end procedure
```

- X_N: indica che il trigger è stato innescato, il classificatore ha prodotto delle etichette ma il risultato del majority voting non è concorde con il target della Sezione 5.3. L'etichetta viene comunque inserita mentre il target viene indicato con N per cui il modulo di trading non opererà in questo giorno.
- X_Y: indica che il target è stato trovato e che quindi il modulo di trading potrà operare.

Per riassumere e visualizzare i concetti espressi nelle ultime sezioni si prendano in esempio le seguenti figure.

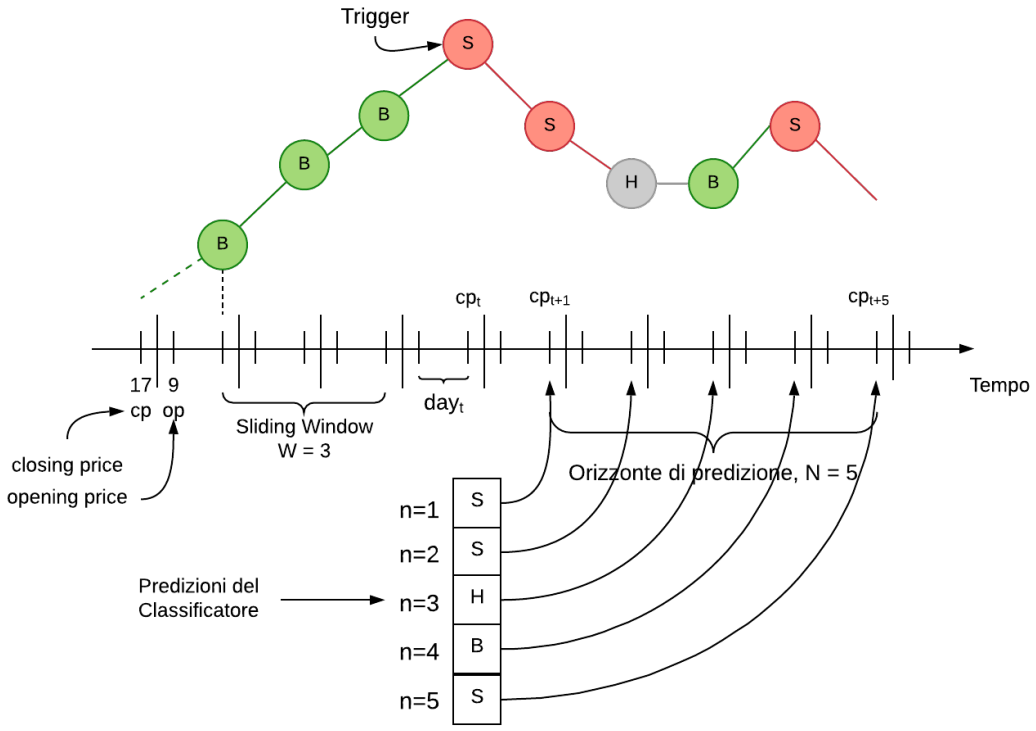

Figura 5.3: Strategia consecutive

La figura 5.3 mostra il comportamento che il sistema segue utilizzando la strategia *consecutive*. Nel giorno day_t, utilizzando una sliding window con W=3, viene rilevato un uptrend in corso dato che negli ultimi 3 giorni la direzione del prezzo dell'azione è stata costante e in aumento. A questo punto, poco prima delle ore 17 del giorno corrente, il classificatore viene addestrato e predice, scegliendo una finestra di predizione con N=5, la direzione del prezzo dell'azione per i successivi 5 giorni. La direzione predetta fa riferimento al prezzo di chiusura che l'azione avrà nei giorni successivi, per cui se per n=1 l'etichetta predetta è S, vuol dire che si prevede che il prezzo con cui l'azione chiuderà nel giorno immediatamente successivo sarà più basso del prezzo attuale. Nella Sezione successiva verrà spiegato meglio questo concetto. A questo punto il majority voting riceve in ingresso le etichette predette e restituisce il valore -1 (equazione 5.51). Siccome il trend identificato precedentemente era un uptrend, si sarebbe dovuto scommettere su un downtrend (target). Dato che il risultato del majority voting è concorde

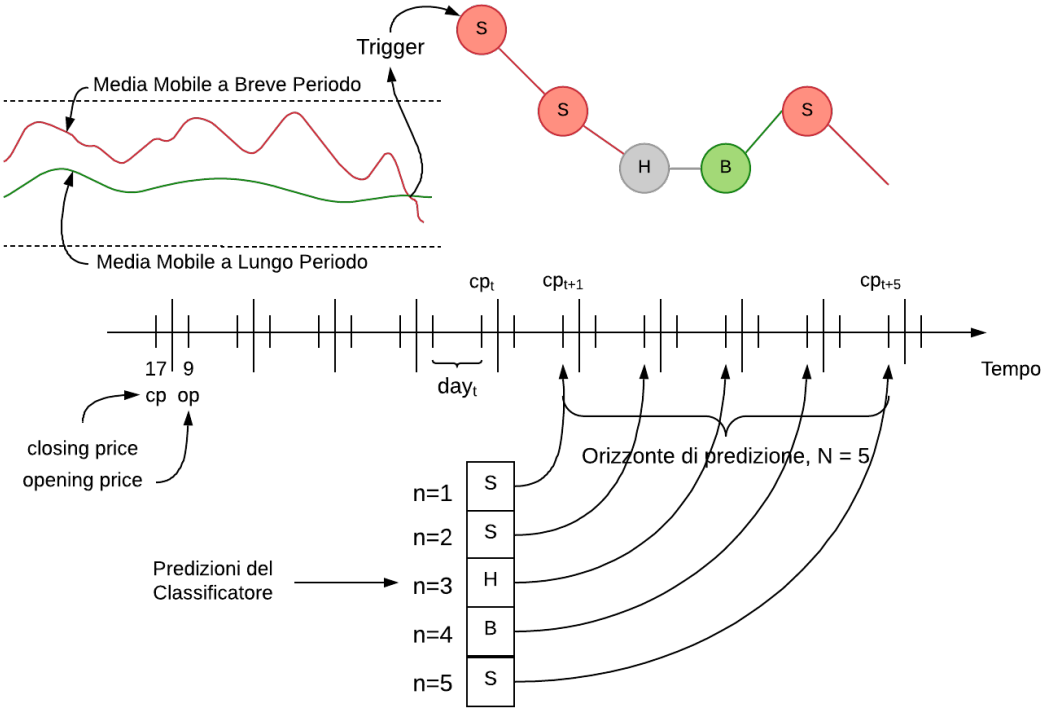

Figura 5.4: Strategie basate su medie mobili

con il target ipotizzato in precedenza, si può procedere con l'apertura di una posizione short-selling nel modulo di trade.

La figura 5.4 mostra il comportamento che il sistema segue nei casi delle strategie basate sulle medie mobili, ovvero la strategia SMA e quella MACD. L'unica cosa che cambia ovviamente è che la linea rossa (periodo breve) e la linea verde (periodo lungo) rappresentate in figura corrispondono in un caso alla SMA a 5 e 20 periodi e nell'altro al MACD calcolato su 12 e 26 periodi. In questo caso, il valore della differenza delle medie letto nel giorno day_t è negativo (media breve sotto la media lunga). Siccome nel giorno precedente il valore era positivo (media breve sopra alla media lunga) si può considerare questa situazione come un possibile ribasso nei prezzi. Il target diventa S e si procede esattamente come per la figura precedente.

La figura 5.5 infine mostra il filtro sul volume applicato a una qualsiasi delle strategie precedenti. Come si evince, per il giorno day_t, considerando

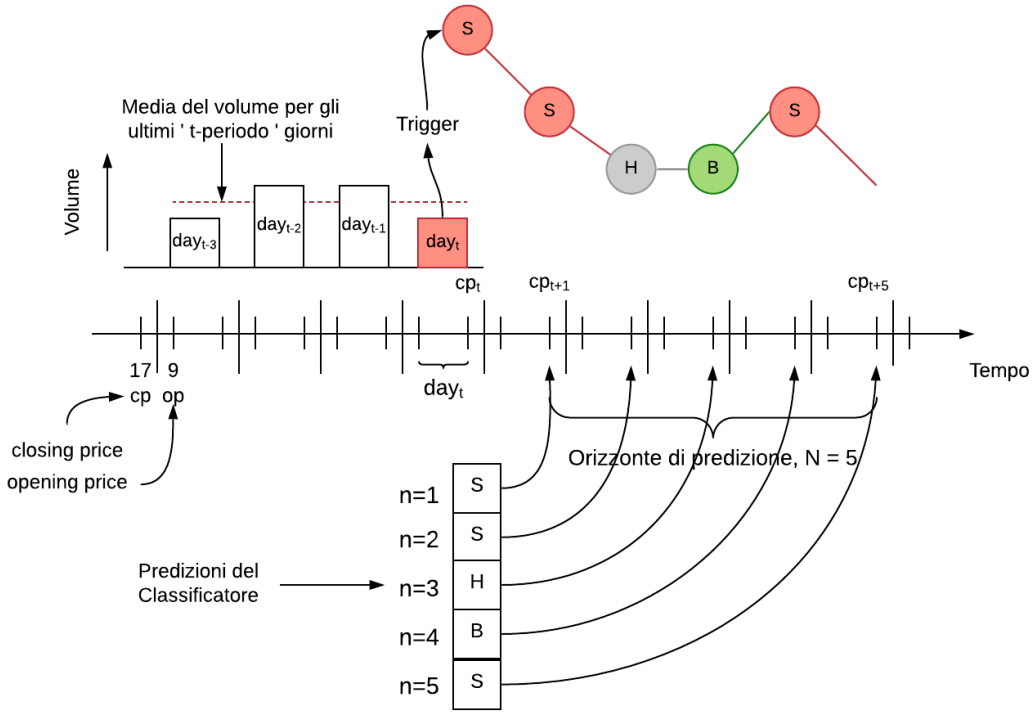

Figura 5.5: Filtro sul volume

un periodo di 3 giorni precedenti al corrente, viene calcolata la media del volume scambiato, rappresentata in figura dalla linea rossa tratteggiata. Dato che il volume scambiato nel giorno corrente è minore del volume medio degli ultimi 3 giorni, ed essendo S il target delle strategie precedenti, il filtro risulta concorde e quindi si può procedere con gli altri step.

5.6 Gestione del trade

Il modulo di trading utilizza solamente il *target* come informazione per aprire una posizione. Nonostante questo l'etichetta predetta del giorno viene comunque passata, sia per non perdere l'informazione su che tipo di segnale il classificatore ha predetto (B, S o H), sia per raccogliere statistiche che serviranno nelle analisi finali.

Al modulo di trading viene passato un dataset caratterizzante un dato anno con all'interno i segnali generati per ciascuna azione per ciascun giorno di mercato (tabella 5.4).

Tabella 5.4: Dataset utilizzato come input del modulo di trading

Giorno	...	$\mathbf{S_i}$	$\mathbf{S_{i+1}}$	$\mathbf{S_{i+2}}$	...
...	...	...	...	...	...
day_{t-1}	...	N_N	N_N	N_N	...
day_t	...	N_N	L_N	H_S	...
day_{t+1}	...	N_N	H_N	S_S	...
day_{t+2}	...	N_N	L_N	S_S	...
day_{t+3}	...	N_N	N_N	N_N	...
...	...	...	...	...	...

Come si può evincere da questa tabella, il modulo di trading:

- per l'azione S_i non aprirà alcuna posizione.
- per l'azione S_{i+1} non aprirà alcuna posizione perché la direzione assunta dalle etichette predette dal classificatore non coincide con il $\text{target}_{5.3}$.
- per l'azione S_{i+2} aprirà una posizione short-selling nel giorno day_t.

Le posizioni vengono sempre aperte poco prima della chiusura del mercato, questo per evitare i relativi problemi nel dover aprire la posizione all'inizio del giorno dopo. Si faccia riferimento alla figura d'esempio 5.6. Nel giorno d_t viene predetta l'etichetta B, buy, che indica che nel giorno d_{t+1} il prezzo dell'azione s sarà più alto.

- Se si apre la posizione alle ore 17 del giorno d_t, ovvero subito dopo che l'etichetta venga predetta, si arriverebbe alla chiusura del mercato del

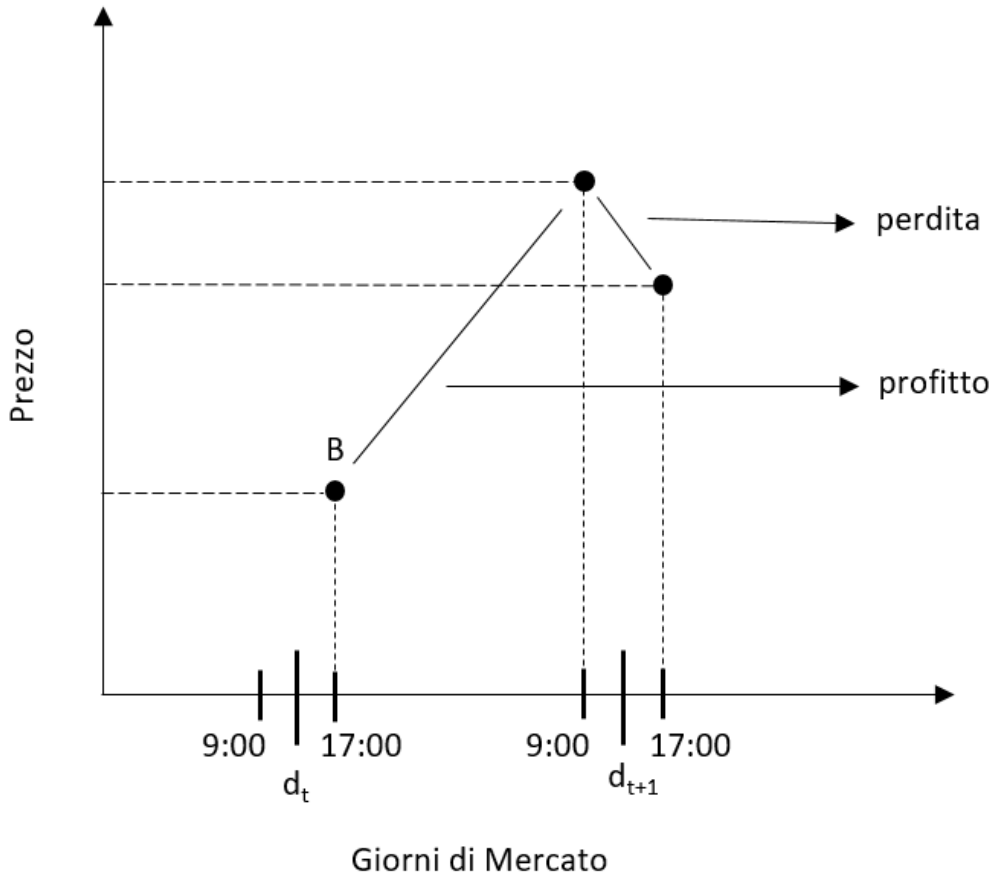

Figura 5.6: Andamento del prezzo di un'azione s

giorno d_{t+1} avendo riscontrato un profitto poiché il prezzo dell'azione s, alla fine del giorno d_{t+1} risulta essere più alto di quello della stessa azione quando la posizione è stata aperta nel giorno precedente.

- Se si apre la posizione alle ore 9 del giorno d_{t+1} si aprirebbe una posizione long-selling con un prezzo dell'azione s molto più alto rispetto al giorno precedente e infatti, una volta arrivati alla chiusura del giorno, la posizione verrebbe chiusa, generando una perdita, poiché avendo aperto long-selling ci si aspetta che il prezzo dell'azione salga quando invece il prezzo, valutato alle 17, è sceso.

Il sistema si comporta così perché le etichette sono predette basandosi sui dati più recenti a disposizione e quindi danno un'informazione su come sarà

il prezzo di un'azione rispetto all'istante corrente non avendo né controllo né informazioni su ciò che accade tra la chiusura di un giorno di mercato e l'apertura del giorno seguente.

Il modulo di trading si comporta nella maniera seguente: per ogni giorno d_t (ogni riga, algoritmo 7, righe 5 e 17) del dataframe in ingresso e conseguentemente per ogni azione $s(d_t)$ (ogni colonna, algoritmo 7, righe 7 e 12) del dataframe, si segue il flusso descritto in figura 5.7.

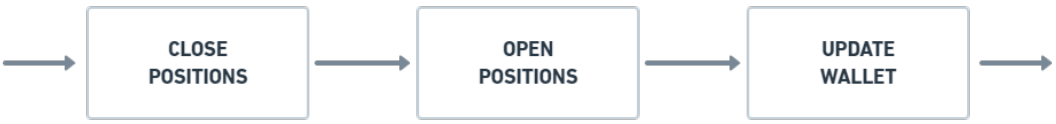

Figura 5.7: Processo di trading dato un giorno d_t e un'azione $s(d_t)$.

- **Close positions**: come prima cosa il sistema controlla tra le posizioni correntemente aperte se ce ne sono alcuna da chiudere. Una posizione può essere chiusa se:

  - non è più presente un target. Questo capita quando non ci si trova più all'interno della finestra di etichette predette dal classificatore. Osservando la tabella 5.4 si può notare che questa condizione si verifica nel giorno d_{t+3} per l'azione S_{i+2}.

  - si è superata la soglia costituita dallo *stop_loss*, ovvero un valore percentuale che indica entro che soglia una perdita è considerata accettabile.

    Mettiamo il caso che una posizione short-selling venga aperta nel giorno d_t per l'azione s_i con il prezzo d'acquisto dell'azione indicato come $pa(d_t, s_i)$. Se nel giorno d_{t+1} la differenza percentuale tra il valore massimo assunto dal prezzo dell'azione s_i e il $pa(d_t, s_i)$ è maggiore della stop loss allora la posizione viene chiusa. Viceversa, per una posizione long-selling, se la differenza percentuale tra il valore minimo assunto dal prezzo dell'azione e il $pa(d_t, s_i)$ è minore della stop loss, negativa, allora la posizione viene chiusa. Indicando con $p_{high}(d_t, s_i)$ e $p_{low}(d_t, s_i)$ rispettivamente il prezzo più alto e più basso assunto dall'azione s_i nel giorno d_t, il caso in cui una posizione

possa essere chiusa in un giorno d_{t+1} a causa della stop loss si può schematizzare come:

$$\begin{cases} true & \text{if } target == B \textbf{ and } \frac{p_{low}(d_{t+1},s_i) - pa(d_t,s_i)}{pa(d_t,s_i)} \leq -stop\_loss \\ true & \text{if } target == S \textbf{ and } \frac{p_{high}(d_{t+1},s_i) - pa(d_t,s_i)}{pa(d_t,s_i)} \geq stop\_loss \\ false & \text{altrimenti} \end{cases} \tag{5.53}$$

Una volta chiusa una posizione, il budget verrà ridimensionato con ciò che la posizione ha ritornato, perdita o guadagno che sia (algoritmo 7, riga 14) più il costo di una tassa calcolata come lo 0.5% sul capitale investito per quella posizione.

- **Open positions**: controlla innanzitutto che per l'azione corrente non ci sia già una posizione aperta e in caso negativo ne apre una in base al valore del target, quindi long-selling, in caso il target sia B, short-selling nel caso il target sia S (algoritmo 7, riga 15). L'investimento dedicato a ciascuna azione è pari al 10% del budget corrente diviso equamente per il numero delle azioni che devono essere aperte nel giorno di mercato corrente. Se in un dato giorno d_t ci sono X azioni che devono essere aperte, l'investimento per ciascuna azione s_i è :

$$\text{investimento}(d_t, s_i) = \frac{\text{budget corrente}}{10 \cdot X} \tag{5.54}$$

- **Update Wallet**: sottrae al budget l'investimento complessivo della giornata corrente (algoritmo 7, riga 16).

Si può dedurre come il ciclo di vita delle posizioni sia limitato solamente alle finestre di predizione e che quindi non si può aprire, o tenere aperta, una posizione quando il target è N. Inoltre non è possibile aprire contemporaneamente per una singola azione più di una posizione. Ovviamente questo non vuol dire che si potrà avere una sola posizione aperta per finestra di predizione, perché se una posizione viene chiusa in uno specifico giorno si potrà sempre aprirne un'altra purché non si ecceda la finestra stessa. Questo porta il numero minimo di possibili posizioni aperte in una finestra ad essere pari a 1 (si apre una posizione il primo giorno e la si tiene aperta fino alla fine della finestra) e porta il numero massimo a coincidere con la dimensione della

finestra di predizione stessa (ogni giorno si chiude la posizione precedente per stop loss e se ne riapre un'altra).

L'algoritmo che descrive brevemente il tutto è:

Algoritmo 7 Gestione del trading automatica.

procedure TRADING($df, stop\_loss$)
    ▷ df è il dataframe descritto dalla tabella 5.4
    $pos\_aperte \leftarrow []$
    $d_t \leftarrow d_0$ ▷ Si inizia dal primo giorno disponibile.
    while $d_t \leq d_{last}$ **do**
        $pos\_da\_aprire \leftarrow []$
        $s_i(d_t) \leftarrow s_0(d_t)$ ▷ Si parte dalla prima azione nel dataframe.
        while $s_i(d_t) \leq s_{last}(d_t)$ **do**
            if $target_{5.5}(s_i(d_t))$ **then**
                $pos\_da\_aprire \leftarrow s_i(d_t)$
            end if
            $s_{i++}(d_t)$
        end while
        $close\_positions(pos\_aperte, target, stop\_loss)$ ▷ Si chiude una posizione se non è più presente un target o si supera la soglia della stop loss
        $open\_positions(pos\_da\_aprire)$
        $update\_budget$
        d_{t++}
    end while
end procedure

Capitolo 6

Esperimenti

L'obiettivo di questa tesi è di proporre un sistema di trading quantitativo che investe sul mercato azionario mediante una strategia multiday di tipo trend reversal basata sull'addestramento di algoritmi di machine learning al fine di dimostrare l'efficacia rispetto ad un approccio tradizionale basato sull' analisi tecnica e definire le configurazioni più appropriate del sistema analizzando l'impatto di vari fattori sui risultati della simulazione trading.

Partendo dal lavoro svolto in [41] questa tesi si propone di ampliarne il modello tramite (i) l'applicazione di nuove strategie ai fini del riconoscimento di un reversal e tramite (ii) l'utilizzo di una strategia di trading basata sul majority voting in cui il tipo di posizione da aprire (long-selling o short-selling) viene determinato non più in base alla singola etichetta prodotta dal classificatore ma in base al tipo di reversal ipotizzato sull'intera finestra di predizione. Inoltre parte del lavoro svolto è servito come sperimentazione volta a comprendere l'impatto dei principali parametri del sistema e la corretta configurazione degli algoritmi in esso integrati.

Nella prima parte di questa Sezione vengono spiegate le configurazioni del sistema utilizzate mentre nella seconda parte vengono analizzati i risultati ottenuti.

6.1 Configurazione degli algoritmi

Dati: sono stati raccolti dati di azioni corrispondenti a 7 anni di mercato, più precisamente dall'inizio dell'anno 2011 alla fine dell'anno 2017. I dati sono relativi alle azioni dell'indice americano Standard & Poor 500 raccolti

tramite le API di Yahoo Finance [43]. Per addestrare i classificatori sono stati utilizzati i descrittori analizzati nella Sezione 5.1. Le informazioni riguardanti le news invece sono state reperite tramite il sito *Reuters* [44]. Siccome le condizioni di mercato cambiano nel tempo, si è deciso di analizzare separatamente i dataset differentemente per ogni anno.

Algoritmi: gli algoritmi utilizzati sono stati spiegati nella Sezione 3. Qui vengono spiegati i settaggi utilizzati. Per l'implementazione degli algoritmi è stata utilizzata la libreria Scikit-learn [45].

- Support Vector Classifier (SVC): kernel = rbf; gamma = $\frac{1}{|D| \cdot \sigma_x^2}$, dove D rappresenta il numero di features mentre σ_x^2 è la varianza; C = 1; random_state = 6.
- Multinomial Naïve Bayes (MNB): $\alpha = 1.0$.
- Gaussian Naïve Bayes (GNB): var_smoothing = 1e-9.
- Random Forest Classification (RFC): criterion = $gini$; n_estimators = 300; random_state = 6.
- K Nearest Neighbor (KNN): k = 5; weights = $uniform$.
- Multilayer Perceptron (MLP): hidden_layers = 1; hidden_layer_size = 23; solver = $lbfgs$.

Nella seconda parte gli algoritmi, così come le features descritte nella Sezione 5.1, verranno indicati con le loro abbreviazioni.

Si è scelto inoltre di fissare il 20% dei giorni dell'anno (all'incirca i primi due mesi) come training set minimo per i classificatori.

Configurazione dei parametri relativi alla strategia di riconoscimento di un trend reversal: La sliding window W assume i valori 3, 4 e 5, l'orizzonte di predizione assume i valori 3, 5.

Trading: Si assume, per ogni anno:

- Un budget iniziale di 100.000 USD.
- Una tassa per ogni operazione pari al 0.5% del capitale investito.

- Un investimento per ogni giorno di mercato pari al 10% del budget corrente.
- Una distribuzione uniforme dell'investimento tra ciascuna azione per ogni giorno di mercato in cui ci sia almeno una posizione aperta.
- Stop loss fissata a 1%.
- Durata massima di una posizione di N giorni dove N è la dimensione della finestra di predizione.

Misure di Valutazione: per valutare le performance del sistema si fa uso di due misure:

- ***R**itorno percentuale **M**edio di **I**nvestimento per **A**zione* (rmia), calcolato per anno come:

$$rmia = \frac{\text{profitto relativo medio per azione}}{\text{investimento medio per azione}} \tag{6.1}$$

- ***P**rofitto **R**elativo **T**otale percentuale* (prt), calcolato per anno come:

$$prt = \frac{\text{equity}_{\text{end}} \text{ - equity}_{\text{start}}}{\text{equity}_{\text{start}}} \tag{6.2}$$

  dove $\text{equity}_{\text{end}}$ e $\text{equity}_{\text{start}}$ rappresentano rispettivamente il valore dell'equity line a fine e a inizio anno (budget finale e budget iniziale).

Tutti gli esperimenti sono stati condotti su un computer con Intel® Xeon® X5650, 32 GB di RAM e con sistema Ubuntu 18.04.1 LTS integrato.

6.2 Risultati

Le simulazioni sono state fatte utilizzando una configurazione composta da 6 parametri variabili quali:

- **Strategie**: consecutive (Cons), SMA, MACD, volume consecutive (vCons), volume SMA (vSMA) e volume MACD (vMACD).
- **Anni**: dal 2011 al 2017.
- **Dimensione sliding window**: w=3, w=4 e w=5.

- **Dimensione finestra di predizione**: n=3 e n=5.
- **Classificatori**: SVC, RFC, MNB, GNB, KNN e MLP (vedere Sezione 3).
- **Features**: TEMP, VOL, NEWS+TEMP, NEWS+VOL, OSC+TEMP, OSC+VOL, OSC+TEMP+NEWS, OSC+VOL+NEWS e ALL (vedere Sezione 4.1).

Una configurazione comprende quindi un valore per ciascuna categoria fatta eccezione per le strategie SMA, MACD, vSMA e vMACD che non comprendono il parametro w (Sezioni 5.3.2 e 5.3.3).

6.2.1 Analisi della significatività statistica

Per ciascuna simulazione, una volta ottenuti i risultati, è stato applicato il test statistico di Friedman. Questo è un test non-parametrico utilizzato per vedere se due o più gruppi di dati, chiamati anche popolazioni, siano significativamente diversi. Il confronto viene effettuato sulla mediana delle popolazioni. Il test di Friedman si basa su due ipotesi: (i) l'ipotesi nulla rappresenta il fatto che non ci sia una diversità tra le popolazioni e che di conseguenza gli effetti delle popolazioni siano gli stessi, mentre (ii) l'ipotesi alternativa rappresenta invece il fatto che le popolazioni siano significativamente diverse.

Se tramite Friedman è possibile rigettare l'ipotesi nulla si procede con l'utilizzo di un test post hoc per valutare il livello percentuale di diversità tra le popolazioni . Per questo lavoro è stato utilizzato il test di Wilcoxon [46], implementato dalla libreria *numpy* [47].

Wilcoxon confronta le popolazioni a coppie e a seconda della strategia utilizzata restituisce un determinato valore. Per questo lavoro è stata utilizzata la strategia $two - sided$ la quale restituisce come valore un numero decimale rappresentante la percentuale entro cui le due popolazioni possono essere definite significativamente diverse.

Per esempio: date due popolazioni denominate P1 e P2, se *Wilcoxon(P1, P2, two-sided)* restituisce il valore 0.02 (2%) vuol dire che P1 e P2 possono essere considerate significativamente diverse fino ad un livello percentuale di significance del 98% (100 - 2). Più il risultato è vicino allo zero e più le popolazioni si possono considerare diverse .

I risultati sono organizzati in tre diverse sottosezioni: nella prima si confrontano le strategie di identificazione del trend reversal trigger, al fine di trovare la migliore (S_{WIN}), sia in termini di *prt* che in termini di *rmia*. Nella seconda sottosezione, fissata la strategia S_{WIN}, si confrontano le combinazioni di n e w, al fine di trovare i valori migliori (n_{WIN} e w_{WIN}). Nell'ultima sottosezione, fissati S_{WIN}, n_{WIN} e w_{WIN}, si confrontano i classificatori e le features, al fine di trovare il $class_{WIN}$ e la fid_{WIN}.

Ogni sottosezione è divisa per *prt* e *rmia* e per ciascuna sono presenti:

- grafici che mostrano l'andamento del *prt* / *rmia* sui 7 anni per le configurazioni confrontate.

- una tabella statistica ordinata in modo decrescente rispetto al valore di *prt* / *rmia* medio, composta da:

  - **Strategia**: nome della configurazione confrontata.
  - **Average Ranking Value**: posizione in classifica media attribuita alla configurazione considerata.
  - **Valore di rmia/prt medio at 85%**: rappresenta il valore medio di *prt* / *rmia* della configurazione considerata su 7 anni. Inoltre se è presente un asterisco vicino al dato vuol dire che il test di Wilcoxon *two-sided* ha confermato che il valore considerato è significativamente diverso rispetto al primo classificato per un valore percentuale di significance superiore all'85%. Per semplicità i confronti sono stati fatti solo con il valore classificatosi come migliore.

- una tabella riassuntiva composta da:

  - **Strategia**: nome della configurazione confrontata.
  - **Numero di Giorni di Mercato**: numero di giorni in cui viene aperta almeno una posizione di mercato.
  - **Numero Medio di Operazioni per Giorno**: numero medio di posizioni aperte per giorno di mercato (ovvero giorni in cui viene aperta almeno una posizione).
  - **Profitto Relativo Percentuale**: guardare la definizione di *prt* del paragrafo 'Misure di Valutazione' della Sezione 6.1.

- **Profitto Medio per Operazione**: rappresenta il valore del guadagno medio ottenuto per la singola posizione ed è calcolato come *profitto relativo totale / numero di posizioni totali*.
- **Investimento Medio per Operazione**: indica la quantità di dollari investiti mediamente per ciascuna posizione.
- **Ritorno Percentuale Medio per Operazione**: guardare la definizione di *rmia* del paragrafo 'Misure di Valutazione' della Sezione 6.1.

6.2.2 Confronto tra le strategie di identificazione del trend reversal trigger

Come prima cosa le strategie vengono confrontate tra loro per capire quale sia la migliore. Indicando una configurazione come la combinazione di $strategia + horizon + window + classificatore + feature + anno$ i valori medi rappresentati in questa sezione sono calcolati come la media dei valori di tutte le combinazioni possibili di una configurazione tenendo fissa solo la strategia (e l'anno nel caso dei grafici) per un totale di, non considerando gli anni, 324 combinazioni per le strategie consecutive e volume consecutive e 108 combinazioni per le rimanenti (dato che per le strategie basate su medie mobili il valore della window non è presente).

Per i grafici quindi il valore *prt / rmia* medio di una strategia in un preciso anno è calcolato come la media dei *prt / rmia* delle combinazioni costituite dai parametri rimanenti (324 e 108, fissata solo la strategia). Per le tabelle invece il valore *prt / rmia* medio di una strategia è calcolato come la media dei *prt / rmia* precedenti su 7 anni.

Profitto relativo totale percentuale: la figura 6.1 mostra l'andamento del *prt* medio per ciascuna delle strategie su 7 anni. Si può notare come la strategia volume consecutive raggiunga i risultati migliori 7 anni su 7.

Si è applicato in seguito il test di Friedman tramite cui è stato possibile scartare l'ipotesi nulla sottolineando quindi come le strategie siano significativamente diverse tra loro. Nella tabella 6.1 si può notare infatti come la strategia volume consecutive sia la migliore sia come posizione media in classifica (1.17) sia come *prt* medio. In più tramite Wilcoxon è stato possibile riscontrare come l'utilizzo di questa strategia sia significativamente diverso rispetto alle altre (condizione dettata dall'asterisco).

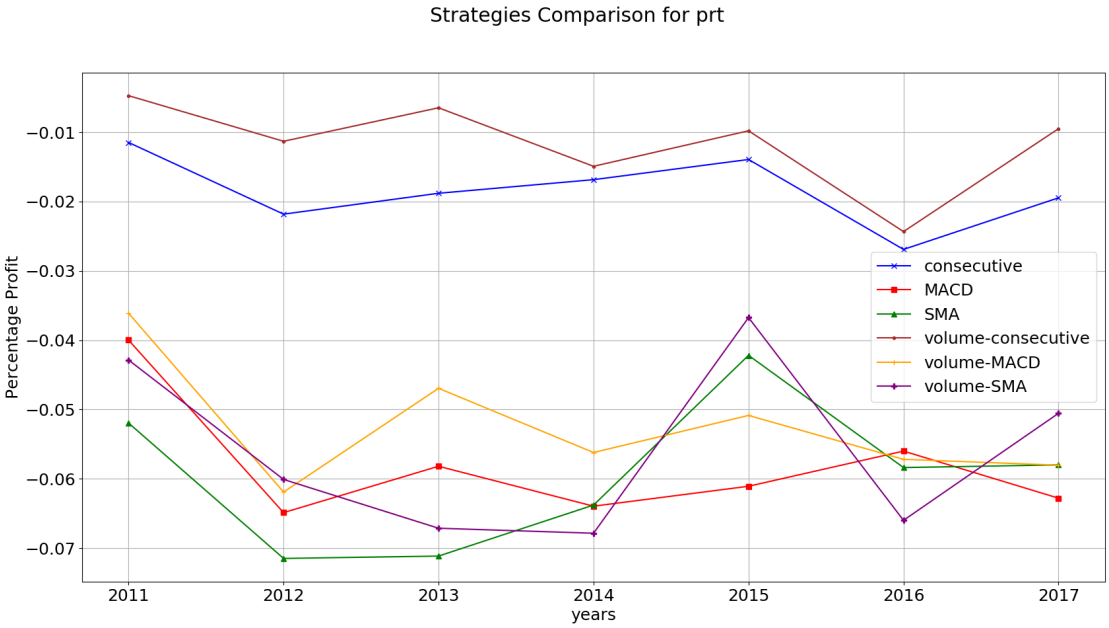

Figura 6.1: Andamento del *prt* medio delle strategie su 7 anni

Dalla tabella 6.1 si può ancora notare come già solo l'utilizzo della strategia consecutive sia migliore rispetto alle strategie SMA (+4.10%) e MACD (+3.96%). Inoltre l'utilizzo di un filtro sul volume migliora il ritorno di ciascuna strategia: si registra un +0.69% per la strategia volume consecutive (rispetto alla consecutive), un +0.57% per la volume MACD (rispetto alla MACD) e infine un +0.36% per la volume SMA (rispetto alla SMA).

Strategia	**Average Rank Value**	**prt at 85%**
volume consecutive	1.17	-1.16%
consecutive	2.33	-1.85%*
volume MACD	4.5	-5.24%*
volume SMA	5.17	-5.59%*
MACD	5.5	-5.81%*
SMA	5.83	-5.95%*

Tabella 6.1: Confronto tra le strategie sul *prt* medio calcolato su 7 anni. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata

Ritorno percentuale medio di investimento per azione: la figura 6.2 mostra l'andamento del *rmia* medio per ciascuna delle strategie su 7 anni e si può notare come la strategia volume consecutive raggiunga i risultati migliori 4 anni su 7 (nel 2016 è il risultato peggiore di tutti), mentre la strategia consecutive raggiung i risultati migliori negli altri 3 anni. Il test

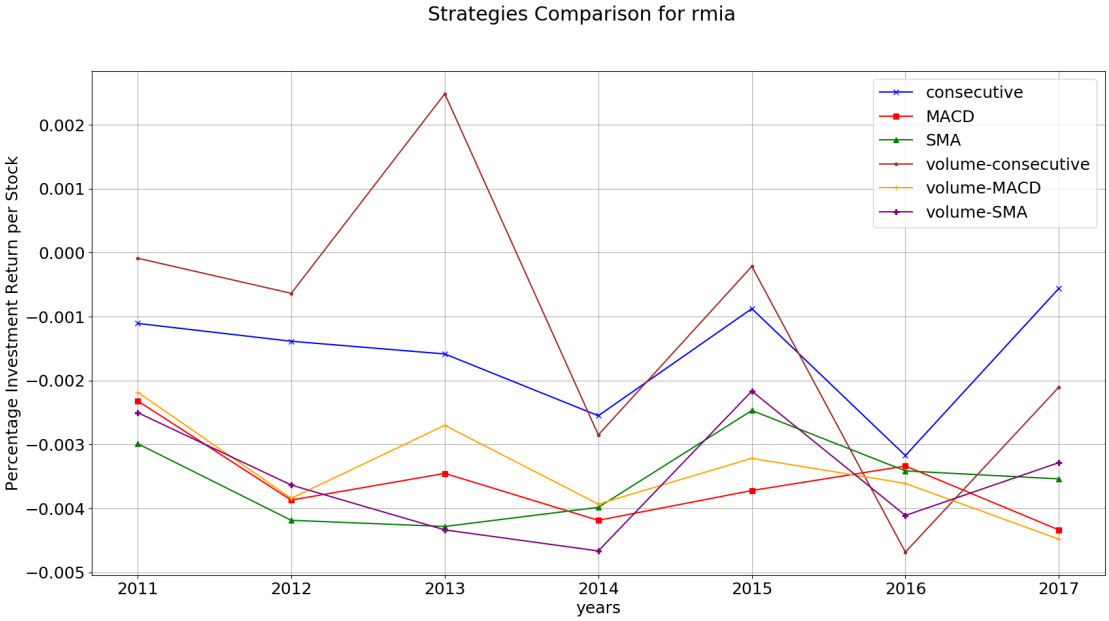

Figura 6.2: Andamento del *rmia* medio delle strategie su 7 anni

di Friedman ha permesso di rigettare l'ipotesi nulla sottolineando quindi come le strategie siano significativamente diverse tra loro. Nella tabella 6.2 possiamo notare come la strategia volume consecutive risulti essere la migliore nonostante sia la seconda migliore classificata nel ranking medio. Inoltre Wilcoxon ha evidenziato, diversamente dal *prt*, come l'utilizzo della strategia volume consecutive sia significativamente diverso rispetto alle strategie SMA, MACD, vSMA e vMACD ma non comporti una significativa differenza rispetto alla strategia consecutive.

Esattamente come per il *prt*, la tabella 6.2 ha evidenziato come la strategia consecutive sia migliore rispetto alle strategie SMA (+0.20%) e MACD (+0.19%) e come l'utilizzo del filtro sul volume comporti un miglioramento, seppur molto più attenuato, nel ritorno di ciascuna strategia: +0.04% per la strategia volume consecutive (rispetto alla consecutive) e +0.01% sia per la

Strategia	**Average Rank Value**	**rmia at 85%**
volume consecutive	2.33	-0.11%
consecutive	1.83	-0.16%
volume MACD	4.67	-0.34%*
volume SMA	5.17	-0.35%*
MACD	5.17	-0.35%*
SMA	5.33	-0.36%*

Tabella 6.2: Confronto tra le strategie sul *rmia* medio calcolato su 7 anni. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata

volume MACD (rispetto alla MACD) che per la volume SMA (rispetto alla SMA).

Nella tabella riassuntiva 6.3 sono riportati i valori più significativi mediati sui 7 anni.

Da questo primo confronto possiamo indicare la strategia volume consecutive come la migliore. Nella prossima sottosezione verranno confrontate le combinazioni dei valori di n (orizzonte di predizione, o horizon) e di w (sliding window) per la strategia volume consecutive.

Strategia	Numero di Giorni di Mercato	Numero Medio di Posizioni Aperte per Giorno	Profitto Relativo Percentuale	Profitto Medio per Operazione	Investimento Medio per Operazione	Ritorno Percentuale Medio per Operazione
vCons	62.80 (16.2)	2.29 (1.0)	-1.16% (1487.8)	-4.56 (38.9)	5591.2 (1812.9)	-0.11%
Cons	81.57 (17.0)	2.80 (1.2)	-1.85% (1646.8)	-5.23 (28.9)	5032.4 (1629.1)	-0.16%
vMACD	169.94 (11.2)	5.67 (1.4)	-5.25% (1776.9)	-7.93 (6.2)	2409.0 (736.4)	-0.34%
vSMA	174.64 (10.3)	6.63 (1.8)	-5.59% (1799.3)	-6.88 (4.9)	2134.3 (670.5)	-0.35%
MACD	182.69 (8.9)	9.21 (2.2)	-5.81% (1736.2)	-5.88 (4.1)	1655.1 (579.1)	-0.35%
SMA	186.00 (7.2)	12.56 (3.3)	-5.96% (1571.9)	-3.88 (2.7)	1279.9 (490.6)	-0.36%

Tabella 6.3: Tabella riassuntiva dei valori medi relativi alle strategie su 7 anni e delle relative deviazioni standard (indicate tra parentesi)

6.2.3 Confronto tra le combinazioni di w e n per la strategia volume consecutive

Dalla sottosezione precedente la strategia volume consecutive si è rivelata essere la scelta migliore. Per essa si contano un totale di: 6 (classificatori) · 9 (features) · 3 (window) · 2 (horizon) = 324 possibili configurazioni.

In questa sottosezione verranno confrontate le combinazioni di n e w per la strategia volume consecutive al fine di identificare i valori migliori. I valori che verranno rappresentati nei grafici e nelle tabelle sono da considerarsi come i valori medi di *prt* / *rmia* calcolati su 54 configurazioni (54 · 7 anni nel caso delle tabelle), dato che, fissati la strategia, n e w, rimangono solo le combinazioni di classificatori (6) e features (9).

Profitto relativo totale percentuale: la figura 6.3 mostra l'andamento del *prt* medio delle combinazioni di n e w su 7 anni per la strategia volume consecutive. Si può notare innanzitutto che l'andamento delle combinazioni con n=3 sia di solito superiore a quello con n=5. Per quanto riguarda w invece si può notare come gli andamenti delle combinazioni migliorino all'aumentare del valore di w.

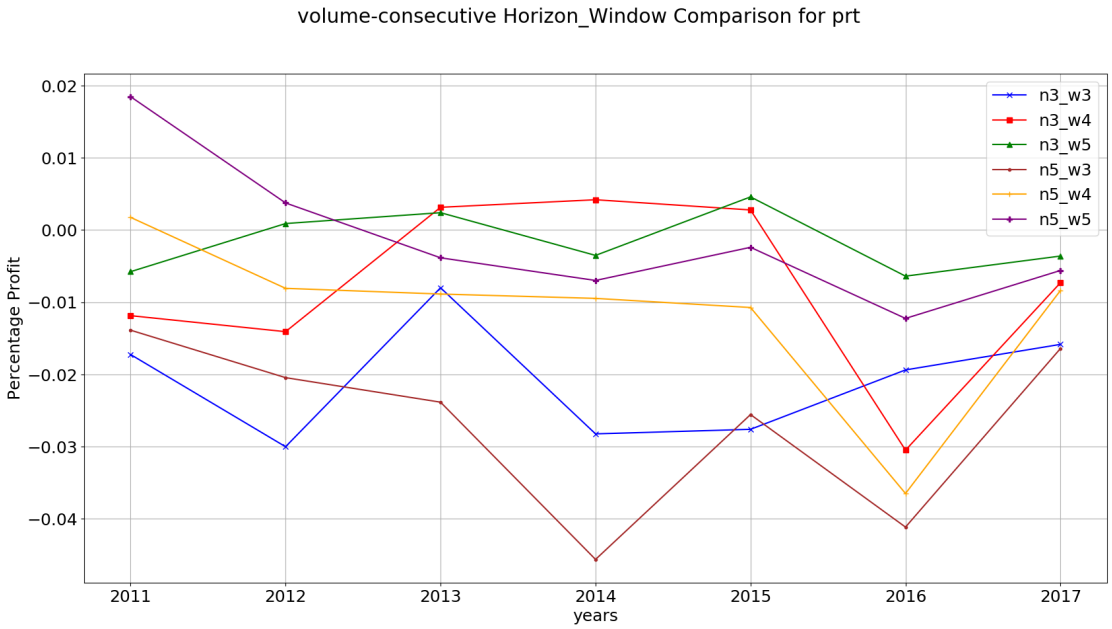

Figura 6.3: Andamento del *prt* medio delle combinazioni di n e w su 7 anni per la strategia volume consecutive

Le migliori combinazioni risultano essere n5_w5 per i primi due anni (2011, 2012), n3_w4 per gli anni intermedi (2013, 2014) e n3_w5 per gli ultimi tre (2015, 2016 e 2017).

Dal grafico quindi è possibile notare come l'andamento del *prt* migliori all'aumentare del valore w e che per ciascuna w, la combinazione con n=3 risulti migliore di quella con n=5.

Dalla tabella 6.4 si può notare come la precedente osservazione sia quasi del tutto concorde con i risultati medi ottenuti sui 7 anni. Infatti le combinazioni con w=5 risultano le migliori seguite da quelle con w=4 e infine da quelle con w=3. L'unica differenza è che la combinazione migliore risulta essere n5_w5 nonostante n3_w5 sia prima nella classifica del ranking medio. Questo è dovuto dal fatto che la strategia n5_w5 nel primo anno ha riscontrato un grande profitto positivo che ne ha alzato la media.

Grazie a Friedman è stato possibile scartare l'ipotesi nulla sottolineando quindi come le combinazioni siano significativamente diverse tra loro. Tramite Wilcoxon invece è stato possibile notare come la combinazione migliore n5_w5 sia significativamente diversa rispetto a n5_w4 (+1.02%), n3_w3 (+1.56%) e n5_w3(+2.54%) mentre non vi è differenza con n3_w5 (+0.03%) e n3_w4(+0.63%).

Strategia	**Average Rank Value**	**prt at 85%**
n5_w5	2.5	-0.13%
n3_w5	2.0	-0.16%
n3_w4	3.17	-0.76%
n5_w4	4.5	-1.15%*
n3_w3	5.83	-2.09%*
n5_w3	6.5	-2.67%*

Tabella 6.4: Confronto tra le combinazioni di n e w sul *prt* medio calcolato su 7 anni per la strategia volume consecutive. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata

Ritorno percentuale medio di investimento per azione: la figura 6.4 mostra invece l'andamento del *rmia* medio delle combinazioni di n e w su 7 anni per la strategia volume consecutive. Si può osservare come, a parte per n3_w5, non ci siano differenze marcate negli andamenti delle altre combinazioni.

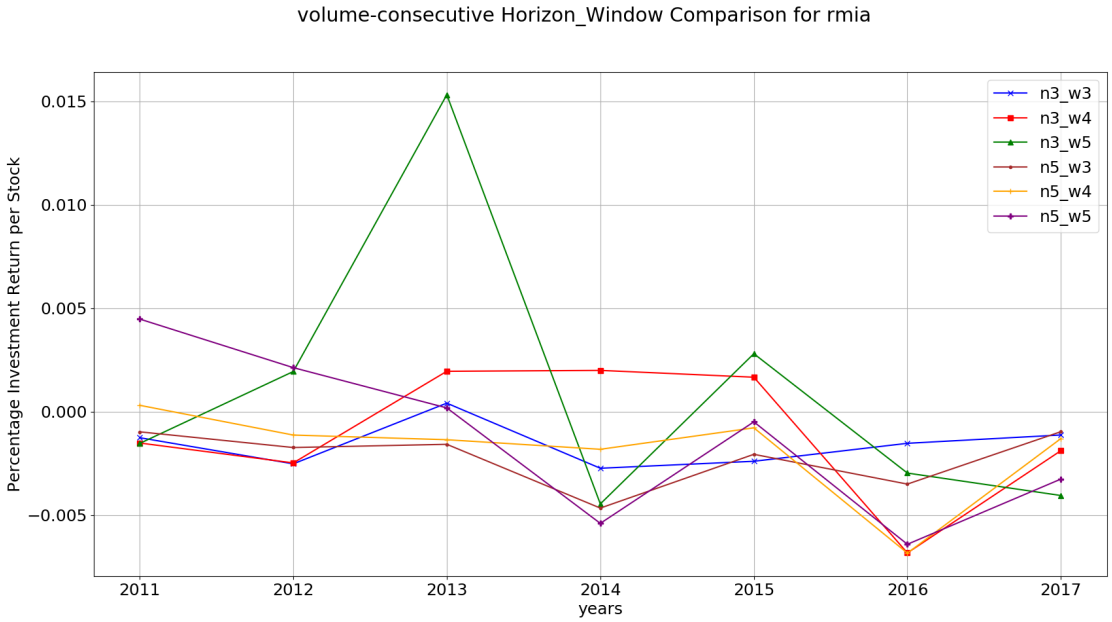

Figura 6.4: Andamento del *rmia* medio delle combinazioni di n e w su 7 anni per la strategia volume consecutive

Dalla tabella 6.5 si può notare come la combinazione n3_w5 sia la migliore e sia l'unica a raggiungere un *rmia* medio positivo. A supporto di quanto dedotto dal grafico si può notare come lo scarto tra la seconda e l'ultima combinazione sia solo di +0.12%. Questo comportamento lo si poteva già notare nella sottosezione precedente dove il range di valori del *prt* medio era molto più ampio (+4.79%) rispetto a quello del *rmia* medio (+0.25%). E' possibile inoltre notare come a parità di w le combinazioni con n=3 siano migliori rispetto a quelle con n=5: n3_w5 registra un +0.22% rispetto a n5_w5, n3_w4 registra un +0.08 rispetto a n5_w4 e infine n3_w3 registra un +0.04% rispetto a n5_w3.

Strategia	**Average Rank Value**	**rmia at 85%**
n3_w5	3.67	+0.10%
n3_w4	4.0	-0.10%
n5_w5	4.0	-0.12%
n3_w3	4.17	-0.16%
n5_w4	4.17	-0.18%
n5_w3	4.5	-0.22%

Tabella 6.5: Confronto tra le combinazioni di n e w sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive

L'applicazione di Friedman in questo caso ha dato esito negativo poiché non è stato possibile rigettare l'ipotesi nulla non riuscendo quindi a trovare una differenza significativa tra le combinazioni.

Nella tabella riassuntiva 6.6 sono riportati i valori più significativi delle combinazioni di n e w per la strategia volume consecutive mediati sui 7 anni. Si può notare come all'aumentare del valore di w diminuiscano i giorni di mercato. Questo è dovuto al fatto che è più comune trovare 3 giorni di mercato in cui la direzione di prezzo è concorde piuttosto che in 5 giorni.

Da questo secondo confronto possiamo evincere che n=3 sia migliore in media rispetto a n=5 e che w=5 sia migliore rispetto a w=3 e w=4 indicando così la combinazione volume consecutive n3 w5 come la migliore.

Strategia	Numero di Giorni di Mercato	Numero Medio di Posizioni Aperte per Giorno	Profitto Relativo Percentuale	Profitto Medio per Operazione	Investimento Medio per Operazione	Ritorno Percentuale Medio per Operazione
n3_w3	111.07 (28.2)	3.32 (0.9)	-2.09% (1543.2)	-4.16 (7.0)	3826.3 (1125.2)	-0.16%
n3_w4	48.38 (14.4)	2.04 (0.4)	-0.76% (742.2)	-4.97 (14.0)	5745.7 (977.4)	-0.10%
n3_w5	16.71 (6.4)	1.41 (0.2)	-0.16% (378.0)	10.82 (57.1)	7538.6 (776.8)	+0.10%
n5_w3	120.32 (32.4)	3.56 (1.2)	-2.67% (1338.0)	-7.06 (5.7)	3578.2 (1223.0)	-0.22%
n5_w4	59.03 (19.2)	2.03 (0.4)	-1.15% (941.4)	-10.39 (15.7)	5611.1 (1067.3)	-0.18%
n5_w5	21.27 (8.2)	1.40 (0.2)	-0.13% (291.3)	-11.59 (48.8)	7247.3 (867.5)	-0.12%

Tabella 6.6: Tabella riassuntiva dei valori medi assunti dalle combinazioni di n e w della strategia volume consecutive su 7 anni e delle relative deviazioni standard (indicate tra parentesi)

6.2.4 Confronto tra i classificatori e le features per la configurazione volume consecutive n3 w5

Questa sottosezione è organizzata nel seguente modo: partendo dalla strategia e dalle n e w definite nelle sottosezioni precedenti, per ogni misura di valutazione verranno prima confrontati i classificatori e in seguito le features. I valori che verranno rappresentati nei grafici e nelle tabelle sono da considerarsi per i classificatori come i valori medi di *prt* / *rmia* calcolati su 9 configurazioni ($9 \cdot 7$ anni nel caso delle tabelle, dato che rimangono solo le features come parametro variabile della configurazione) e per le features come i valori medi calcolati su 6 configurazioni ($6 \cdot 7$ anni nel caso delle tabelle, dato che rimangono solo i classificatori come parametro variabile della configurazione).

Infine tramite il confronto con una tabella che combina, per *prt* e *rmia*, le features con i classificatori si osserva se le ipotesi fatte in precedenza concordano con la classifica stilata dalla tabella stessa.

Per questioni di spazio e semplicità le features verranno indicate tramite le loro abbreviazioni: ALL (A), NEWS+TEMP (NT), NEWS+VOL (NV), OSC+TEMP (OT), OSC+TEMP+NEWS (OTN), OSC+VOL (OV), OSC+VOL+NEWS (OVN), TEMP (T), VOL (V).

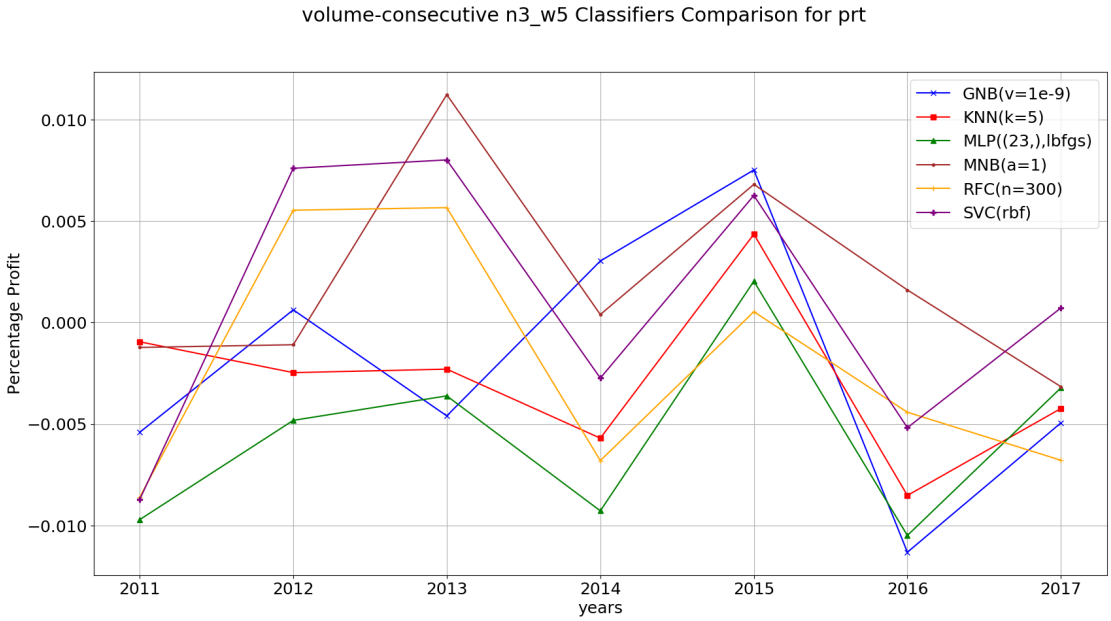

Figura 6.5: Andamento del *prt* medio dei classificatori su 7 anni per configurazione volume consecutive n3 w5

Profitto relativo totale percentuale: la figura 6.5 mostra l'andamento del *prt* medio per ciascuno dei classificatori della configurazione volume consecutive n3 w5 su 7 anni. Osservando il grafico si può notare come i classificatori MNB e SVC sembrino avere gli andamenti migliori dovuti al maggior numero di anni chiusi in positivo (4) e al fatto che le perdite sembrino essere contenute mentre si possono identificare in MLP e KNN gli andamenti peggiori (un solo anno chiuso in positivo, 2015, per entrambi).

Nella tabella 6.7 si può notare come il classificatore MNB si classifichi al primo posto sia nel ranking medio e sia per *prt* con un valore medio positivo, seguito dal classificatore SVC, positivo anch'esso.

Il test di Friedman specifica che l'uso dei classificatori è significativamente diverso e tramite Wilcoxon si riesce a capire che l'uso del primo classificato è significativamente diverso dal RFC (+0.42%), dal KNN (+0.49%) e dal MLP (+0.77%) ma non è diverso dal SVC (+0.13%) e dal GNB (+0.42%).

Strategia	**Average Rank Value**	**prt at 85%**
MNB	2.33	+0.21%
SVC	3.0	+0.08%
GNB	4.17	-0.21%
RFC	4.67	-0.21%*
KNN	4.33	-0.28%*
MLP	6.0	-0.56%*

Tabella 6.7: Confronto tra i classificatori sul *prt* medio calcolato su 7 anni per la strategia volume consecutive n3 w5. L'asterisco indica una significativa differenza, misurata con Wilcoxon, della configurazione considerata con la prima classificata

La figura 6.6 invece mostra l'andamento del *prt* medio per ciascuna delle features della configurazione volume consecutive n3 w5 su 7 anni. Osservando questo grafico si può notare come le differenze tra gli andamenti delle features sembrino non essere marcate. Non c'è mai un andamento che risulti essere migliore rispetto agli altri su tutti e 7 gli anni. Si può però notare come OSC+VOL raggiunga i valori migliori negli anni 2013 e 2015. L'andamento peggiore risulta invece essere NEWS+VOL (1 solo valore positivo).

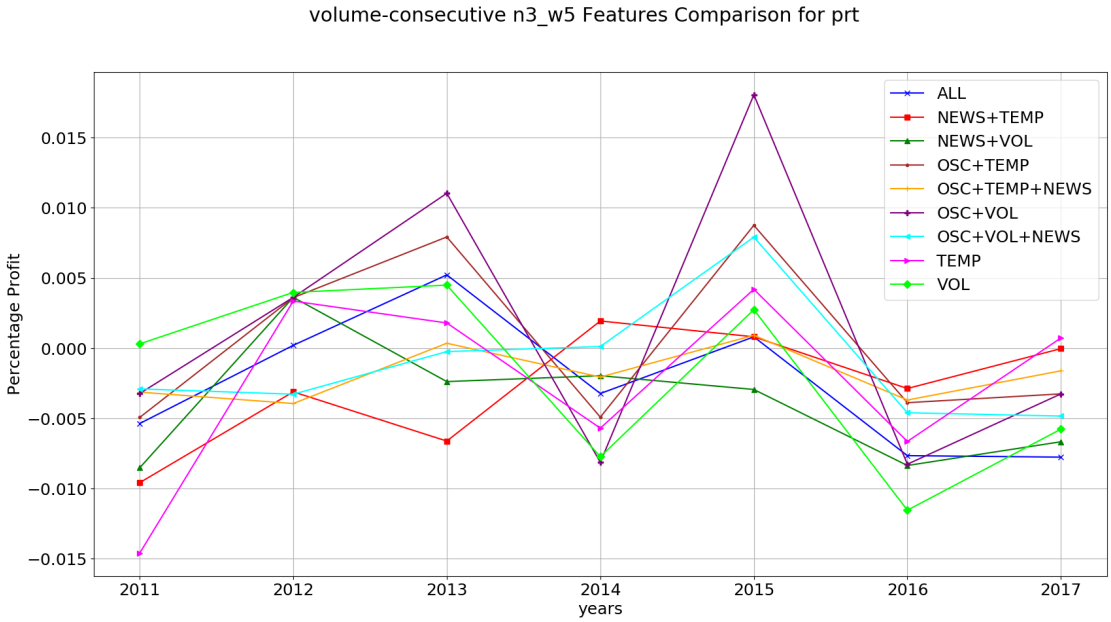

Figura 6.6: Andamento del *prt* medio delle features su 7 anni per la configurazione volume consecutive n3 w5

Dalla tabella 6.8 si può notare come i risultati siano concordi con l'affermazione fatta in precedenza. Le migliori features, le uniche con *prt* positivo, risultano infatti essere OSC+VOL (+0.14%) seguito da OSC+TEMP (+0.05%). Friedman inoltre non ha dato esito positivo sottolineando dunque come non sia stato possibile trovare significative differenze tra le features.

Strategia	Average Rank Value	prt at 85%
OSC+VOL	3.33	+0.14%
OSC+TEMP	2.89	+0.05%
OSC+VOL+NEWS	3.67	-0.11%
OSC+TEMP+NEWS	3.56	-0.19%
VOL	3.89	-0.19%
TEMP	4.0	-0.24%
ALL	4.78	-0.25%
NEWS+TEMP	3.89	-0.28%
NEWS+VOL	5.0	-0.39%

Tabella 6.8: Confronto tra la features sul *prt* medio calcolato su 7 anni per la strategia volume consecutive n3 w5

Ritorno percentuale medio di investimento per azione: la figura 6.7 mostra l'andamento del *rmia* medio per ciascuno dei classificatori della configurazione volume consecutive n3 w5 su 7 anni. Osservando il grafico si può notare come i classificatori SVC e MNB presentino l'andamento migliore (4 anni in positivo per SVC e 3 per MNB) mentre MLP e KNN quello peggiore (per entrambi un solo anno chiuso in positivo).

Dalla tabella 6.9 si può evincere come le osservazioni fatte sul grafico trovino qui un riscontro, indicando SVC (+0.88%) e MNB (+0.41%) come i classificatori migliori. Si nota anche una corrispondenza tra i classificatori migliori per *prt* con quelli per *rmia*. Nonostante questo Friedman non ha dato esito positivo indicando quindi che non vi è una significativa differenza tra i classificatori.

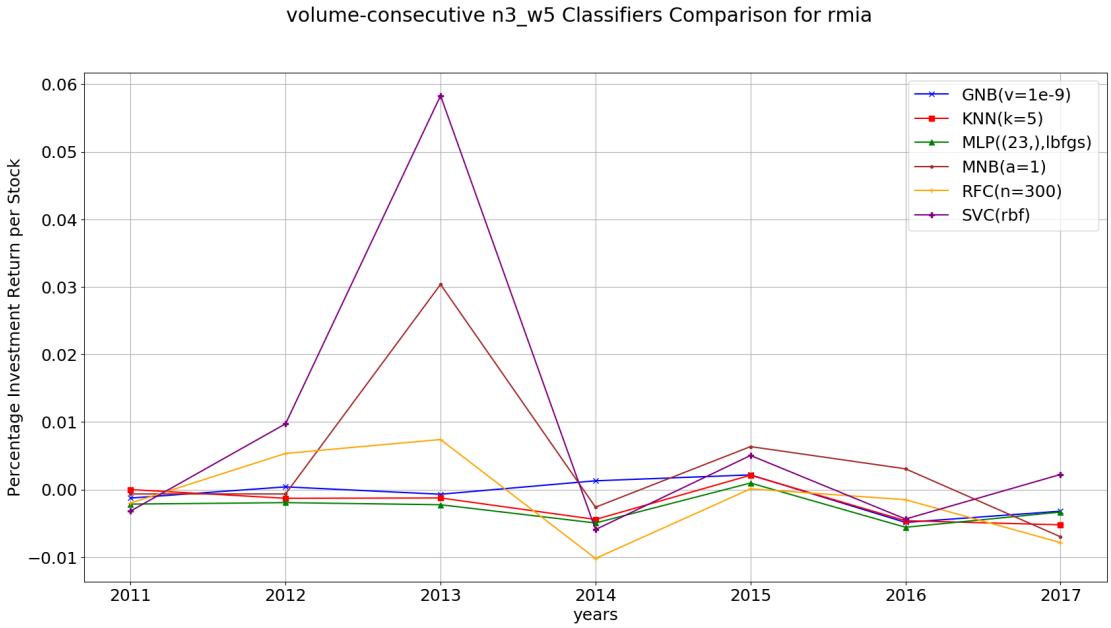

Figura 6.7: Andamento del *rmia* medio dei i classificatori su 7 anni per la configurazione volume consecutive n3 w5

Strategia	**Average Rank Value**	**rmia at 85%**
SVC	3.17	+0.88%
MNB	2.83	+0.41%
GNB	3.5	-0.09%
RFC	4.83	-0.12%
KNN	4.33	-0.21%
MLP	5.83	-0.27%

Tabella 6.9: Confronto tra i classificatori sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive n3 w5

La figura 6.8 invece mostra l'andamento del *rmia* medio per ciascuna delle features della configurazione volume consecutive n3 w5 su 7 anni. Anche per questo grafico si può osservare che per la maggior parte gli andamenti non vi siano significative differenze. Fa eccezione l'anno 2013 dove si nota un *rmia* molto alto per le features basate sugli oscillatori.

Dalla tabella 6.10 si può notare come le features basate sugli oscillatori

abbiano raggiunto il *rmia* più alto. Anche in questo caso Friedman ha dato esito negativo.

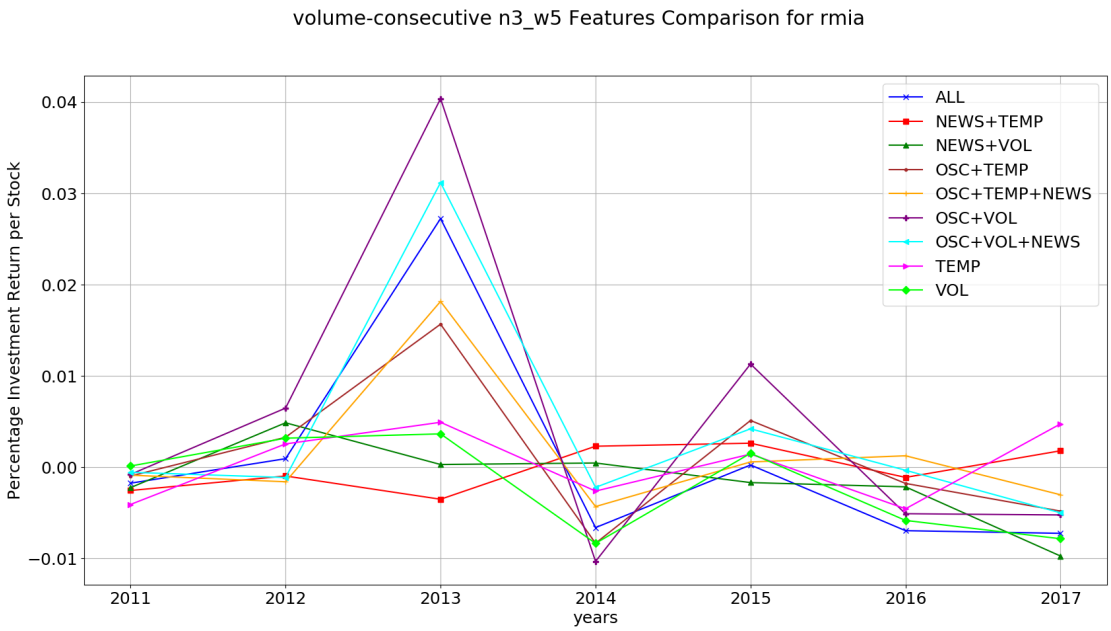

Figura 6.8: Andamento del *rmia* medio delle features su 7 anni per la configurazione volume consecutive n3 w5

Strategia	**Average Rank Value**	**rmia at 85%**
OSC+VOL	3.11	+0.52%
OSC+VOL+NEWS	2.78	+0.37%
OSC+TEMP+NEWS	3.67	+0.14%
OSC+TEMP	3.33	+0.11%
ALL	5.0	+0.08%
TEMP	4.11	+0.03%
NEWS+TEMP	3.78	-0.02%
NEWS+VOL	4.67	-0.15%
VOL	4.56	-0.19%

Tabella 6.10: Confronto tra le features sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive n3 w5

Dai confronti precedenti i migliori classificatori e le migliori features sono risultati essere: per *prt*

- classificatori: MNB (+0.21%), SVC (+0.08%)
- features: OSC+VOL (+0.14%), OSC+TEMP (+0.05%)

mentre per *rmia*

- classificatori: SVC (+0.88%), MNB (+0.41%)
- features: OSC+VOL (+0.52%), OSC+VOL+NEWS (+0.37%), OSC+TEMP+NEWS (+0.14%), OSC+TEMP (+0.11%), ALL (+0.08%), TEMP (+0.03%)

Per avere una visione più generale su quale siano i classificatori e le features migliori, sono stati riportati in seguito i grafici che confrontano gli andamenti medi generali dei classificatori e delle features su 7 anni. Ciascun valore (per anno) rappresenta quindi il *prt* / *rmia* medio calcolato come la media dei *prt* / *rmia* di tutte le possibili configurazioni con $classificatore$ / $feature$ + *anno* fissati, per un totale di 180 possibili combinazioni per i classificatori (i parametri variabili sono features + window + horizon + strategie) e 120 combinazioni per le features (i parametri variabili sono classificatori + window + horizon + strategie).

Le figure 6.9 e 6.10 mostrano l'andamento del *prt* e del *rmia* medio per i classificatori. Si può notare come MNB e SVC siano i migliori evidenziando una correlazione con quanto ottenuto in precedenza per la configurazione volume consecutive n3 w5.

Le figure 6.11 e 6.12 mostrano invece l'andamento del *prt* e del *rmia* medio per le features. Si può notare come, a differenza dei classificatori, non ci siano differenze marcate tra gli andamenti. Si può dunque dedurre che la scelta di una feature per una determinata configurazione dipenda dalla configurazione stessa non essendoci in generale una feature migliore.

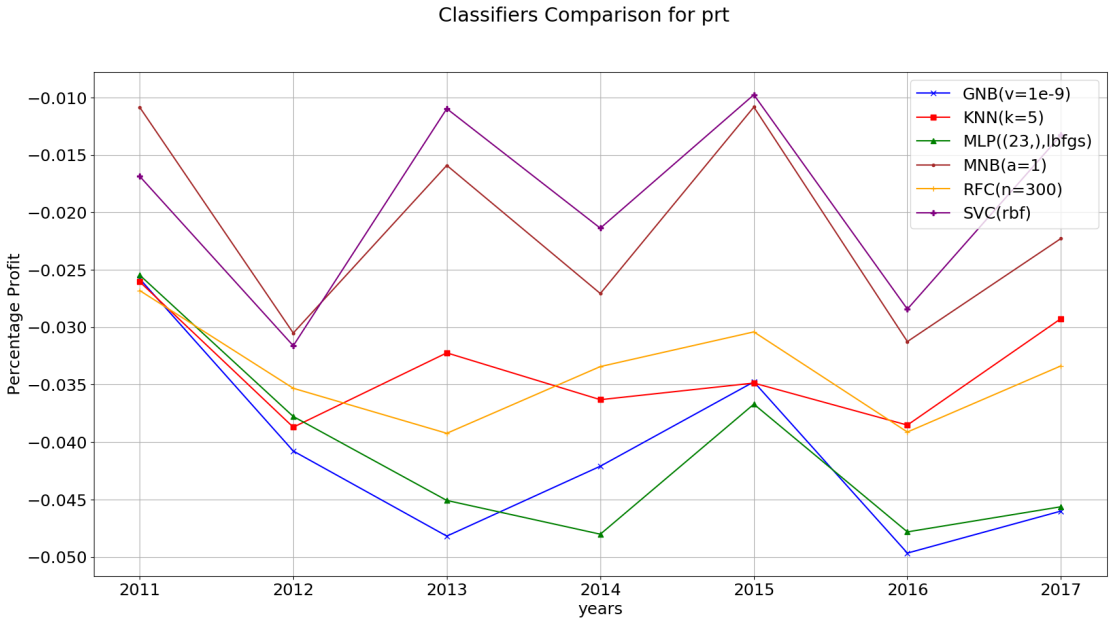

Figura 6.9: Andamento del *prt* medio generale dei classificatori su 7 anni

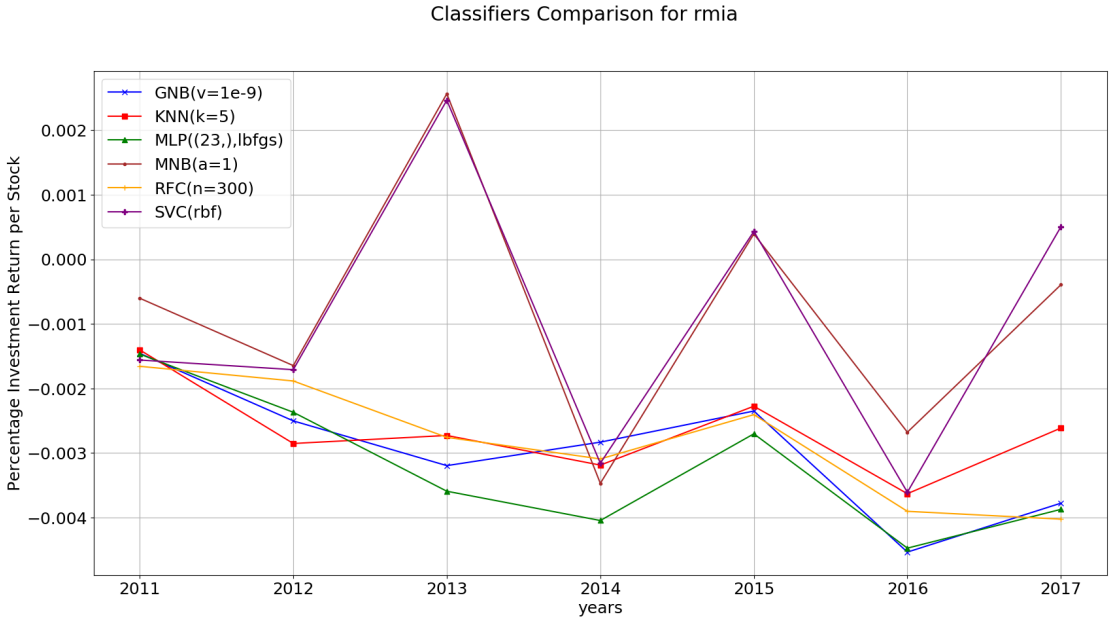

Figura 6.10: Andamento del *rmia* medio generale dei classificatori su 7 anni

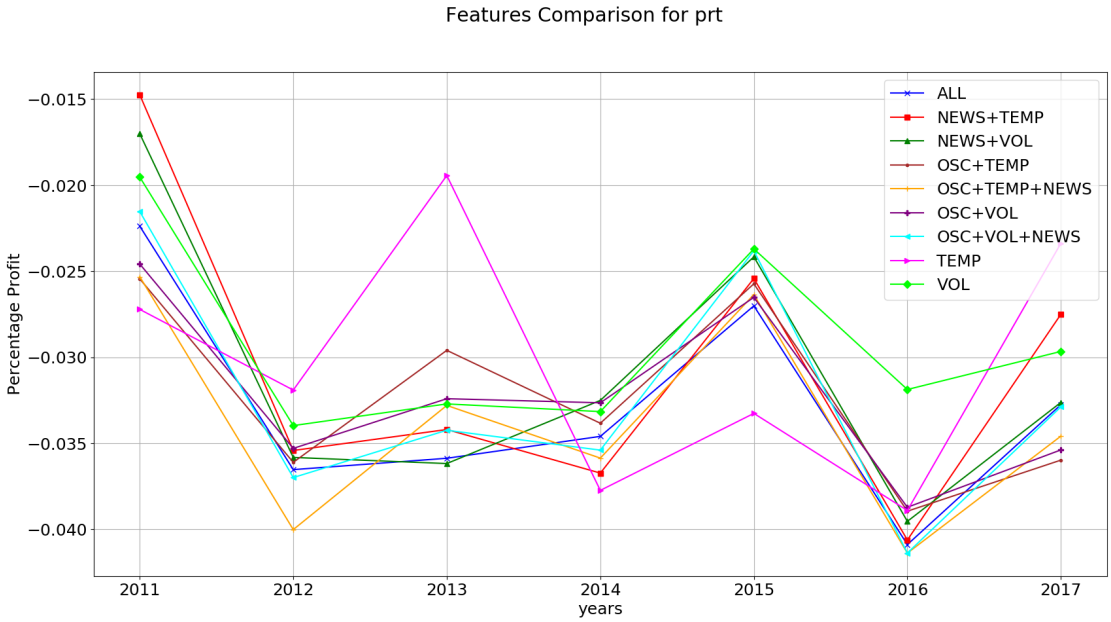

Figura 6.11: Andamento del *prt* medio generale delle features su 7 anni

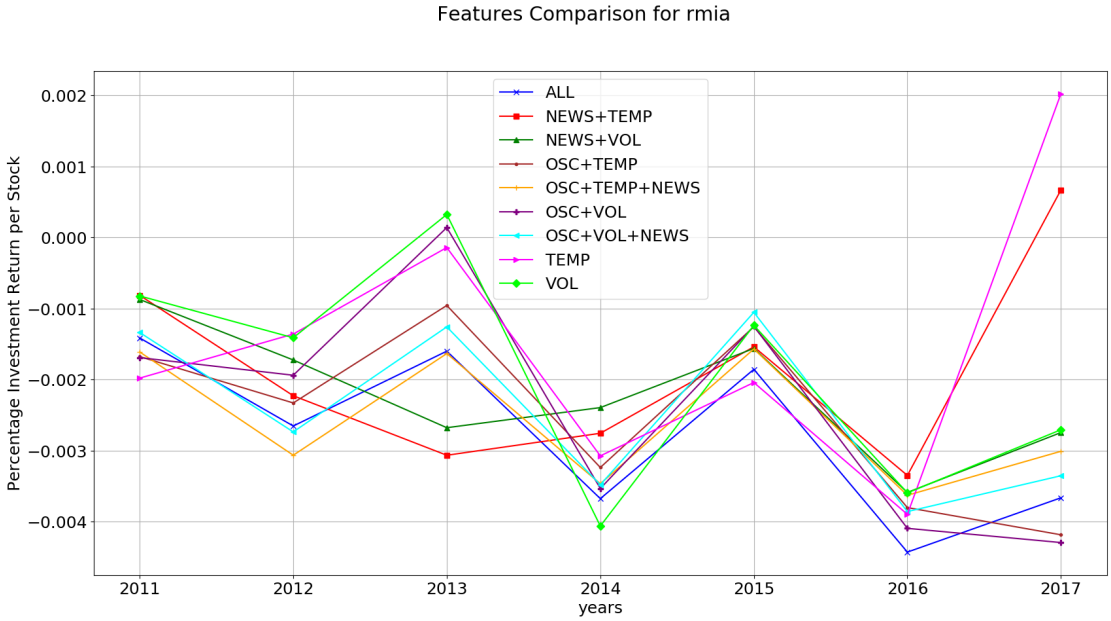

Figura 6.12: Andamento del *rmia* medio generale delle features su 7 anni

A questo punto si procede nell'analizzare quale sia il miglior classificatore e la migliore feature per la configurazione volume consecutive n3 w5. Nel

fare ciò si utilizzano le tabelle 6.11 e 6.12 che riportano i valori medi (su 7 anni) rispettivamente di *prt* e *rmia* delle combinazioni dei classificatori e delle features. Per semplicità sono stati riportati solo i primi 10 valori in ordine decrescente (su 54 combinazioni).

Dalla tabella 6.11 si può dedurre come il classificatore migliore sia MNB seguito da SVC e come le features migliori siano quelle relative agli oscillatori esattamente come dedotto in precedenza dallo studio delle figure 6.5, 6.6 e dalle tabelle 6.7 e 6.8.

Dalla tabella 6.12 invece si può dedurre come il classificatore migliore sia SVC seguito da MNB e come le features migliori anche qui siano quelle relative agli oscillatori, risultato concorde con quello ottenuto dall'analisi delle figure 6.7, 6.8 e dalle tabelle 6.9 e 6.10.

Da queste osservazioni possiamo indicare, per la configurazione volume consecutive n3 w5, OSC+VOL come la feature migliore e, in ordine, SVC e MNB come i classificatori migliori.

Strategia	**Average Rank Value**	**prt at 85%**
MNB_OSC+VOL	2.03	+0.65%
MNB_OSC+TEMP	2.04	+0.41%
SVC_OSC+TEMP	2.08	+0.39%
SVC_NEWS+TEMP	1.85	+0.39%
MNB_OSC+VOL+NEWS	2.33	+0.33%
SVC_OSC+VOL	2.29	+0.32%
MNB_NEWS+TEMP	2.3	+0.32%
MNB_NEWS+VOL	2.83	+0.26%
KNN_OSC+TEMP	2.5	+0.26%
MNB_OSC+TEMP+NEWS	2.57	+0.24%

Tabella 6.11: Confronto tra le combinazioni di features e classificatori sul *prt* medio calcolato su 7 anni per la strategia volume consecutive n3 w5

Strategia	**Average Rank Value**	**rmia at 85%**
SVC_OSC+VOL	2.75	+2.54%
SVC_OSC+VOL+NEWS	2.81	+2.28%
SVC_OSC+TEMP+NEWS	2.52	+1.14%
SVC_ALL	3.55	+0.95%
MNB_OSC+VOL	2.19	+0.90%
MNB_OSC+VOL+NEWS	2.46	+0.66%
MNB_ALL	4.63	+0.65%
MNB_OSC+TEMP	2.31	+0.62%
SVC_NEWS+TEMP	1.8	+0.62%
MNB_OSC+TEMP+NEWS	2.96	+0.55%

Tabella 6.12: Confronto tra le combinazioni di features e classificatori sul *rmia* medio calcolato su 7 anni per la strategia volume consecutive n3 w5

L'andamento del *prt* e del *rmia* delle configurazioni volume consecutive n3 w5 SVC OSC+VOL e volume consecutive n3 w5 MNB OSC+VOL può essere visualizzato nelle figure 6.13 e 6.14 mentre nelle tabelle riassuntive 6.13 e 6.14 sono riportati i valori medi (su 7 anni) più significativi dei classificatori e delle features per la configurazione volume consecutive n3 w5.

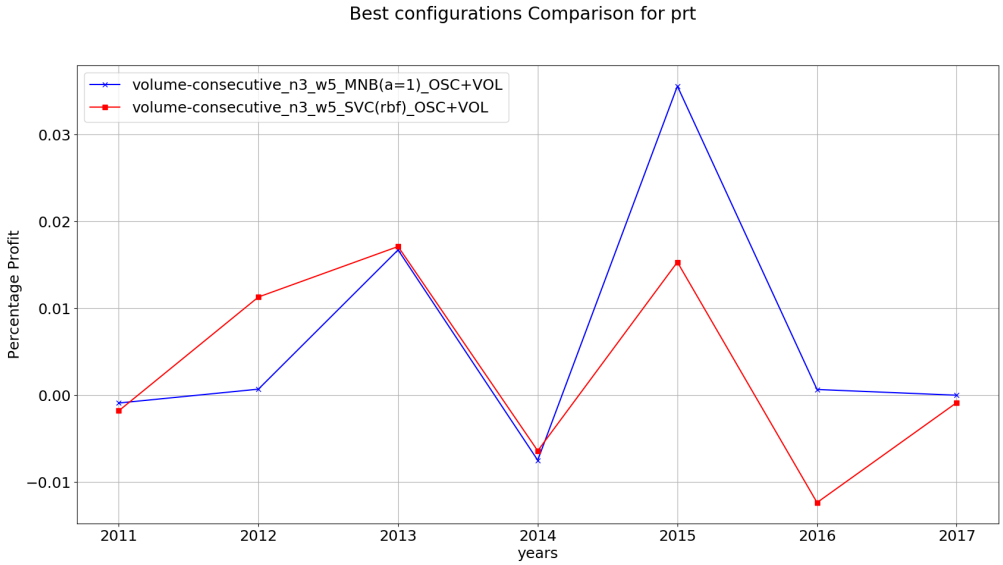

Figura 6.13: Andamento del *prt* su 7 anni delle configurazioni volume consecutive n3 w5 SVC OSC+VOL e volume consecutive n3 w5 MNB OSC+VOL

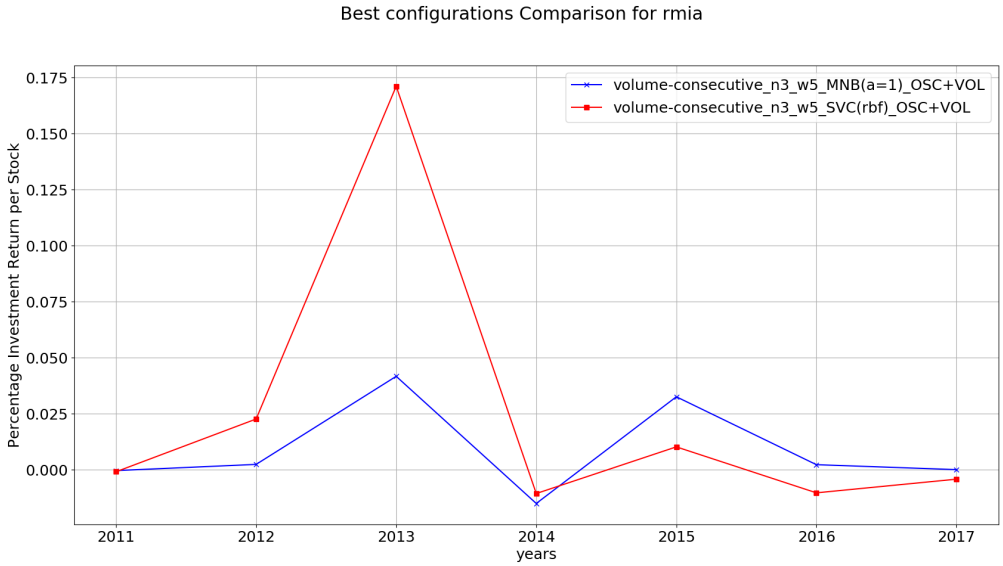

Figura 6.14: Andamento del *rmia* su 7 anni delle configurazioni volume consecutive n3 w5 SVC OSC+VOL e volume consecutive n3 w5 MNB OSC+VOL

Strategia	Numero di Giorni di Mercato	Numero Medio di Posizioni Aperte per Giorno	Profitto Relativo Percentuale	Profitto Medio per Operazione	Investimento Medio per Operazione	Ritorno Percentuale Medio per Operazione
GNB	24.08 (2.38)	1.60 (0.04)	-0.21% (198.7)	-7.64 (11.8)	6844.7 (180.0)	-0.09%
KNN	18,24 (2.03)	1.63 (0.06)	-0.28% (374.3)	-19.41 (22.3)	6995.2 (250.1)	-0.21%
MLP	22.48 (2.89)	1.52 (0.08)	-0.56% (289.5)	-22.19 (11.1)	7132.0 (451.7)	-0.27%
MNB	8.73 (2.51)	1.12 (0.16)	+0.21% (255.4)	37.86 (36.1)	7980.7 (786.5)	+0.41%
RFC	18.86 (4.04)	1.39 (0.12)	-0.21% (244.3)	-11.47 (11.9)	7890.8 (573.7)	-0.12%
SVC	9.90 (1.63)	1.20 (0.07)	+0.08% (305.8)	87.76 (88.8)	8388.3 (624.7)	+0.88%

Tabella 6.13: Tabella riassuntiva dei valori medi assunti dai classificatori della configurazione volume consecutive n3 w5 su 7 anni e delle relative deviazioni standard (indicate tra parentesi)

Strategia	Numero di Giorni di Mercato	Numero Medio di Posizioni Aperte per Giorno	Profitto Relativo Percentuale	Profitto Medio per Operazione	Investimento Medio per Operazione	Ritorno Percentuale Medio per Operazione
A	16 (5.42)	1.43 (0.14)	-0.25% (338.6)	12.78 (55.67)	7601.2 (596.4)	+0.08%
NT	18.95 (6.63)	1.45 (0.20)	-0.28% (527.7)	0.09 (31.42)	7241.6 (699.5)	-0.02%
NV	18.31 (6.42)	1.45 (0.21)	-0.39% (356.5)	-11.59 (25.26)	7202.2 (728.6)	-0.15%
OT	15.05 (5.58)	1.40 (0.16)	+0.05% (393.8)	9.24 (24.93)	7711.5 (774.3)	+0.11%
OTN	16.81 (6.57)	1.37 (0.20)	-0.19% (345.8)	16.29 (51.20)	8078.3 (861.8)	+0.14%
OV	13.78 (4.98)	1.33 (0.19)	+0.14% (279.9)	46.44 (95.17)	7782.2 (737.0)	+0.52%
OVN	16.36 (5.48)	1.40 (0.18)	-0.11% (307.3)	36.99 (91.01)	7823.4 (814.6)	+0.37%
T	18.52 (6.79)	1.44 (0.21)	-0.24% (251.0)	5.62 (18.01)	7209.4 (409.5)	+0.03%
V	16.64 (7.23)	1.41 (0.35)	-0.19% (190.8)	-18.48 (15.74)	7197.8 (668.2)	-0.19%

Tabella 6.14: Tabella riassuntiva dei valori medi assunti dalle features della configurazione volume consecutive n3 w5 su 7 anni e delle relative deviazioni standard (indicate tra parentesi)

Capitolo 7

Conclusioni e Lavori Futuri

I mercati finanziari consentono la compravendita (trading) di strumenti finanziari. Esistono diversi tipi di mercati (bond, forex, derivati, ...), ma il più noto è probabilmente quello azionario in cui si scambiano quote di società quotate.

Con l'avvento della digitalizzazione il trading online ha assunto sempre più importanza diventando accessibile per gli utenti di tutto il mondo e offrendo loro la possibilità di fare compravendita di strumenti finanziari tramite il web.

Grazie alla disponibilità di grandi moli di dati storici relativi ai mercati finanziari, una promettente direzione di ricerca è l'utilizzo di tecniche di analisi dei dati per definire le regole con cui i sistemi di trading investono sul mercato azionario in modo automatico. I processi predittivi si basano sull'analisi di dati storici mediante algoritmi di Machine Learning. Essi includono una vasta scelta di algoritmi di classificazione e regressione, finalizzati a predire il valore di una variabile target.

In questa tesi è stato proposto un sistema di trading quantitativo che investe sul mercato azionario mediante una strategia multiday di tipo trend reversal basata sull'utilizzo di tecniche di machine learning quali algoritmi di classificazione per predire la direzione di un'azione nei giorni successivi.

L'obiettivo è stato quello di dimostrare l'efficacia di un approccio innovativo basato su Machine Learning rispetto ad un approccio tradizionale

basato esclusivamente sull' analisi tecnica e definire le configurazioni più appropriate del sistema analizzando l'impatto di vari fattori sui risultati della simulazione trading.

E' stata condotta una campagna sperimentale su un campione di 7 anni, dal 2011 al 2017, di azioni dell'indice azionario americano Standard&Poor 500.

I risultati sperimentali hanno evidenziato che:

- la strategia *consecutive* registra ritorni migliori in termini sia di profitto relativo che di profitto medio per operazione rispetto alle strategie basate sull'incrocio di medie mobili quali *SMA* e *MACD*.
- l'applicazione di un filtro sul volume scambiato per azione sembra migliorare le prestazioni e i ritorni di ogni strategia rendendo così la strategia *volume consecutive* la migliore.
- la finestra di predizione (N) predetta dal classificatore ha registrato andamenti migliori su 3 giorni rispetto a 5.
- nel caso della strategia *volume consecutive* si è potuto notare come sia il profitto relativo che il profitto medio per operazione aumentino all'aumentare del valore di W che, ricordiamo, rappresentare la grandezza della sliding window, ovvero una finestra di W giorni nei quali la direzione del prezzo dell'azione deve essere concorde.
- i classificatori che hanno registrato significativamente il miglior andamento nelle simulazioni sono *SVC* e *MNB*.
- per quanto riguarda le features invece non ce n'è mai stata una che abbia avuto un andamento significativamente migliore rispetto alle altre. Per la strategia migliore, la *volume consecutive*, le feature che si sono comportate meglio sono quelle basate su *VOL*, in particolare *VOL, OSC+VOL e NEWS+VOL*). Mentre, per la configurazione migliore *volume consecutive n3 w5* le features basate sugli oscillatori tecnici si sono rivelate le più efficaci, in particolare *OSC+VOL* sottolineando una corrispondenza con quanto detto in precedenza.

Al di là di queste osservazioni abbiamo notato che i profitti relativi e i profitti medi per operazione ottenuti non sono elevati. Questo può essere dovuto al

ristretto numero di operazioni svolte in un intero anno di mercato.

Il lavoro svolto in questa tesi, sebbene sia stato valutato in molti scenari, può essere ampliato sotto diversi aspetti:

- Provare la portabilità del sistema su altri indici azionari (Dow Jones, ...) o altri tipi di mercato (cryptovalute, Forex, ecc.).
- Utilizzare algoritmi di deep learning (come LSTM) o costruire modelli di predizione più complessi.
- Utilizzare altre strategie ensemble al posto del simple majority voting.
- Modificare la strategia di trading (diverse stop loss, diverse distribuzioni di investimento, ...).
- Introdurre l'utilizzo del machine learning anche nel modulo di trading andando a classificare gli anni passati, o porzioni di essi, in modo tale da saper riconoscere se una situazione di mercato si è già verificata in precedenza e quindi agire di conseguenza.

Bibliografia

[1] A. Saunders e M. M. Cornett. *Financial Markets and Institutions*. 5ª ed. McGraw-Hill/Irwin, 2012 (cit. alle pp. 19, 22).

[2] F. S. Mishkin e S. G. Eakins. *Financial Markets and Institutions*. 7ª ed. Prentice Hall, 2012 (cit. a p. 19).

[3] *Trade.com*. `https://www.trade.com/` (cit. a p. 21).

[4] *Plus500*. `https://www.plus500.it/` (cit. a p. 21).

[5] *eToro*. `https://www.etoro.com/` (cit. a p. 21).

[6] J. Murphy. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York Institute of Finance Series. New York Institute of Finance, 1999 (cit. alle pp. 21, 47).

[7] I. H. Witten e E. Frank. *Data Mining, Pratical Machine Learning Tools and Techniques*. 2ª ed. 500 Sansome Street, Suite 400, San Francisco, CA 94111: Morgan Kaufmann, 2005 (cit. alle pp. 24, 26, 33).

[8] P. N. Tan, M. Steinbach e V. Kumar. *Introduction to Data Mining*. 2ª ed. Pearson, 2005 (cit. alle pp. 24, 26–32, 34, 35).

[9] *Classifying data with decision trees*. `https://elf11.github.io/2018/07/01/python-decision-trees-acm.html`. Lug. 2018 (cit. a p. 33).

[10] *A hands-on tutorial on the Perceptron learning algorithm*. `http://abhay.harpale.net/blog/machine-learning/a-hands-on-tutorial-on-the-perceptron-learning-algorithm/` (cit. a p. 34).

[11] D. Ballabio, R. Todeschini e V. Consonni. «Chapter 5 - Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data». In: *Data Handling in Science and Technology* 31 (2019), pp. 129–155 (cit. a p. 36).

[12] L. S. Malagrino, N. T. Roman e A. M. Monteiro. «Forecasting stock market index daily direction: A Bayesian Network approach». In: *Expert Systems With Applications* 105 (2018), pp. 11–22 (cit. alle pp. 37, 39, 41).

[13] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan e K. P. Soman V. K. Menon. «Stock price prediction using lstm, RNN and cnn-sliding window model». In: *ICACCI. IEEE* (2017), pp. 1643–1647 (cit. alle pp. 37, 41).

[14] X. Chenand e Z. He. «Prediction of Stock Trading Signal Based on Support Vector Machine». In: *8th International Conference on Intelligent Computation Technology and Automation* (2015) (cit. alle pp. 37, 39, 41).

[15] L. A. Teixeira e A. L. I. Oliveira. «Predicting stock trends through technical analysis and nearest neighbor classification». In: *2009 IEEE International Conference on Systems, Man and Cybernetics* (2009) (cit. alle pp. 37, 39, 41).

[16] K. Alkhatib, N. Hassan, I. Hmeidi e K. A. M. Shatnawi. «Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm». In: *International Journal of Business, Humanities and Technology* 3 (2013) (cit. alle pp. 37, 39, 41).

[17] T. Manojlović e I. Štajduhar. «Predicting stock market trends using random forests: A sample of the Zagreb stock exchange». In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (2015), pp. 1189–1193 (cit. alle pp. 37, 39, 41).

[18] S. Basak, S. Kar, S. Saha, L. Khaidem e S. R. Dey. «Predicting the direction of stock market prices using tree-based classifiers». In: *The North American Journal of Economics and Finance* 47 (2019), pp. 552–567 (cit. alle pp. 37, 39, 41).

[19] W. Li e J. Liao. «A comparative study on trend forecasting approach for stock price time series». In: *11th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID)* (2017), pp. 74–78 (cit. alle pp. 37, 41).

[20] S. Pryima, R. Vovk e V. Vovk. «Using Artificial Neural Networks to Forecast Stock Market Indices». In: *XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT)* (2019), pp. 108–112 (cit. alle pp. 37, 39, 41).

[21] J. Patel, S. Shah, P. Thakkar e K. Kotecha. «Predicting stock market index using fusion of machine learning techniques». In: *Expert Systems with Applications* 42 (2015), pp. 2162–2172 (cit. alle pp. 37, 41).

[22] C.-F. Tsai, Y.-C. Lin, D. C. Yen e Y.-M. Chen. «Predicting stock returns by classifier ensembles». In: *Applied Soft Computing* 1 (2011), pp. 2452–2459 (cit. alle pp. 37, 41).

[23] Y. Chen e Y. Hao. «A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction». In: *Expert Systems with Applications* 80 (2017), pp. 340–355 (cit. alle pp. 37, 41).

[24] L. Cagliero, P. Garza, G. Attanasio e E. Baralis. «Training ensembles of faceted classification models for quantitative stock trading». In: *Computing* 102 (2020), pp. 1213–1225 (cit. alle pp. 38–40).

[25] H. Talebi, W. Hoang e M. L. Gavrilova. «Multi-Scale Foreign Exchange Rates Ensemble for Classification of Trends in Forex Market». In: *Procedia Computer Science* 29 (2014), pp. 2065–2075 (cit. alle pp. 38, 40).

[26] A. Adegboye, M. Kampouridis e C. G.Johnson. «Regression genetic programming for estimating trend end in foreign exchange market». In: *IEEE Symposium Series on Computational Intelligence (SSCI)* (2017), pp. 1–8 (cit. alle pp. 38, 40).

[27] Z. Wang, S.-B. Ho e Z. Lin. «Stock market prediction analysis by incorporating social and news opinion and sentiment». In: *IEEE International Conference on Data Mining Workshops* (2018), pp. 1375–1380 (cit. alle pp. 38, 39, 41).

[28] A. Picasso, S. Merello, Y. Ma, L. Oneto e E. Cambria. «Technical analysis and sentiment embeddings for market trend prediction». In: *Expert Systems with Applications* 135 (2019), pp. 60–70 (cit. alle pp. 38, 39, 41).

[29] D. Heyman, M. Lescrauwaet e H. Stieperaere. «Investor attention and short-term return reversals». In: *Finance Research Letters* 29 (2019), pp. 1–6 (cit. alle pp. 38, 40).

[30] N. T. Son, L. V. Thanh, T. Q. Ban, D. X. Hoa e B. N. Anh. «An Analyze on Effectivness of Candlestick Reversal Patterns for Vietnamese Stock Market». In: *Proceedings of the 2018 International Conference on Information Management Management Science. ACM* (2018), pp. 89–93 (cit. alle pp. 38, 40).

[31] K. Lee e G. Jo. «Expert system for predicting stock market timing using a candlestick chart». In: *Expert Systems with Applications* 16.4 (1999), pp. 357–364 (cit. alle pp. 38, 40).

[32] Q. Lan, D. Zhang e L. Xiong. «Reversal Pattern Discovery in Financial Time Series Based on Fuzzy Candlestick Lines». In: *Systems Engineering Procedia* 2 (2011), pp. 182–190 (cit. alle pp. 38, 40).

[33] X. Chu, Z. Gu e H. Zhou. «Intraday momentum and reversal in chinese stock market». In: *Finance Research Letters* 30 (2019), pp. 83–88 (cit. alle pp. 38, 40).

[34] R. J. Balvers e Y. Wu. «Momentum and mean reversion across national equity markets». In: *Journal of Empirical Finance* 13 (2006), pp. 24–48 (cit. alle pp. 38, 40).

[35] T.L. Chen e F.-Y. Chen. «An intelligent pattern recognition model for supporting investment decisions in stock market». In: *Information Sciences* 346-347 (2016), pp. 261–274 (cit. alle pp. 38, 40).

[36] E. Papacostantis e A. P. Engelbrecht. «Coevolutionary Particle Swarm Optimization for Evolving Trend Reversal Indicators». In: *IEEE Symposium on Computational Intelligence for Financial Engineering and Economics* (2011) (cit. alle pp. 38, 40).

[37] I. O. Kravets, V. O. Kozlovskaya e V. S. Tumko. «The forecastings of future changes in the trend's direction of stock quotes by neural networks and fuzzy systems». In: *IEEE First International Conference on Data Stream Mining Processing* (2016) (cit. alle pp. 38, 40).

[38] J. U, P. Lu, C. Kim, U. Ryu e K. Pak. «A new LSTM based reversal point prediction method using upward/downward reversal point feature sets». In: *Chaos, Solitons Fractals* 132 (2020), pp. 1–8 (cit. alle pp. 38, 40).

[39] R. K. Nayak, D. Mishra e A. K. Rath. «A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices». In: *Applied Soft Computing* 35 (2015), pp. 670–680 (cit. alle pp. 39, 40).

[40] S. Zhao, Y. Tong, X. Meng, X. Yang e S. Tan. «Predicting return reversal through a two-stage method». In: *7th IEEE International Conference on Software Engineering and Service Science* (2016) (cit. alle pp. 39, 40).

[41] L. Cagliero, G. Attanasio, P. Garza e E. Baralis. «Combining news sentiment and technical analysis to predict stock trend reversal». In: *2019 International Conference on Data Mining Workshops* (2019) (cit. alle pp. 39, 40, 77).

[42] *Medie mobili: cosa sono e come utilizzarle per il trading*. `https://www.borsaitaliana.it/notizie/sotto-la-lente/analisi-tecnica-medie-mobili-trading.htm`. Giu. 2018 (cit. alle pp. 47, 62).

[43] *Yahoo Finance*. `https://finance.yahoo.com/` (cit. a p. 78).

[44] *Reuters*. `https://www.reuters.com/` (cit. a p. 78).

[45] *Scikit-learn*. `https://scikit-learn.org/` (cit. a p. 78).

[46] Denise Rey e Markus Neuhäuser. «Wilcoxon-Signed-Rank Test». In: *International Encyclopedia of Statistical Science*. A cura di Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1658–1659. ISBN: 978-3-642-04898-2. DOI: `10.1007/978-3-642-04898-2_616`. URL: `https://doi.org/10.1007/978-3-642-04898-2_616` (cit. a p. 80).

[47] *Wilcoxon Test*. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html` (cit. a p. 80).