Università
Ca'Foscari
Venezia

Corso di Laurea magistrale
(*ordinamento ex D.M. 270/2004*)
in Economia e Diritto

Tesi di Laurea

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Bilancio delle società sportive e previsioni: un approccio sperimentale

Relatore
Ch. Prof. Andrea Borghesan

Laureando
Silvio Nalli
Matricola 845460

Anno Accademico
2013 / 2014

Indice

Indice delle Figure

Indice delle Tabelle

Introduzione

Questo lavoro nasce dal riconoscimento della crescente rilevanza della performance economica nel mondo dello sport, ormai posta allo stesso piano della prestazione sportiva.

Il motore della competizione nel calcio è da sempre il perseguimento del risultato sportivo, che coindice con la vittoria del maggior numero possibile di gare disputate. Soggiogata dalla necessità di vincere trofei, la gestione economico-finanziaria delle società calcistiche è stata trascurata per molti decenni. Lo scenario è mutato negli anni Novanta del secolo scorso, quando i club professionistici sono diventati imprese a tutti gli effetti: alla ricerca della prestazione sportiva si è aggiunto l'obiettivo del conseguimento dell'equilibrio economico e finanziario della società.

La tesi qui elaborata si pone l'ambizioso obiettivo di sviluppare e testare un modello previsionale dei risultati delle squadre calcistiche partecipanti al campionato di serie A nell'arco di una stagione. Per la definizione delle variabili necessarie alla misurazione dei risultati sportivi ci si baserà sul database Opta, l'azienda leader mondiale di raccolta e analisi di dati sportivi. Sintetizzati dal Loro database gli indici più significativi, verrà implementata una rete neurale artificiale a fine di stimare le prestazioni sportive delle società disputanti la massima serie del campionato di calcio italiano.

Successivamente, si valuteranno le potenzialità del nostro modello come uno strumento di supporto decisionale per le società calcistiche in fase di costruzione delle rosa di giocatori. Infatti, fissato un traguardo di punti da raggiungere, la rete neurale, attraverso una previsione, stimerà se la squadra sia in grado di raggiungere la posizione desiderata. In caso contrario, verranno valutati alcuni scenari di modifica della composizione della rosa (inserendo giocatori provenienti da altre società di serie A che lo società possa permettersi in termini economici) e stimato il punteggio che la rosa modificata è in grado di ottenere.

All'analisi matematico-statistica verrà affiancato uno studio sulle variabili economico-finanziarie delle società calcistiche, ricavate dai rispettivi bilanci sportivi, allo scopo di valutare se sono presenti gruppi di società caratterizzati da una situazione economico-finanziaria simile. L'analisi verrà condotta attraverso l'implementazione di un *cluster analysis*, in cui verrà inserito come dataset una serie di indici economici, patrimoniali e

finanziari desunti dalla riclassificazione dei bilanci sportive delle società partecipanti alla serie A nelle stagioni 2010/11, 2011/12, 2012/13.

Il lavoro si suddivide quindi in 5 sezioni. Il primo capitolo descriverà l'evoluzione dello scenario calcistico italiano, dall'importazione del gioco dall'Inghilterra negli anni a cavallo tra la fine del XIX secolo e l'inizio del XX secolo fino al contesto attuale. Successivamente, esporrà quali variabili sono state scelte per l'analisi economico-finanziaria e quali invece sono state ritenute significative per il modello previsionale.

Il secondo e il terzo capitolo saranno dedicati alla presentazione teorica delle tecniche di *data mining* utilizzate nei due diversi modelli. Nello specifico, nella seconda sezione si descriveranno le reti neurali, dai cenni storici sul loro sviluppo alle tipologie di architetture e algoritmi elaborate. Esse, basandosi su una architettura simile alla struttura neurale del cervello umano, rappresentano un utile strumento per la risoluzione di modelli non lineari, caratterizzati cioè da un'elevata complessità della relazione tra le variabili. Il terzo capitolo si occuperà invece della descrizione delle tecniche di classificazione e raggruppamento, in particolare la *cluster analysis*. Essa consiste nel dividere un dataset in gruppi più piccoli chiamati *cluster* in base a criteri di dissimilarità decisi a priori, consentendo di valutare la composizione dei gruppi così formati a posteriori.

Nel quarto e nel quinto capitolo si implementeranno i due modelli di analisi proposti. Nello specifico, la quarta sezione esporrà l'analisi economico-finanziarie delle società disputanti la serie A per le stagioni 2010/11, 2011/12, 2012/13. Verranno quindi inseriti gli indici desunti dalla riclassificazione dei bilanci societari all'interno del modello di *cluster analysis*, al fine di valutare se sono presenti società con valori economici e finanziari simili ed eventualmente se alcune di esse sono caratterizzate da un trend migliorativo o peggiorativo della situazione economico-finanziaria nell'arco delle stagioni considerate. Iel quinto capitolo si occuperà, invece, della descrizione ed elaborazione del modello pratico di rete neurale. L'obiettivo sarà quello di valutare se le variabili ottenute dall'analisi dei match calcistici sono significative ai fine della previsione del risultato sportivo. In caso affermativo, si determinerà l'eventuale capacità dello strumento implementato a fornire supporto decisionale al management in fase di costruzione della rosa di calciatori necessaria ad affrontare la competizione.

1 Le variabili in gioco

1.1 Da gioco a business

Il calcio è solo da pochi decenni considerato un business: fin dagli albori, è sempre stato considerato solo un gioco. I primi a calciare un pallone sembra siano stati i Cinesi, praticanti lo *Tsu-chu* già dal IX secolo a.C., ma ne rivendicano l'origine anche i Greci, inventori di un gioco chiamato *episkyros* nel IV secolo a.C. Quest'ultimo gioco, un mix tra il calcio e il rugby moderno ma molto più violento, fu adottato dai Romani, che lo chiamarono *Harpastum* e fissarono alcune regole precise: le squadre, che dovevano essere in egual numero, si contendevano con le mani o con i piedi la palla in un campo rettangolare, delimitato da linee di contorno e da una linea centrale. Lo scopo era poggiare la palla sulla linea di fondo del campo avversario. Il gioco era molto popolare tra i legionari, che lo diffusero in tutte le province, permettendone una capillare diffusione. Nel Medioevo se ne ricordano due varianti: la *Soule*, praticata nei villaggi francesi, e il *Calcio Fiorentino*, diffuso a Firenze e nelle città limitrofe.

La patria del calcio moderno è l'Inghilterra: nel 1857 fu fondato il primo club della storia, lo *Sheffield Football Club*, mentre nell'ottobre del 1863 i rappresentanti delle 11 squadre allora esistenti diedero vita alla *Football Association* (FA) e codificarono il regolamento del gioco. Una ventina d'anni più tardi, nel 1886, le quattro federazioni britanniche (a quella inglese si erano aggiunte la scozzese nel 1873, la gallese nel 1876 e l'irlandese nel 1880) istituirono *l'International Football Association Board*. Nato per promulgare un regolamento comune mediando in un testo unico le specificità dei vari movimenti nazionali, l'IFAB rappresenta tuttora l'unico organo dotato del potere di stabilire qualsiasi modifica ed innovazione delle regole del gioco del calcio a livello internazionale e nazionale, vincolando alla loro osservanza tutte le federazioni, organizzazioni ed associazioni calcistiche.

Sempre nel 1886 viene ufficialmente riconosciuto il professionismo sportivo: i calciatori sono cioè equiparati alle altre categorie di lavoratori e devono conseguentemente percepire un compenso per l'opera prestata. Il denaro necessario a remunerare i calciatori veniva guadagnato attraverso la vendita di biglietti per assistere al match: la trasformazione

del calcio da sport dilettantistico a sport professionistico è funzionale all'evoluzione del calcio da sport a business[1].

Il gioco divenne sempre più popolare nel corso dei primi anni del Novecento e, nonostante il trend di crescita dei ricavi fosse positivo, il denaro incassato non riusciva a coprire i costi di compravendita dei calciatori e gli stipendi a loro assegnati. Le questioni finanziarie non erano tenute in gran considerazione a quei tempi: la maggior parte degli spettatori assisteva alle partite da un terrapieno a cui potevano accedere senza pagare, mentre le tribune erano poco capienti e riservate al pubblico più facoltoso. Fu per questo necessario l'apporto di capitale da parte di uomini d'affari locali per salvare le squadre dai debiti sempre crescenti e per finanziare opere strutturali necessarie al club, quali nuovi impianti sportivi e stadi più capienti, gestendo il club con un'ottima imprenditoriale per cercare di massimizzare le fonti di guadagno.

Nel corso dei successivi decenni del secolo scorso il calcio è diventato lo sport più popolare al mondo, sia per numero di spettatori che per numero di praticanti[2], grazie all'istituzione di competizioni continentali, come la *UEFA Champions League*, e intercontinentali, tra tutte il Campionato del Mondo FIFA.

Riassumendo, gli elementi fondamentali del business del calcio, rimasti sostanzialmente tali dagli albori del professionismo, sono[3]:

- la presenza di un prodotto, ovvero l'intrattenimento offerto dal gioco stesso;
- la vendita del prodotto ai clienti, rappresentati dai tifosi e gli appassionati;
- i lavoratori interessati, tra cui si annoverano i calciatori, i tecnici e i membri del loro staff, i dirigenti e i consulenti della società;
- i terreni, gli equipaggiamenti e i fabbricati utilizzati;
- un ambiente competitivo che necessita però anche un certo grado di cooperazione con i concorrenti.

1.2 La nascita del movimento calcistico italiano

In Italia, il movimento calcistico prese piede nell'ultimo decennio del XIX secolo: in

[1] U. Lago, A. Baroncelli e S. Szimanski, *Il business del Calcio. Successi sportivi e rovesci finanziari*, Egea, Milano, 2004, pag. 19.
[2] 1.098.450 di calciatori solo in Italia. Fonte: Report Calcio FIGC 2014.
[3] U. Lago, A. Baroncelli e S. Szimanski, *cit.*, pag. 20

quegli anni il calcio era praticato all'interno di polisportive, che prevedevano sezioni differenti a seconda dello sport praticato. Il gioco era in parte diverso a quello attuale, ed era comunemente chiamato calcio ginnastico: le prime manifestazioni di calcio furono infatti durante i tornei della Federazione Ginnastica Nazionale Italiana (FGNI)[4]. Il primo club fondato ufficialmente fu il Genoa Cricket and Athletic Club nel 1893[5], a cui seguirono lo Sport-Club Juventus nel 1897, Il *Milan Foot-Ball and Cricket Club* nel 1899 e il Football Club Pro Vercelli nel 1903 (facente parte della Società Ginnastica Pro Vercelli, fondata nel 1892).

La promulgazione e codificazione del gioco, ad opera del professore di educazione fisica Federico Gabrielli, prese piede nella città di Rovigo nel 1893 e in quella di Treviso nel 1896, dove si svolsero le prime gare nel Concorso interprovinciale Ginnastico del 1896. Il regolamento fu pubblicato per la prima volta nel 1985, e conteneva le regole della *Football Association* tradotte in italiano. Fu successivamente e gradualmente uniformato a quello stilato dall'*International Football Association Board* (IFAB), al quale si aderì formalmente nel 1903. La FGNI continuò ad operare fino al 1907, in concomitanza con la Federazione Italiana Football (FIGC dal 1909), fondata nel 1898 per promuovere il gioco nel Paese.

L'8 maggio 1898 si disputò il primo campionato italiano di calcio, a cui parteciparono 4 squadre: il titolo fu conquistato dal Genoa, che vinse anche per i due anni successivi. A partire dalla stagione 1909/10 la FIGC decise di modificare il torneo, adottando un Girone Unico la cui classifica determinava la vincente del titolo.

Il sistema fu perfezionato nel 1929/30, anno in cui il torneo prese la denominazione di Serie A. Questa data segnò di fatto uno spartiacque nella storia del campionato: da allora, la maggior parte delle affermazioni è infatti stata appannaggio delle cosiddette "grandi", ovvero la Juventus (il club attualmente più scudettato), l'Inter e il Milan, cui solo il Bologna e il Torino, negli anni trenta e quaranta, seppero porre un serio contrasto.

1.3 L'evoluzione della società calcistica professionistica

Il fine principe delle società calcistiche è sempre stato il perseguimento del risultato sportivo, relegando quello economico spesso in secondo piano. Ciò ha causato una situazione instabile e deficitaria dal punto di vista economico-finanziario, che la FIGC e il legislatore nazionale hanno cercato di risolvere attraverso l'approvazione di diversi provvedimenti atti a strutturare il modello di gestione assunto dai club professionistici.

[4] M. Romanato, F*rancesco Gabrielli (1857-1899). Le origini del calcio in Italia: dalla ginnastica allo sport*, Treviso, Antilia, 2008.

[5] *La storia del ritrovamento dell'atto fondativo del Genoa CFC*, in `fondazionegenoa.com`.

Nel dettaglio, si possono individuare tre fasi storiche che hanno portato all'attuale fisionomia giuridica ed economica del sistema calcio italiano.[6]

Gli anni Sessanta e Settanta

Il passaggio da gioco svolto nel tempo libero ad attività organizzata si ebbe in Italia solo intorno agli anni Sessanta, quando l'innalzamento del livello tecnico delle competizioni nazionali, la nascita di tornei internazionali e il progressivo interessamento dei media dell'epoca spinse diverse squadre ad aumentare l'impegno economico, nel tentativo di affrontare in modo soddisfacente le nuove sfide a livello internazionale. La dimensione societaria dell'associazione sportiva, istituita dal *Comitato Olimpico Nazionale Italiano* (CONI)[7], divenne del tutto inadeguata: le associazioni non riuscirono più a far fronte alle spese crescenti attraverso il solo contributo volontario dei soci, e cercarono di reperire fondi rivolgendosi al mercato, assumendo gradualmente connotati di natura imprenditoriale. Ciò portò alla scomparsa della figura dell'associato-praticante a favore dell'atleta professionista, i cui elevati costi di acquisto e gestione acuirono la situazione di deficit finanziario di molti club.

La non più sostenibile condizione delle associazioni sportive richiese la revisione della forma giuridica assunta: con una serie di delibere del 1966, la FIGC attuò la trasformazione delle associazione sportive a società di capitali, aventi uno statuto indicato dalla Federazione stessa e obbligandole a tenere un sistema contabile economico-patrimoniale. In particolare, lo statuto escludeva (art. 3) che l'oggetto sociale potesse avere finalità diverse da quella sportiva e vietava (art. 22) lo scopo di lucro soggettivo, prevedendo la destinazione di eventuali utili di bilancio alle attività di carattere sportivo della società stessa.

La riforma non ebbe gli effetti sperati: la causa è da ricercare nei limiti statuari imposti dalla Federazione. L'impossibilità di ottenere una remunerazione economica del capitale investito porta ad una scarsa attenzione degli investitori verso la gestione economico-finanziaria equilibrata della società, spingendo quest'ultimi a ricercare esasperatamente il successo sportivo per ottenere forme di remunerazioni indirette, quali popolarità e prestigio sociale. Ciò innescò un meccanismo a spirale di aumento dei costi: le perdite di complessive delle società di serie A e B passarono da 18 milioni di euro nel 1972 a più di 44 milioni di euro nel 1980[8].

[6] M. Mancin, *Il bilancio delle società sportive professionistiche. Normativa civilistica, principi contabili nazionali e internazionali (IAS/IFRS)*, CEDAM, Venezia, 2009, pag. 3.
[7] Legge n. 426 del 16 dicembre 1942
[8] M. Mancin, *cit.*, pag. 11.

Dal 1981 alla metà degli anni Novanta

Alla difficile situazione finanziaria delle società si affiancò, nel 1978, il blocco del calcio mercato imposto dalla Pretura di Milano, riscontrando il possibile reato di mediazione di manodopera a scopo di lucro. Il fatto richiese un intervento tempestivo del legislatore per assicurare il regolare svolgimento del torneo: i lavori presero corpo nella Legge n. 91 del 23 marzo 1981 sulle *"Norme in materia di rapporti tra società e sportivi professionisti"*. Nel dettaglio, la disposizione legislativa stabiliva:

- la forma giuridica, riconoscendo le sole società sportive costituite come società per azioni o società a responsabilità limitata la capacità di stipulare contratti con atleti professionisti;
- l'oggetto sociale delle società, che obbliga a reinvestire gli eventuali utili «per il perseguimento della attività sportiva»;
- il divieto di lucro soggettivo, riconoscendo all'impresa la facoltà di generare utili, ma non di distribuirli tra i soci;
- più stringenti controlli sulla gestione e sull'operato delle società attribuiti al CONI e alla Federazione.

La riforma, pur compiendo dei passi avanti rispetto allo statuto del 1966, non riuscì a porre rimedio alla difficile situazione finanziaria in cui versavano le società calcistiche italiane: il grave squilibrio esistente tra ricavi e i costi della gestione corrente, causa principale della crisi del tempo, non fu oggetto del sistema dei controlli sopra citato[9].

Al contrario, l'intervento del legislatore mutò in maniera radicale il rapporto esistente tra atleta e società. Nella situazione preesistente, il legame tra club e calciatore era disciplinato da due rapporti diversi: il rapporto di lavoro sportivo, che traeva origine dal contratto di ingaggio ed in cui si stabiliva il compenso spettante all'atleta in corrispettivo della prestazione sportiva fornita; ed il rapporto di vincolo sportivo, che garantiva alla società il diritto esclusivo di assicurarsi le prestazioni sportive del giocatore per tutta la sua carriera agonistica. Quest'ultimo vincolava di fatto il calciatore, al momento della stipula del contratto, alla volontà della società, non lasciandoli nessuna libertà di scelta di trasferimento verso altre società. Il trasferimento di un giocatore, infatti, poteva avvenire solamente attraverso una cessione del vincolo verso un altro club, a fronte di

[9] M. Mancin, *cit.*, pag. 20.

un corrispettivo concordato dalle parti[10]. Inoltre, tutelando i diritti alle prestazioni sportive degli atleti delle società, assumeva una notevole rilevanza patrimoniale per le stesse società calcistiche.

La legge 91/81 prevede, all'articolo 16, la graduale estinzione del rapporto di vincolo sportivo nell'arco dei 5 anni successivi alla sua entrata in vigore, creando due fattispecie riguardanti la cessione di un calciatore:

- cessione del contratto, che stabilisce una durata massima del contratto tra calciatore e società pari a 5 anni e consente la cessione dello stesso ad altra società, purché vi sia il consenso dell'atleta e si rispettino le norme federali fissate in materia;
- trasferimento del calciatore a scadenza del contratto, che consente al calciatore, giunto in scadenza di contratto, di trasferirsi ad una società di suo gradimento, previo il versamento da parte di quest'ultima di un'Indennità di Preparazione e Promozione(IPP).

L'alto valore dell'indennità richiesta dai club (non era infatti previsto un importo limite) ripropose di fatto l'appena estinto vincolo sportivo: l'atleta in scadenza di contratto si trovava costretto a scegliere tra una delle poche società aventi la disponibilità economica necessaria a versare l'indennità, oppure a rinnovare il proprio contratto con il club in cui militava. Ciò produsse un ulteriore aumento del costo del cartellino dei calciatori e dei rispettivi ingaggi: le difficoltà ad acquistare un altro giocatore sul mercato per l'alto valore delle IPP costrinsero le società ad offrire remunerazioni maggiori per indurre i calciatori a rinnovare il contratto in scadenza.

Tutto questo, unito alla conferma del divieto di lucro soggettivo, acuì ulteriormente la già difficile situazione finanziaria in cui il sistema calcio si trovava: in questo contesto, il colpo definitivo lo assestò la sentenza Bosman.

Dal 1996 ai giorni nostri

Lo scenario descritto si mantenne sostanzialmente inalterato anche durante la prima metà degli anni Novanta, nonostante le crescenti fonti di introito determinate dalle cosiddette *pay-tv*: i bilanci con data di chiusura giugno 1995 rilevavano, in termini aggregati, perdite superiori ai 54 milioni di euro per i soli club di serie A[11].

[10] U. Lago, A. Baroncelli e S. Szimanski, *cit.*, pag. 47.
[11] M. Mancin, *cit.*, pag. 31

La sentenza Bosman, pronunciata alla fine dello stesso anno, mutò completamente il panorama calcistico italiano ed europeo: la questione riguardava la causa intentata dal calciatore professionista Jean Marc Bosman, di nazionalità belga e tesserato per la società Royal Liegeois, presso la Corte d'appello di Liegi contro L'UEFA, la Federazione belga e il suo club di appartenenza. L'oggetto del contendere esaminava il diritto delle società a richiedere un'indennità per il trasferimento di giocatori giunti a scadenza di contratto e la libertà concessa alle federazioni sportive di limitare, attraverso norme statuarie, l'accesso di cittadini stranieri appartenenti agli Stati aderenti alla Comunità Europea. La corte d'appello belga rinviò la questione in via pregiudiziale alla Corte di Giustizia europea, che si pronunciò sulle questioni sollevate con la sentenza C-415/93 del 15 dicembre 1995[12]. La sentenza, produttiva di effetti verso tutti gli Stati membri, accolse il ricorso del calciatore sulle questioni sollevate, provocando la decadenza immediata di tutte le norme contrarie e quindi un vuoto normativo in tutte le legislazioni dei Paesi UE: si venne a creare una discriminazione tra la cessione di giocatori all'interno di un paese (in cui le indennità erano ancora valide) e il trasferimento di un atleta nel mercato comunitario, in cui le IPP erano state abolite e che quindi poteva avvenire a parametro zero.

La sentenza ebbe effetti deleteri anche riguardo il metodo di contabilizzazione dei diritti alle prestazioni dei calciatori: esso consisteva nell'iscrizione in Stato patrimoniale del costo di acquisto, al netto dell'IPP che si presumeva di incassare alla scadenza del contratto (si riteneva l'indennizzo il probabile valore di realizzo). L'abrogazione delle indennità avrebbe quindi appesantito ulteriormente la già gravosa situazione economica delle società sportive. Ciò richiese un tempestivo intervento del legislatore, che non tardò ad arrivare. Con il D.L 485 del 20 settembre 1996 si attua la riforma della Legge 91/8, attraverso:

- l'eliminazione dell'obbligo di versare l'IPP per il trasferimento di atleti professionisti a scadenza di contratto[13]. L'indennità rimane applicabile solo nel caso di firma del primo contratto professionistico di un calciatore;
- la possibilità di «spalmare», derogando dai principi civilistici vigenti, la perdita causata dall'eliminazione dell'IPP nei tre esercizi successivi;
- introduzione della finalità di lucro per le società sportive professionistiche;

[12] Per una lettura integrale della sentenza: `http://curia.europa.eu/juris/showPdf.jsf;jsessionid=9ea7d2dc30db71e267fcd10c4da5a5caa3076107f57d.e34KaxiLc3qMb40Rch0SaxuKaNr0?text=&docid=99445&pageIndex=0&doclang=IT&mode=lst&dir=&occ=first&part=1&cid=900213`

[13] Già previsto dal D.L 272 del 17 maggio 1996 "Disposizioni urgenti per le società sportive", non convertito in legge entro il termine previsto di 60 giorni.

- la nomina obbligatoria di un collegio sindacale per le società sportive;
- la limitazione al solo ambito sportivo dell'incidenza dei controlli svolti dalla Federazione.

La nuova disciplina, convertita nella legge 586/1996, completava definitivamente lo sviluppo dei club in società di capitali: essi potevano svolgere, oltre all'attività sportiva, qualsiasi altra attività economica strumentale a quella tipica[14].

Gli effetti della sentenza Bosman non si circoscrissero al solo livello normativo: toccarono infatti l'aspetto gestionale dei club, contribuendo alla nascita di un nuovo modello di business, legato maggiormente a dinamiche di sponsorizzazione, di investimento e di visibilità internazionale.

Il motore di questo trend di crescita fu l'aumento vertiginoso dei ricavi causato dall'ingresso delle pay-tv nel mondo calcistico, trasformandolo in spettacolo televisivo[15].

La figura 1.1 riassume i ricavi da diritti televisivi nell'arco dell'ultimo trentennio.

Osservando l'andamento del grafico, si notano alcuni punti di svolta:

- nella stagione 93/94 la RAI perde l'esclusiva sui diritti televisivi del campionato di calcio, in parte acquistati dalla piattaforma satellitare Tele +. L'ingresso nel mercato della *pay tv* fa balzare i ricavi da 55,9 milioni di euro della stagione 92/93 a 92,9 milioni di euro in una sola stagione. La negoziazione avveniva collettivamente: La Lega Calcio contrattava, per tutti i club a lei affiliati, i diritti televisivi sia per le trasmissioni in chiaro che per quelle criptate;
- dalla stagione 96/97 ai telespettatori viene fornito lo strumento della *pay per view*, ovvero la possibilità di acquistare il singolo evento sportivo che desiderano vedere. I ricavi per quella stagione salgono a 204 milioni di euro a fronte dei 104 milioni di euro della stagione precedente. La contrattazione resta collettiva e gestita dalla Lega Calcio;

[14] U. Lago, A. Baroncelli e S. Szimansky, cit., pag 52.
[15] U. Lago, A. Baroncelli e S. Szimansky, *cit.*, pag 53.

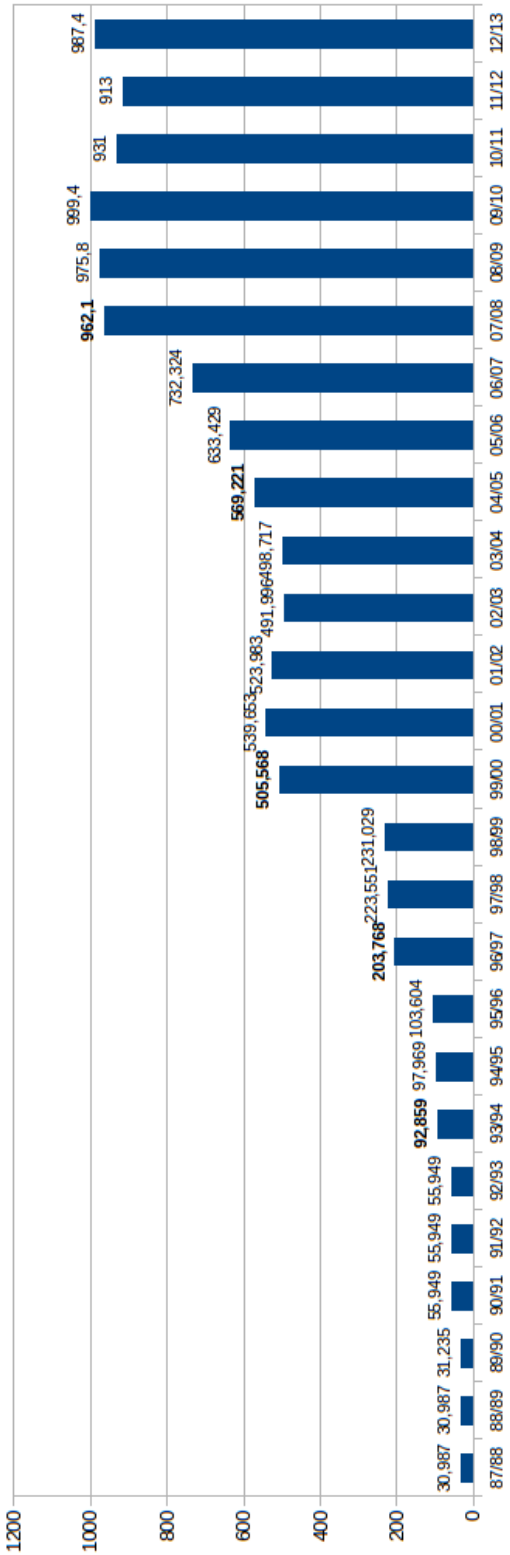

Figure 1.1: Proventi da diritti televisivi in Italia del sistema calcio (dati in milioni di euro)
Fonte: M. Mancin, cit., ed elaborazione personale su dati Lega Calcio e Report Calcio

- per legge[16], dalla stagione 99/00 un singolo operatore può acquistare massimo il 60% dei diritti in forma criptata per le partite di serie A: i maggiori club possono ora contrattare autonomamente con le emittenti televisivi (a Tele + si era affiancata Stream) l'importo per la cessione dei diritti per i match giocati "in casa". Ciò provoca la crescita esponenziale del valore dei diritti, che sfondano quota 500 milioni di euro per la stagione 99/00, più che raddoppiando i 231 milioni di euro della stagione precedente;

- dalla stagione 04/05 entra nel mercato, affiancando la piattaforma satellitare Sky (subentrata alle precedenti), il digitale terrestre che contribuisce, anche grazie allo sviluppo delle reti UMTS e Internet, a una nuova fase espansiva di questi ricavi, superando la fase di assestamento avvenuta nelle stagioni precedenti;

- dalla stagione 07/08, con l'implementazione di Premium Calcio da parte di Mediaset Premium e il ritorno della contrattazione collettiva decisa dal legislatore nella legge n. 9/08[17], i ricavi da diritti tv si impennano fino ad avvicinare il miliardo di euro. Tali valori si sono poi assestati nelle stagioni successive intorno ai 950 milioni di euro.

Nonostante l'enorme mole di introiti causati dai diritti tv, le perdite di bilancio non hanno accennato a diminuire: i club, costretti ad offrire ai giocatori stipendi sempre maggiori nel tentativo di vincolarli alla società o strapparli alla concorrenza alla scadenza di contratto, sono stati investiti dal circolo vizioso diritti televisivi – stipendi, rappresentato dalla figura 1.2. Quest'ultimi sono così cresciuti quasi allo stesso ritmo dei ricavi totali, nella convinzione che gli introiti delle stagioni successive avrebbero sostenuto la loro folle corsa. Il punto di non ritorno viene raggiunto nella stagione 02/03, quando i ricavi da diritti televisivi subiscono un importante ridimensionamento: le società, per contenere le perdite, attuano politiche di doping amministrativo e contabilità creativa[18].

[16] D.L. n. 15 del 30 gennaio 1999, contenente "*Disposizioni urgenti per lo sviluppo equilibrato dell'emittenza televisiva e per evitare la costituzione o il mantenimento di posizioni dominanti nel settore radiotelevisivo*", convertito nella legge 78/99.

[17] Sui risvolti di tale provvedimento si veda: A. de Martini, *La disciplina dei diritti televisivi nello sport*, in Rivista di Economia dello Sport, Vol. VII, fasc. 2, 2011.

[18] L. A. Bianchi e D. Corrado, *Bilanci delle società di calcio. Le ragioni di una crisi.* Egea, Milano, 2004.

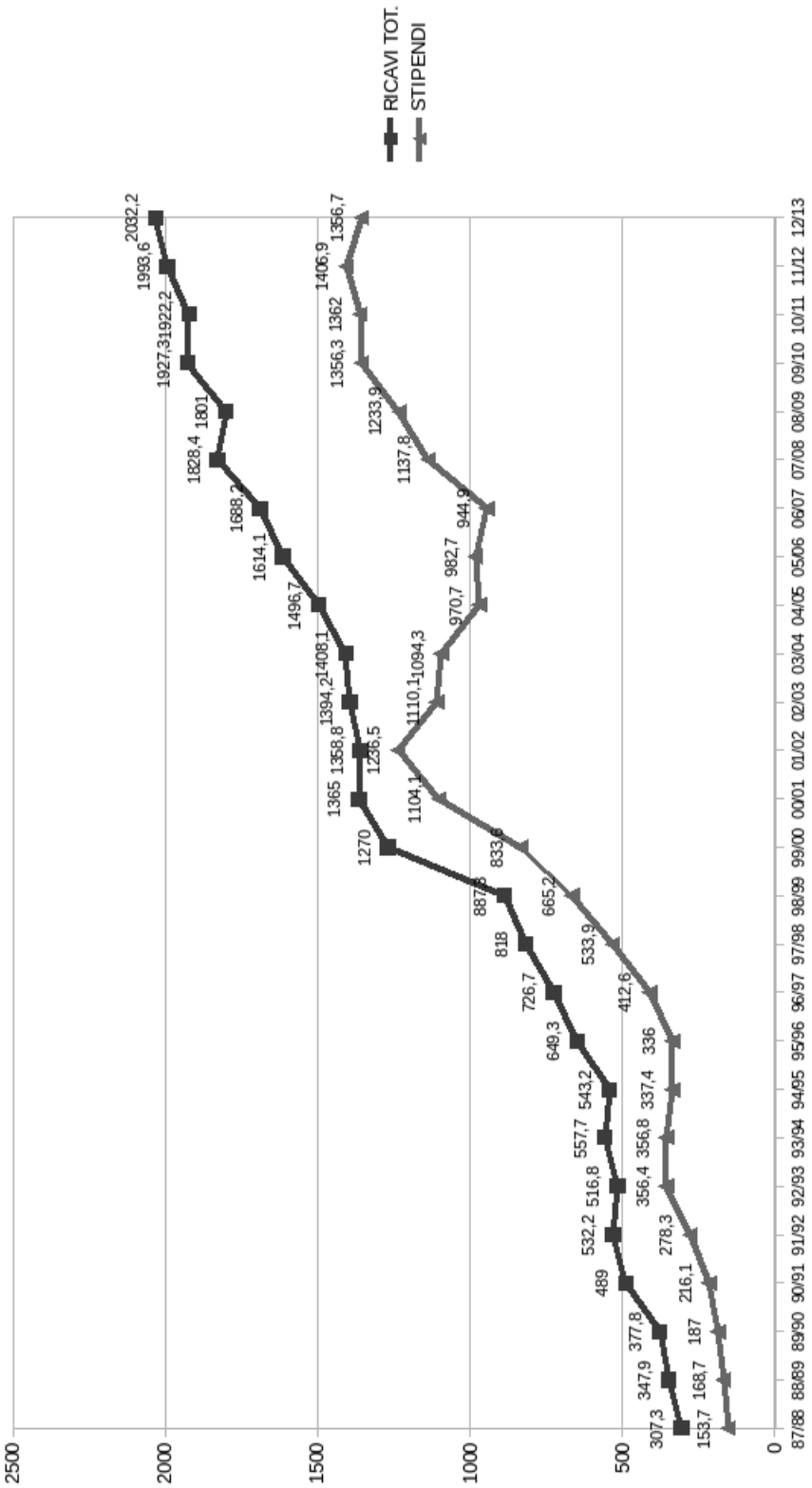

Figure 1.2: Rapporto-ricavi-totali/stipendi per valori aggregati delle società di serie A e B (valori in milioni di euro)
Fonte: M. Mancin, cit.*, ed elaborazione personale su dati Report Calcio.*

La prassi era di gonfiare, al momento della cessione, il valore del cartellino di un calciatore, creando artificialmente plusvalenze: la tecnica copriva in parte le perdite nel breve periodo, ma appesantiva la situazione economica nel lungo per le pesanti quote di ammortamento che gravavano sui bilanci futuri. Il legislatore fu quindi costretto ad intervenire nuovamente, emanando il cosiddetto «decreto salva calcio»[19] che consentiva alle società sportive di ripartire nell'arco di un decennio (termine poi ridotto a 5 anni) le svalutazioni causate dalle perdite durevoli di valore dei DPC anziché spesarle nell'esercizio in cui erano maturate.

Nell'ultimo decennio, anche L'UEFA ha preso a cuore la salute finanziaria delle società sportive, adottando due nuovi strumenti di controllo gestionale: Il Manuale delle Licenze UEFA e il *fair play* finanziario.

Il Manuale delle licenze introduce un sistema di licenze, a livello europeo, necessario all'ammissione delle società alle competizioni organizzate dalla UEFA (Champions League e Europa League[20]). Esso rappresenta l'ambizioso tentativo di introdurre un sistema di certificazione della qualità della gestione di una società di calcio professionistica, in tutti gli aspetti che la compongono: attività sportiva della prima squadra, attività giovanile, gestione dello stadio, organizzazione interna, gestione economico-finanziaria. In base a tale sistema, sono ammesse alle competizioni UEFA le sole società che, oltre ad aver conseguito il titolo sportivo durante la competizione nazionale, siano in possesso di specifici requisiti di natura legale, infrastrutturale, organizzativa ed economico-finanziaria. La licenza ha validità per una sola stagione sportiva ed è rilasciata, in Italia, dalla FIGC su delega dell'UEFA, conformemente ai principi sanciti dal Manuale delle Licenze. Gli obiettivi che si è posta l'UEFA con tale strumento sono[21]:

- ottimizzare la gestione economica-finanziaria dei club;
- fornire maggiore trasparenza e credibilità al sistema calcio;

[19] Legge n. 27 del 21 febbraio 2013.

[20] Nel 2000, anno della sua redazione da parte del Comitato Esecutivo dell'Uefa, le competizioni interessate erano Champions League, Coppa Uefa e Intertoto.

[21] A. Bernoldi, C. Sottoriva, *La disciplina della redazione del bilancio di esercizio delle società di calcio. Confronto con l'esperienza internazionale e impatto del cd. «Financial fair play»*, in Rivista di diritto ed economia dello sport, Vol VII, Fasc 1, 2011, pag. 174.

- fornire garanzie ai creditori;
- dare continuità nella partecipazione alle competizioni UEFA;
- aumentare la correttezza delle competizioni europee sotto il profilo economico-finanziario;
- creare un mercato più attraente per gli investitori e i partner commerciali.

Questo strumento è stato fondamentale per il mantenimento di un adeguato equilibrio economico-finanziario nella delle società di Serie A. In particolare, fondamentale è stata l'introduzione del criterio di assenza di debiti scaduti da trasferimento di calciatori[22], che ha contribuito a riformare l'approccio gestionale di molti club europei: le società (quelle che non avevano un Presidente in grado di ripianare personalmente i debiti contratti), pur di non perdere la remunerazione economica associata alle competizioni UEFA, hanno cercato di pagare i debiti contratti cedendo alcune loro attività patrimoniali.

L'altro progetto posto in essere dal Comitato Esecutivo dell'UEFA nel 2009 è il fair play finanziario, per cercare di risollevare i risultati economici negativi delle società sportive. Nonostante l'introduzione del sistema delle licenze avesse parzialmente sanato la situazione debitoria di molti club, su di essi pendeva ancora la spada di Damocle dell'altissimo monte ingaggi dei calciatori. L'idea di fondo è di far rientrare le società dai debiti contratti nell'arco di un triennio e condurle poi sulla via dell'auto-sostentamento finanziario. Gli obiettivi prefissati sono[23]:

- introdurre maggior rigore e razionalità nel sistema finanziario delle società;
- incentivare l'auto-sostenibilità delle società, soprattutto a lungo termine;
- stimolare gli investimenti in infrastrutture e nei settori giovanili;
- incoraggiare la società a competere entro i propri introiti;
- accertarsi che le società onorino gli impegni finanziari nei tempi prestabiliti;

[22] Criterio F.03: "La Società richiedente la Licenza deve dimostrare di non avere, alla data del 31 marzo che precede la Stagione della Licenza, debiti scaduti nei confronti di altre società di calcio, derivanti da trasferimenti di calciatori, intervenuti fino al 31 dicembre precedente", consultabile all'indirizzo: `http://www.figc.it/it/105/3816/Norme.shtml`

[23] Il testo integrale del Manuale ufficiale UEFA sul *finalcial fair play* può essere consultato qui: `http://www.uefa.org/MultimediaFiles/Download/Tech/uefaorg/General/01/80/54/10/1805410_DOWNLOAD.pdf` .

- contenere le pressioni sulle richieste salariali e sui trasferimenti;
- limitare gli effetti dell'inflazione nel mondo calcistico.

Le regole sulle licenze per club e sul fair play finanziario, approvate a maggio 2010 dopo un lungo periodo di consultazione e aggiornate nell'edizione 2012, vengono implementate per un periodo triennale. Gli stipendi e le spese di ingaggio dei club che partecipano alle competizioni UEFA vengono monitorate dall'estate 2011. La prima valutazione dei bilanci è stata condotta nel 2013/14. Le società vengono monitorate attraverso il rispetto di 3 punti fondamentali:

1. Assenza di debiti arretrati verso altre società, dipendenti o autorità nazionali;
2. Fornitura di informazioni di natura finanziaria qualora richieste;
3. Obbligo del pareggio di bilancio.

Il requisito del «pareggio di bilancio» (calcolato come la differenza tra le voci di ricavo rilevanti e le voci di costo rilevanti) costituisce l'elemento fondamentale della norma ed è valutato nell'arco di tre stagioni sportive. Il Regolamento prevede alcune eccezioni:

- una soglia di tolleranza del deficit, posta nel limite di 5 milioni di euro;
- Il limite appena descritto può essere superato solo se il deficit è coperto attraverso l'impiego di mezzi propri e non supera i 45 milioni di euro nei primi 3 anni. Dalla stagione 2015/16, il limite di accettabilità verrà abbassato a 30 milioni di euro;
- In caso di deficit, sarà conteggiato l'eventuale surplus accumulato nelle stagioni precedenti.

Le società che non rispettano i 3 punti fondamentali sopra descritti, subiscono sanzioni di diversa natura da parte dell'UEFA, fino all'esclusione dalle competizioni europee per una o più stagioni.

Il successo riscosso dalla prima fase attuativa, con una riduzione dell'80% dei ritardi sui pagamenti e di 900 milioni di euro dei debiti complessivi dei club europei, ha risposto

alle critiche di coloro che giudicavano il progetto troppo ambizioso e di difficile implementazione, raccogliendo il supporto pressoché totale di tutti i portatori d'interesse.

1.4 Le variabili analizzate nel modello

Nel contesto calcistico italiano appena descritto, abbiamo analizzato i bilanci delle 20 società disputanti il campionato di serie A per le stagioni 2010/2011, 2011/12, 2012/13 al fine di individuare se le squadre possono essere raggruppate secondo criteri di similarità. I *cluster* così creati saranno poi valutati in base all'analisi matematico-statistica implementata con le variabili «di gioco».

L'elaborazione dei bilanci societari ha portato alla definizione di una serie di variabili da utilizzare come input per la cluster analysis:

- Totale Attività;
- EBITDA;
- EBIT;
- Utile (perdita) di esercizio;
- ROS;
- EBITDA/RICAVI;
- ROA;
- Indice di Capitalizzazione;
- Indice di solidità patrimoniale.

Nella descrizione di questi indici, si farà riferimento agli schemi di bilancio applicati alle società sportive riportati nella tabella 1.1 e nella tabella 1.2.

STATO PATRIMONIALE	
ATTIVO	**PASSIVO**
A. CREDITI VERSO SOCI PER VERSAMENTI ANCORA DOVUTI	**A. PATRIMONIO NETTO**
I) CAPITALE SOTTOSCRITTO NON RICHIAMATO	I. CAPITALE
II) CAPITALE RICHIAMATO NON VERSATO	II. RISERVA DA SOVRAPPREZZO AZIONI
B. IMMOBILIZZAZIONI con separata indicazione di quelle concesse in locazione finanziaria	III. RISERVE DI RIVALUTAZIONE
I. IMMOBILIZZAZIONI IMMATERIALI	IV. RISERVA LEGALE
1) Costi di impianto e ampliamento	V. RISERVE STATUARIE
2) Costi di ricerca,di sviluppo e pubblicità	VI. RISERVA PER AZIONI PROPRIE IN PORTAFOGLIO
3) Diritto di brevetto Industriale e diritti di utilizzazione delle opere dell'ingegno	VII. ALTRE RISERVE, DISTINTAMENTE INDICATE
4) Concessioni, licenze, marchi e diritti simili	1) Riserva di rivalutazione ex art. 2426, n. 4, c.c.
5) Avviamento	2) Riserva per deroghe ex art. 2423, 4° comma, c.c.
6) Immobilizzazioni in corso e acconti	3) Riserva ammortamenti anticipati
7) Capitalizzazione costi vivaio	4) Riserva straordinaria
8) Diritti pluriennali alle prestazioni dei calciatori	5) Riserva ex art. 4 legge n 586/ 96
9) Oneri pluriennali da rettifiche di valore ex art. 18 bis legge 91/1981	6) Riserva per versamenti in c/futuro aumento di capitale
10) Altre	7) Riserva per copertura perdite esercizi precedenti
TOTALE (I)	8) Riserva per copertura perdite esercizio in corso
II. IMMOBILIZZAZIONI MATERIALI	VIII. UTILI (PERDITE) PORTATI A NUOVO
1) Terreni e fabbricati	IX. UTILE (PERDITA) DELL'ESERCIZIO
2) Impianti e macchinari	**TOTALE (A)**
3) Attrezzature industriali e commerciali	**(I+II+III+IV+V+VI+VII+VIII+IX)**
4) Altri beni	
5) Immobilizzazioni in corso e acconti	**B. FONDO RISCHI E ONERI**
TOTALE (II)	1) Per trattamento di quiescenza e obblighi simili
III. IMMOBILIZZAZIONI FINANZIARIE	2) Per imposte, anche differite
1) Partecipazioni in	3) Altri
a) imprese controllate	**TOTALE FONDO RISCHI E ONERI (B)**
b) imprese collegate	
c) imprese controllanti	**C. TRATTAMENTO DI FINE RAPPORTO LAVORO SUBORDINATO**
d) altre imprese	**TOTALE (C)**
e) Compartecipazioni ex art. 102 bis N.O.I.F.	
2) Crediti con separata indicazione, per ciascuna voce, degli importi esigibili entro l'esercizio successivo	**D. DEBITI con separata indicazione, per ciascuna voce, degli importi esigibili oltre l'esercizio successivo**
a) verso imprese controllate	1) Obbligazioni ordinarie
b) verso imprese collegate	2) Obbligazioni convertibili
c) verso imprese controllanti	3) Debiti verso soci per finanziamenti
d) verso altri	4) Debiti verso banche
3) Altri titoli	5) Debiti verso altri finanziatori
4) Azioni proprie (di valore nominale complessivo pari a Euro …)	6) Acconti
TOTALE (III)	7) Debiti verso fornitori
TOTALE IMMOBILIZZAZIONI (B) (I+II+III)	8) Debiti rappresentati da titoli di credito
	9) Debiti verso imprese controllate
C. ATTIVO CIRCOLANTE	10) Debiti verso imprese collegate
I. RIMANENZE	11) Debiti verso imprese controllanti
1) Materiale di consumo	12) Debiti tributari
2) Prodotti in corso di lavorazione e semilavorati	13) Debiti verso istituti di previdenza e di sicurezza sociale
3) Lavori in corso su ordinazione	*14) Debiti per compartecipazioni ex art. 102 bis N.O.I.F*
4) Prodotti e finiti e merci	*15) Debiti verso enti-settore specifico*
5) Acconti	16) Altri debiti
TOTALE (I)	**TOTALE DEBITI (D)**
II. CREDITI CON SEPARATA INDICAZIONE, PER CIASCUNA VOCE, DEGLI IMPORTI ESIGIBILI OLTRE L'ESERCIZIO SUCCESSIVO	
1) Verso clienti	*(segue)*

2) Verso imprese controllate	**E. RATEI E RISCONTI PASSIVI**
3) Verso imprese collegate	I) RATEI PASSIVI
4) Verso imprese controllanti	II) RISCONTI PASSIVI
4-bis) Crediti tributari	III) AGGIO SUI PRESTITI
4-ter) Imposte anticipate	**TOTALE RATEI E RISCONTI PASSIVI (E)**
5) Crediti verso enti-settore specifico	
6) Verso altri	**TOTALE PASSIVO**
TOTALE (II)	
III. ATTIVITÀ FINANZIARIE CHE NON COSTITUISCONO IMMOBILIZZAZIONI	
1) Partecipazioni in imprese controllate	
2) Partecipazioni in imprese collegate	
3) Partecipazioni in imprese controllanti	
4) Altre partecipazioni	
5) Azioni proprie (di valore nominale complessivo pari ad Euro …)	
6) Altri titoli	
TOTALE (III)	
IV. DISPONIBILITÀ LIQUIDE	
1) Depositi bancari e postali	
2) Assegni	
3) Denaro e valori in cassa	
TOTALE (IV)	
TOTALE ATTIVO CIRCOLANTE (C) (I+II+III+IV)	
D. RATEI E RISCONTI ATTIVI	
I) RATEI ATTIVI	
II) RISCONTI ATTIVI	
III) DISAGGIO SUI PRESTITI	
TOTALE RATEI E RISCONTI ATTIVI (D)	
TOTALE ATTIVO	

Table 1.1: Schema di bilancio dello stato patrimoniale per le società sportive

CONTO ECONOMICO

A. VALORE DELLA PRODUZIONE

1) Ricavi delle vendite e delle prestazioni
   a) ricavi da gare in casa
   b) percentuale su incassi da gare da squadre ospitanti
   c) abbonamenti
2) Variazione delle rimanenze di materiali di prodotti in corso di lavorazione, semilavorati e finiti
3) Variazione dei lavori in corso su ordinazione
4) Incrementi di immobilizzazione per lavori interni e di capitalizzazione costi vivaio
5) Altri ricavi e proventi
   a) contributi in conto esercizio
   b) proventi da sponsorizzazioni
   c) proventi pubblicitari
   d) proventi commerciali e royalties
   e) proventi da cessione di diritti televisivi
      - proventi televisivi
      - percentuale su diritti televisivi da squadre ospitanti
      - proventi televisivi da partecipazione competizioni U.E.F.A.
   f) proventi vari
   g) ricavi da cessione temporanea prestazioni calciatori
   h) plusvalenze da cessione diritti pluriennali prestazioni calciatori
   i) altri proventi gestione calciatori
   l) ricavi e proventi diversi

TOTALE VALORE DELLA PRODUZIONE (A)

B. COSTI DELLA PRODUZIONE

6) Per acquisti materiale di consumo e di merci
7) Per servizi
8) Per godimento di beni di terzi
9) Per il personale
   a) salari e stipendi
   b) oneri sociali
   c) trattamento di fine rapporto
   d) trattamento di quiescenza e simili
   e) altri costi
10) Ammortamenti e svalutazioni
   a) ammortamenti immobilizzazioni immateriali
   b) ammortamenti immobilizzazioni materiali
   c) altre svalutazioni delle immobilizzazioni
   d) svalutazioni dei crediti compresi nell'attivo circolante e nelle disponibilità liquide
11) Variazioni delle rimanenze di materiale di consumo e di merci
12) Accantonamenti per rischi
13) Altri accantonamenti
14) Oneri diversi si gestione
   a) spese varie organizzazione gare
   b) tasse iscrizione gare
   c) oneri specifici verso squadre ospitate
      - percentuale su incassi gare a squadre ospitate
      - percentuale su diritti televisivi a squadre ospitate
   d) costi per acquisizione temporanea prestazioni calciatori
   e) minusvalenze da cessione diritti pluriennali prestazioni calciatori
   f) altri oneri da gestione calciatori
   g) altri oneri diversi di gestione

TOTALE COSTI DELLA PRODUZIONE (B)

DIFFERENZA TRA VALORE E COSTI DELLA PRODUZIONE (A-B)

(segue)

C. PROVENTI E ONERI STRAORDINARI
15) Proventi da partecipazioni
a) imprese controllate
b) imprese collegate
c) in altre imprese
16) Altri proventi finanziari
a) da crediti iscritti nelle immobilizzazioni
- imprese controllate
- imprese collegate
- imprese controllanti
b) da titoli iscritti nelle immobilizzazioni che non costituiscono partecipazioni
c) da titoli i scritti nell'attivo circolante che non costituiscono partecipazioni
d) da proventi diversi dai precedenti
- imprese controllate
- imprese collegate
- imprese controllanti
e) proventi da compartecipazioni ex art. 102 bis N.O.I.F.
17) Interessi ed altri oneri finanziari
a) verso imprese controllate
b) verso imprese collegate
c) verso imprese controllanti
d) altri oneri finanziari
e) oneri da compartecipazioni ex art. 102 bis N.O.I.F.
17 bis) Utile e predite su cambi
a) utile su cambi
b) perdite su cambi
TOTALE PROVENTI ED ONERI FINANZIARI (C)
D. RETTIFICHE DI VALORE DI ATTIVITÀ FINANZIARIE
18) Rivalutazioni
a) di partecipazioni
b) di immobilizzazioni finanziarie che non costituiscono partecipazioni
c) di titoli iscritti all'attivo circolante che non costituiscono partecipazioni
19) Svalutazioni
a) di partecipazioni
b) di immobilizzazioni finanziarie che non costituiscono partecipazioni
c) di titoli iscritti all'attivo circolante che non costituiscono partecipazioni
TOTALE DELLE RETTIFICHE (D)
E. PROVENTI E ONERI STRAORDINARI
20) Proventi
a) plusvalenze da alienazioni
b) sopravvenienze attive straordinarie
c) altri proventi straordinari
21) Oneri straordinari
a) minusvalenze da alienazioni
b) imposte relative ad esercizi precedenti
c) sopravvenienze passive straordinarie
d) altri oneri straordinari
TOTALE DELLE PARTITE STRAORDINARIE (E)
RISULTATO PRIMA DELLE IMPOSTE (A-B+ -C +-D+-E)
22) Imposte sul reddito d'esercizio
a) imposte correnti
b) imposte differite
c) imposte anticipate
UTILE (PERDITA) DELL'ESERCIZIO

Table 1.2: Schema di bilancio del conto economico per le società sportive

1.4.1 Totale Attività

Lo stato patrimoniale di una società fornisce una rappresentazione della consistenza quantitativa e della composizione qualitativa degli investimenti e delle fonti correlate di finanziamento esistenti alla chiusura dell'esercizio[24].

In questo indice vengono raccolte tutte le voci inserite nell'attivo dello Stato patrimoniale di una società sportiva. Esso può essere diviso in 4 macro-classi:

A) Crediti verso soci per versamenti ancora dovuti: rappresentano l'impegno dei soci di sottoscrizione di nuove quote di capitale sociale dei quali, in quanto crediti, non è ancora stato effettuato il conferimento;

B) Immobilizzazioni: sono considerati tali tutti i beni destinati a perdurare nella società. Esse sono a loro volta suddivise in:

- immobilizzazioni materiali: sono costituite dai beni della società tangibili di utilità pluriennale che vengono ammortizzati ogni anno. Tra essi figurano terreni, fabbricati, attrezzature, impianti;
- immobilizzazioni finanziarie, ovvero le attività finanziarie che costituiscono investimenti destinati ad essere utilizzati durevolmente;
- immobilizzazioni immateriali: quest'ultime incidono notevolmente sul valore totale dell'attivo delle società calcistiche, in quanto annoverano i diritti pluriennali alle prestazioni dei calciatori (DPC) e i costi di capitalizzazione del vivaio;

C) Attivo circolante: al suo interno sono considerati tutti i beni destinati ad essere utilizzati nell'esercizio successivo. Compongono questa voce le rimanenze, i crediti della società e le disponibilità liquide della stessa;

D) Ratei e i risconti attivi: derivano da operazioni comuni a due periodi consecutivi e raffigurano ricavi o costi di competenza all'esercizio, ma che sono maturati o rilevati in un altro esercizio.

[24] M. Valeri, *Standard IAS/IFRS e nuove esigenze di disclosure nel bilancio delle società di calcio*, Giappichelli editore, Torino, 2008, pag. 135.

1.4.2 EBITDA

Acronimo dell'inglese *Earnings Before Interest, Taxes, Depreciation and Amortization*, questo indice rappresenta l'utile prima degli interessi passivi, imposte, svalutazioni e ammortamenti su beni materiali e immateriali. Esso ha una duplice valenza informativa[25]:

1. è un indicatore sufficientemente oggettivo dell'andamento economico della gestione caratteristica;
2. rappresenta un flusso monetario potenziale, in quanto dal suo calcolo restano esclusi i costi non monetari (ammortamenti, accantonamenti e svalutazioni), ovvero a cui non corrispondono, nel periodo di sostenimento, equivalenti esborsi monetari.

Nel caso in cui l'EBITDA assuma un valore negativo, la società ha problemi di redditività e di flussi di cassa; viceversa, un suo valore positivo non indica necessariamente che l'azienda genera valore: questo perché vengono ignorate le variazioni del capitale circolante, in conto capitale, le imposte versate e gli interessi corrisposti.

Questo indice di performance finanziaria sia calcola sottraendo alla somma del valore della produzione (VP) e dei proventi da compartecipazione (PC) di un esercizio i relativi costi produzione diminuiti del valore degli ammortamenti e delle svalutazioni ($CP - AS$) sommati agli oneri da compartecipazione (OC):

$$EBITDA = (VP + PC) - (CP - AS + OC)$$

dove CP rappresenta il costo della produzione e AS gli ammortamenti e le svalutazioni (voce B.10 conto economico).

Il valore della produzione per il settore è caratterizzato da:

- ricavi delle vendite e delle prestazioni: in questa voce (A.1 del conto economico) sono contenuti i proventi derivanti dalla vendita dei biglietti e degli abbonamenti per le partite casalinghe e il 18% dell'incasso delle partite disputate in trasferta;
- altri ricavi e proventi: sono la parte più consistente dei componenti positivi di reddito delle società calcistiche. Contenuti nella voce A.5, si suddividono a loro volta in:

[25] U. Sostero, P. Ferrarese, M. Mancin, C. Marcon, *Elementi di bilancio e di analisi economico-finanziaria*, Libreria Editrice Cafoscarina, Venezia, 2011, pag. 217.

- contributi in conto di esercizio, in cui sono inseriti i contributi federali erogati nel corso della stagione sportiva;
- proventi da sponsorizzazioni, corrispettivi versati alla società dallo sponsor ufficiale, dallo sponsor tecnico, dagli sponsor istituzionali, da fornitori ufficiali e dai partner commerciali;
- proventi pubblicitari, ovvero i ricavi derivanti dalla cessione di spazi pubblicitari all'interno dello stadio nel corso delle partite casalinghe;
- proventi commerciali e *royalties*, ottenuti attraverso le attività di *merchandising* e *licensing*;
- proventi da cessione di diritti televisivi che costituiscono, come abbiamo visto precedentemente, la fonte di ricavo più remunerativa per le società calcistiche italiane;
- proventi vari, in cui sono racchiusi i proventi radiofonici, telefonici, editoriali, derivanti dallo sfruttamento dei diritti di immagine dei giocatori e dell'allenatore, concessioni varie;
- ricavi da cessione temporanea dei diritti alle prestazioni dei calciatori;
- plusvalenze da cessione dei diritti pluriennali alle prestazioni dei calciatori, che sorgono nei casi in cui la società ceda un giocatore ad un'altra prima della scadenza del contratto: se il prezzo di vendita è superiore al valore contabile attribuito al diritto pluriennale al netto degli ammortamenti, la società contabilizza una plusvalenza o, viceversa, una minusvalenza[26];
- altri proventi da gestione dei calciatori, in cui vengono messi a bilancio i premi di valorizzazione, di preparazione, l'indennità di formazione e il contributo di solidarietà;
- ricavi e proventi diversi;

I proventi e gli oneri da compartecipazione, disciplinati dall'ex art. 102 N.O.I.F. misurano la differenza tra il valore fissato per la risoluzione della compartecipazione rispetto a quello stabilito alla stipula della stessa.

I costi considerati comprendono le seguenti voci di spesa specifica:

[26] M. Valeri, *cit*, pag. 185.

- costi del personale: in costante aumento dalla sentenza Bosman, sono iscritti nella voce B.8 del conto economico. Si suddividono a loro volta in salari e stipendi, oneri sociali, trattamento di fine rapporto, trattamento di quiescenza e simili;
- costi per servizi: voce generica in cui rientrano i costi per i tesserati, per l'attività sportiva, i costi specifici tecnici, di vitto, alloggio e locomozione gare, le spese assicurative e previdenziali e quelle amministrative, pubblicitarie e generali;
- oneri diversi di gestione: la voce B.14 comprende diverse voci di spesa specifiche del settore, tra cui quelle per l'organizzazione delle gare e per l'iscrizione alle stesse, oneri specifici verso le squadre ospitate, i costi per acquisizione temporanea dei diritti alle prestazioni di calciatori e le minusvalenze da de cessione dei diritti pluriennali alle prestazioni sportive degli atleti.

1.4.3 EBIT

L'EBIT, acronimo di *Earnings Before Interests and Taxes*, è una misura del risultato operativo di un'azienda. L'indice include tutte le spese sostenute nel corso dell'esercizio tranne gli oneri finanziari e le imposte sul reddito[27]:

$$EBIT = EBITDA - S\&A$$

dove $S\&A$ sono le svalutazioni e gli ammortamenti per l'esercizio (voce B.10 del conto economico).

In termini economici, l'EBIT rappresenta la resa del capitale investito nell'impresa: per questo motivo è un utile strumento di confronto tra le società che operano nello stesso settore.

1.4.4 Utile (perdita) di esercizio

Il risultato netto d'esercizio è chiamato utile se ha un valore superiore allo zero, perdita se viceversa assume un valore negativo. Esso si ottiene sottraendo dal risultato ante imposte (EBT, acronimo inglese di *Earnings Before Taxes*) le imposte sul reddito (voce 22) conto economico):

$$Utile\,(perdita) = EBT - T$$

dove EBT è il risultato del seguente calcolo:

[27] Z. Bodie, A. Kane, e A. J. Marcus, *Essentials of Investments*, McGraw Hill Irwin, 2004, p. 452.

$$EBT = EBIT - Of + PS - OS$$

in cui Of rappresentano gli oneri finanziari, PS e OS sono rispettivamente i proventi e gli oneri derivanti dalla gestione straordinaria.

L'importanza di quest'indice è da ricercarsi nel vincolo del pareggio di bilancio imposto dal fair play finanziario UEFA: un valore positivo (o una perdita contenuta entro 5 milioni di euro) è fondamentale per le società che non vogliono essere escluse dalle competizioni continentali o incappare in pesanti sanzioni pecuniarie.

1.4.5 ROA

Il ROA, acronimo di *Return on Assets*, misura la redditività dell'attivo netto. Si calcola dividendo l'EBIT per il valore delle attività presenti nello stato patrimoniale:

$$ROA = \frac{EBIT}{Tot.\ Attività}$$

Questo indice esprime sostanzialmente il rendimento delle risorse impiegate nell'attività di impresa: esso consente, durante la valutazione, di isolare la capacità della società di far fruttare le risorse investite nella gestione caratteristica e patrimoniale, in quanto non è influenzato dagli effetti economici negativi associati alle scelte di finanziamento e alle politiche fiscali a cui è sottoposta l'impresa[28].

Il ROA è spesso valuto in riferimento al valore assunto dallo stesso riguardo un'impresa concorrente o il valore medio dell'indice nel settore. Tuttavia, un suo valore maggiore o in linea con quello del settore non significa necessariamente che si possa esprimere un giudizio positivo sulla gestione: ciò dipende se il rendimento dell'attivo netto è superiore al costo medio dei mezzi di terzi.

1.4.6 ROS

Il ROS, acronimo dell'inglese *Retun on Sales*, esprime la redditività della società in relazione alla capacità remunerativa del flusso di ricavi. Rappresenta il ricarico percentuale sulle vendite: quanto rimane sul prezzo di vendita dopo aver coperto i costi operativi.

Si calcola dividendo L'EBIT per i ricavi, in questo caso formati dalla somma del valore della produzione sommato ai proventi da compartecipazioni:

$$ROS = \frac{EBIT}{(VP + PC)}$$

Se l'indice assume valori positivi una parte di ricavi è ancora disponibile dopo la coper-

[28] U. Sostero, P. Ferrarese, M. Mancin, C. Marcon, *cit.*, pag 282.

tura di tutti i costi inerenti alla gestione caratteristica: esprime quindi la capacità dei ricavi della gestione caratteristica a contribuire alla copertura dei costi extra-caratteristici, oneri finanziari, oneri straordinari e a produrre un congruo utile quale remunerazione del capitale proprio.

Se invece il ROS si assesta su valori vicini allo zero, la capacità remunerativa del flusso di ricavi caratteristici è limitata alla sola copertura dei costi della gestione caratteristica: in tal caso caso, la copertura degli oneri finanziari, degli oneri straordinari e l'utile dipendono dalla presenza di risorse extra-caratteristiche quali proventi finanziari e proventi straordinari.

Infine, se l'indice assume di valori negativi, esso denota l'incapacità dei ricavi caratteristici a coprire anche i soli costi di gestione. Ciò è il sintomo di una gravissima crisi produttiva e gestionale.

1.4.7 EBITDA Margin

L'*EBITDA Margin* è dato dal rapporto tra l'EBITDA e i ricavi. Nel nostro caso, è dato dalla formula:

$$EBITDA\,Margin = \frac{EBITDA}{(VP + PC)}$$

Questo indice esplica la redditività operativa di una società: un rapporto crescente indica un aumento della redditività lorda delle vendite e una diminuzione dell'incidenza dei costi operativi.

Il margine sulle vendite è legato da un lato alle condizioni concorrenziali del settore nel quale l'impresa si trova a operare e dall'altro ai vantaggi competitivi specifici dell'impresa: la modifica di uno di questi fattori può determinare una sua variazione in positivo o negativo.

1.4.8 Tasso indebitamento

Il tasso di indebitamento complessivo esprime la proporzione tra il valore totale dei Mezzi di terzi e l'ammontare del Patrimonio netto dell'impresa:

$$Tasso\,di\,indebitamento = \frac{Mezzi\,di\,terzi}{PN}$$

dove PN indica il Patrimonio netto.

Esso consente di fornire una misura relativa del valore complessivo delle passività, esprimendo un giudizio sul livello di rischio finanziario[29] associato all'impresa: una maggiore

[29] possiamo definire il rischio finanziario come «il livello di probabilità che l'impresa non sia in grado di disporre dei mezzi di pagamento per onorare i propri impegni futuri man mano che giungano a

consistenza del Patrimonio netto riduce tale rischio, influenzando oltremodo le attese dei portatori d'interesse sulle potenzialità di crescita e di creazione di valore dell'impresa.

1.4.9 Indice di solidità patrimoniale

L'indice di solidità patrimoniale (chiamato anche tasso di indebitamento netto), desumibile dalla riclassificazione funzionale dello Stato patrimoniale, è determinato dal rapporto tra la posizione finanziaria netta e il patrimonio netto:

$$Indice\ solidità\ patrimoniale = \frac{PFN}{PN}$$

La posizione finanziaria netta (PFN) è formata dai prestiti a breve e a lungo termine concessi dal sistema bancario e dal mercato dei capitali, al netto delle attività finanziarie disponibili (crediti finanziari a breve e a lungo termine, titoli, partecipazioni non strategiche, disponibilità liquide)[30]. Se assume un valore negativo, si parla in tal caso di indebitamento finanziario netto, che evidenzia l'esposizione dell'azienda nei confronti dei terzi finanziatori.

L'indice di solidità patrimoniale consente di valutare la proporzione esistente tra i Mezzi di terzi, al netto delle attività finanziarie disponibili, ed il Patrimonio netto per la copertura del fabbisogno generato dal capitale investito netto.

1.5 Variabili Opta

Le variabili descritte nella sezione precedente permetteranno di analizzare la prestazione economica fornita dalle società calcistiche italiane.

Non meno importante della precedente è la performance sportiva che viene, tradizionalmente, fatta coincedere con la posizione e il numero di punti raggiunti alla fine del campionato. In Serie A, esistono delle soglie virtuali basate sul numero di punti conquistati: la «zona salvezza» (ovvero i punti necessari a non retrocedere in Serie B) si assesta intorno ai 40 punti, per la «zona Europa» (il punteggio che permette alla società di accedere alle coppe continentali) bisogna accomulare almeno 60 punti mentre per la zona «scudetto» (i punti occorrenti a vincere il titolo) è necessario spingersi almeno ad 80 punti.

Al fine di prevedere il punteggio che una squadra possa potenzialmente ottenere nell'arco di un campionato, abbiamo selezionato alcune delle più significative variabili contenute nel database Opta.

Opta raccoglie e analizza i dati degli eventi sportivi in tempo reale, per renderli disponibili pochi secondo dopo che l'azione è avvenuta in campo. Ogni dato raccolto e archiviato

scadenza», U. Sostero, P. Ferrarese, M. Mancin, C. Marcon, *cit.*, pag 306.

[30] U. Sostero, P. Ferrarese, M. Mancin, C. Marcon, *cit.*, pag 337.

va ad aggiungersi a un vastissimo archivio in cui sono presenti dati di manifestazioni sportive di tutto il globo. I dati raccolti dall'Azienda hanno un livello di dettaglio molto accurato, superando il dato statistico sportivo tradizionale: ad esempio, non forniscono solamente minuto e marcatore per ogni rete segnata, ma comunicano anche con quale parte del corpo il giocatore ha segnato, come si è sviluppata l'azione che ha portato al goal e il significato dell'evento nel contesto della partita e del campionato. Per ciascun match di serie A, sono raccolti aggiornamenti in tempo reale per tutti i tocchi palla, dati posizionali e statistiche per ciascun giocatore, per un totale di oltre 1600 eventi.

Considerata questa mole di dati, fondamentale è la scelta delle variabili da considerare all'interno del modello neurale da implementare. A questo proposito, è stato deciso di dividere gli indicatori in due macro-categorie: la prima riguarda le variabili a livello di squadra, la seconda interessa i dati caratterizzanti le performance individuali dei giocatori partecipanti all'evento sportivo.

1.5.1 Variabili di squadra

OPTA raccoglie per ogni singola giornata del campionato di serie A dati per tutte le squadre facenti parte della competizione. Questi dati vengono suddivisi ed archiviati nei database in più di 800 variabili di tipo quantitativo. Vista la notevole mole di dati, è stato necessario attuare un processo di selezione al fine di valutare quali variabili utilizzare per l'analisi. Per praticità, i dati sono stati suddivisi in 3 categorie, ovvero quelli riferiti alla fase offensiva di gioco, quelli tipici della fase difensiva e quelli che concernono il possesso della sfera e il mantenimento della stessa.

Variabili: fase offensiva

Prima di elencare e descrivere le variabili che scelte per la rete neurale, è importante chiarire cosa si intende per fase offensiva: è quella fase di gioco nella quale una squadra è in possesso di palla[31], quindi tutti gli 11 giocatori partecipano alla fase. Essa si divide inoltre in altre due fasi, le azioni manovrate e le palle inattive.

Le variabili che utilizzate nel modello sono:

- GOAL: si riferisce alle reti realizzate dalla squadra durante tutto l'arco del campionato. La sua incidenza all'interno modello previsionale è notevole: maggiore sarà il numero di marcature segnate, più elevata sarà la probabilità di ottenere una vittoria in ogni giornata disputata e, di conseguenza, punti in classifica;
- GOAL TO SHOT RATIO: questo indice rappresenta il rapporto percentuale tra i

[31] G. Maiuri, *Un modo diverso di pensare calcio: L'approccio Sistemico e la Periodizzazione Tattica*, Youcanprint Self-Publishing editore, pag. 20.

goal segnati e le conclusioni che abbiano inquadrato la porta avversaria nel corso dell'intero campionato;

- SHOOTING ACCURACY: rappresenta il grado di precisione dei tiri tentati verso la porta avversaria nel corso delle 38 giornate di serie A. È espresso dalla percentuale del rapporto tra le conclusioni inquadranti lo specchio di porta e i tiri totali;
- PENALTY SUCCESS RATE: questa variabile si riferisce alla percentuale di trasformazione dei rigori concessi dalla squadra avversaria;

Variabili: possesso della sfera e mantenimento dello stesso

In questo sott'insieme sono raggruppate tutte le azioni di gioco finalizzate al possesso palla e al suo mantenimento: sono variabili che possono incidere nel modello previsionale sia perché il possesso della sfera limita la pericolosità della formazione avversaria sia perché la costruzione di gioco aumenta le chance di creare un'occasione da rete.

Nel dettaglio, le variabili selezionate sono:

- ASSIST: rappresenta il passaggio volontario compiuto da un giocatore che permette ad un compagno di realizzare un gol senza che questi debba dribblare nessun avversario, eccezion fatta per il portiere;
- KEY PASS: questo indice è comprensivo di tutti gli assist mancati, ovvero quelli che non hanno portato alla segnatura di una rete da parte di un compagno;
- PASSING ACCURACY OPPOSITION HALF: riguarda il grado di successo dei passaggi tentati dalla squadra nella metà campo avversaria nell'arco di tutto il campionato;
- PASSING ACCURACY FINAL THIRD: come la precendente variabile, qui si valuta il rapporto percentuale tra i passaggi effettuati con successo negli ultimi 35 metri della metà campo avversaria e i passaggi totali tentati nella stessa porzione di campo;
- CROSSES & CORNERS ACCURACY: questo indice valuta la percentuale dei cross e dei calci d'angolo giunti a destinazione contro tutti i cross e i calci d'angolo calciati nel corso della gara;

- DRIBBLE SUCCESS RATE: questa variabile rappresenta la somma tutti i dribbling compiuti con successo divisa per i numero totale di dribbling tentati nel corso del campionato. Il dribbling è definito come abile palleggio diretto a liberarsi dell'avversario, eseguito mediante leggeri e rapidi spostamenti impressi alla palla ora a destra ora a sinistra[32]

Variabili: fase difensiva

Si definisce la fase difensiva come il periodo di tempo in cui l'avversario è in possesso della sfera. Obiettivo principe di questa fase è evitare di subire goal e il recupero della palla. Un'efficace fase difensiva permette alla squadra di limitare la pericolosità della formazione avversaria ed aumentare perciò le possibilità di ottenere un risultato positivo dal match.

Per questa fase, sono state selezionate le seguenti variabili:

- TACKLES SUCCESS RATE: questo indice esprime il rapporto percentuale tra le azioni di contrasto tentate per mantenere o guadagnare il possesso della sfera che hanno avuto successo su quelle totali ;
- CLEARANCES, BLOCKS & INTERCEPTIONS: questa variabile raggruppa tutte le azioni difensive diverse dai *tackle.* Le spazzate (*clearances*) sono quegli interventi difensivi che allontanano la palla dalla propria aerea di rigore, senza essere direzionati volontariamente verso un compagno. Le stoppate (*blocks*) indicano gli eventi in cui un giocatore blocca una conclusione effettuata da un avversario. Gli intercetti (*interceptions*) sono azioni di gioco volte a intercettare un passaggio tra due giocatori avversari;
- FOULS CONCEDED IN THE DANGER AREA (INC PENS): tutti i falli commessi ai danni di un avversario e fischiati dal direttore di gara avvenuti all'interno dell'ultima tre quarti campo, area di rigore inclusa;
- FOULS WON IN DANGER AREA (INC PENS): enumera i falli subiti e fischiati a favore nell'ultima tre quarti campo, area di rigore inclusa;
- PENALTIES CONCEDED: questa variabile raggruppa tutti i calci di rigore concessi dall'arbitro alla squadra avversaria;

[32] http://dizionari.repubblica.it

- RED CARDS: il numero di cartellini rossi estratti dal direttore di gara durante le partite nei confronti dei giocatori della squadra. Il cartellino rosso comporta l'espulsione del giocatore dal campo, che costringe la squadra a continuare il match in inferiorità numerica;
- GOALS CONCEDED: sono tutti le reti concesse alla formazione avversaria durante una singola partita di campionato;
- SAVES MADE: rappresenta l'insieme di azioni realizzate con successo che il portiere può compiere per evitare una marcatura avversaria, indipendentemente dalla parte del corpo utilizzata;
- SAVES TO SHOT RATIO: questa variabile indica la percentuale di parate compiute dall'estremo difensore sui tiri totali da egli fronteggiati;
- CLEAN SHEETS: in questo indice viene conteggiato il numero di partite terminate dalla squadra senza subire goal dalla formazione avversaria;
- TOTAL SHOT CONCEDED: indica il numero totale di tiri concessi alle formazioni avversarie nell'arco del campionato;
- DUELS WON %: questo indice considera la percentuale di duelli vinti dai giocatori della squadra nel corso della manifestazione. Duelli sono considerate quelle situazioni di gioco in cui due giocatori rivali si fronteggiano 1vs1 per aggiudicarsi il possesso della sfera;
- OPPONENT 2ND YELLOW: sono tutti i giocatori avversari fronteggiati ed espulsi dall'arbitro per somma di ammonizioni nel corso del campionato;
- OPPONENT REDS: analogamente alla precedente, indica gli avversari espulsi del direttore di gara per rosso diretto nel corso dei match disputati.

1.5.2 Variabili individuali

Le variabili appena descritte sono raccolte, nell'arco dei campionati considerati, anche per ogni singolo giocatore sceso in campo nel corso delle stagioni analizzate. Verranno utilizzate queste variabili per valutare l'impatto che hanno i top-player sui risultati della squadra.

A livello semantico, un top-player è un giocatore di massimo livello, in grado cioè di fornire prestazioni di valore superiore alla normalità. Data questa definizione, un top-player può esserlo in una specifica squadra ma perdere la sua qualifica cambiando casacca: la scelta di chi considerare come tale non si presta a valutazioni oggettive. Nella nostra analisi, abbiamo utilizzato due criteri per selezionare i migliori giocatori di ogni squadra della Serie A:

- OPTA Index, un indice stilato da OPTA che assegna un punteggio in base a tutte le giocate compiute dal giocatore, ponderato rispetto al numero di match giocati nell'arco del campionato;
- La valutazione economica assegnata a ciascun calciatore professionista da transfermarkt.it[33]

Un utilizzo incrociato dei due criteri descritti precedentemente ha permesso di selezionare quali siano, secondo un approccio il più possibile oggettivo, i calciatori in grado di spostare gli equilibri del campionato e aventi un costo accessibile per la squadra selezionata nella nostra analisi.

[33] `transfermarkt.it` è un portale calcistico con database mondiale per la valutazione del costo del cartellino dei calciatori.

Data Mining, Reti Neurali e Cluster Analysis

Il termine *Data Mining*[34] è basato sull'analogia delle operazioni dei minatori che "scavano" all'interno delle miniere grandi quantità di materiale di poco valore per trovare l'oro. Nel DM l'"oro" è rappresentato dall'informazione, nascosta in una moltitudine di dati (il materiale di poco valore): sono quindi necessarie tecniche di esplorazione dei dati per estrarla e renderla usufruibile. Il DM oggi è utilizzato in vari contesti, che spaziano dalla teoria dell'informazione, al calcolo numerico, all'intelligenza artificiale, alla statistica computazionale e multivariata), al calcolo probabilistico, fino alle discipline economico-aziendali. Ai giorni nostri le organizzazioni (come le aziende, i centri di ricerca, le banche, i centri di analisi statistica, etc.) hanno a disposizione sempre crescenti quantità di dati, riguardanti se stesse, l'ambiente con cui interagiscono, i competitor e tutti gli stakeholder. Conseguenza naturale è stata l'emergere dell'esigenza di discriminare le informazioni utili all'organizzazione da quelle superflue, che ha portato all'elaborazione di metodi in grado di sottoporre i dati a processi di analisi al fine di trasformarli in conoscenza utile alle aziende per supportare decisioni e intraprendere soluzioni più efficaci, veloci, economicamente sostenibili e tecnologicamente possibili. Un processo di elaborazione avanzata e di analisi alternativa dei dati è il DM, il cui scopo principale consiste nel recuperare l'informazione nascosta in database di grosse dimensioni. Il DM, dunque, si pone come processo di selezione, esplorazione e modellazione di grosse masse di dati, al fine di scoprire strutture, regolarità o relazioni non note a priori, e allo scopo di ottenere un risultato chiaro e utile. Al contrario delle tecniche statistiche tradizionali di analisi dei dati, come ad esempio la regressione, utilizzate per cercare conferma empirica a fatti ipotizzati o già conosciuti, con le tecniche di DM si cercano tra i dati informazioni ignorate a priori e che possono accrescere il bagaglio di conoscenze: i dati vengono analizzati per individuare pattern frequenti, regole di associazione, valori ricorrenti, senza alcun intervento dell'utente, al quale resta comunque il compito di valutare l'effettiva importanza delle informazioni ricavate automaticamente

[34] G. Zazzaro, *Data Mining: esplorando le miniere alla ricerca della conoscenza nascosta*, in Matematicamente.it, numero 9, maggio 2009.

dal sistema. Le tecniche di DM più frequentemente utilizzate sono il *clustering*, le reti neurali, le reti bayesiane, gli alberi di decisioni, gli algoritmi di apprendimento genetico, le analisi di associazione, schematizzate nella figura sottostante, suddivise in tecniche basate su apprendimento supervisionato e tecniche non-supervisionate.

Gran parte dei progetti di DM sono supervisionati, il loro obiettivo è quello di generare previsioni, stime, classificazioni o caratterizzazioni relativamente al comportamento di alcune variabili target già individuate in funzione di variabili di input. I metodi non-supervisionati si differenziano dai precedenti in quanto in essi non esiste una variabile target da prevedere . Noi concentreremo la nostra analisi calcistica nei metodi supervisionati, in particolare implementeremo una rete neurale, mentre utilizzeremo la *cluster analysis* per analizzare le variabili economico finanziarie descritte precedentemente.

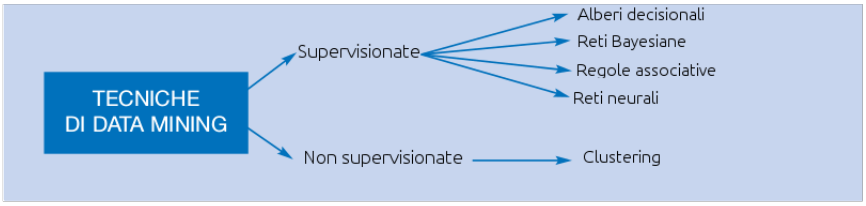

Figure 1.3: Tecniche di Data Mining

2 Reti Neurali

Le reti neurali artificiali sono «sistemi di elaborazione dell'informazione, il cui funzionamento trae ispirazione dai sistemi nervosi biologici[35]». Di quest'ultimi, rivelatisi in numerose circostanze più potenti ed efficaci dei sistemi di computazione classici, si è cercato di simularne le potenzialità riproducendo alcuni aspetti della loro architettura fisica. Queste reti possono infatti essere utilizzate per classificare problemi che sono difficilmente risolvibili dai calcolatori tradizionali[36], in quanto differiscono nel modo di elaborare le informazioni:

- nei sistemi nervosi l'elaborazione avviene in parallelo e quindi distribuita su più elementi, mentre negli elaboratori tradizioni essa è sequenziale;
- flessibilità: i calcolatori devono essere programmati per svolgere un'azione, mentre i sistemi neurali imparano autonomamente in base all'esperienza o con l'aiuto di un insegnante esterno;
- robustezza e tolleranza agli errori: i sistemi neurali possono interagire con dati incompleti senza avere effetti significativi sulla performance totale del sistema.

Le reti neurali riescono a risolvere in modo soddisfacente problemi predittivi, anche quelli in cui vi sono un gran numero di input incompleti legati da funzioni non lineari e quindi difficilmente rappresentabili in equazioni strutturali. Sono quindi dei potenti strumenti di identificazione di regolarità empiriche, grazie alla loro caratteristica di non essere soggette ai vincoli che caratterizzano i processi stocastici generatori di dati. Vengono ora illustrate le tappe fondamentali del processo di sviluppo di queste reti.

[35] D. Floreano e C. Mattiussi, *Manuale sulle reti neurali*, Il Mulino, Bologna, 2004, pag. 16.

[36] I calcolatori di Von Neumann sono composti da uno o più processori che eseguono in successione centinaia di milioni di operazioni al secondo, e da tre tipi di memoria: la prima contiene le istruzioni che il calcolatore deve eseguire, una temporanea in cui vengono letti i dati e depositati i dati dei calcoli effettuati e una permanente in cui vengono registrati i risultati dell'operazione. Qualunque sia il compito da svolgere, questi calcolatori necessitano di programmi, ovvero un insieme di operazioni organizzate gerarchicamente e di tabelle di consultazione in cui sono contenute le regole della procedura analitica che il calcolatore deve seguire.

2.1 Cenni storici

I pionieri della materia furono Warren McCulloch e Walter Pitts[37]. Essi argomentarono che le funzionalità del cervello umano potevano essere simulate attraverso modelli matematici. Durante il loro lavoro presso l'Università di Chicago, dimostrarono nel 1943 che la loro rete artificiale di neuroni, seppur caratterizzata da neuroni con una soglia e una attivazione binaria, era in grado di affrontare e risolvere problemi complessi. Per quanto la rete di McCulloch e Pitts fosse semplice e grezza, essa gettò le basi dei modelli neurali moderni.

Brevemente, le caratteristiche del loro modello possono essere così riassunte:

- le computazioni sono effettuate in intervalli di tempo discreti;
- se un neurone riceve da qualche unità di input un segnale inibitorio, esso non propaga il segnale ai neuroni successivi;
- la computazione segue una legge di soglia lineare: l'unità rilascia un segnali solo quando la sommatoria dei segnali eccitatori ricevuti dalle unità precedenti supera il valore soglia fissato per quel neurone;
- non è presente un algoritmo di apprendimento dall'esperienza.

L'essenza dell'analogia tra il modello di McCulloch e Pitts e il neurone biologico stava quindi nel concetto che una *processing entity*, designata come neurone artificiale, poteva essere programmata per rilasciare un segnale quando i suoi input raggiungono un livello soglia: i dendriti che collegano tra loro i neuroni biologici potevano essere discretamente imitati attraverso delle connessioni programmate. A queste connessioni, infine, potevano essere assegnati dei pesi affinché fossero in grado di simulare efficacemente le sinapsi, così come potevano essere programmate per essere definite positive (e quindi corrispondere a una sinapsi eccitatoria) oppure negative (sinapsi inibitoria).

Alla fine degli anni '40 Donald Hebb propose il primo meccanismo di apprendimento sinaptico per le reti neurali[38]: il suo lavoro, basato soprattutto sull'osservazione del comportamento umano ed animale, era finalizzato a cercare di spiegare in modo plausibile come avviene apprendimento in una rete di neuroni. Nel corso dei suoi studi, formulò la cosiddetta "regola di Hebb" per imitare la memoria a breve termine neuronale: se

[37] W. McCulloch e W. Pitts, *A logical calculus of the ideas immanent in nervous activity,* in Bulletin of Mathematical Biophysics, vol. 5.

[38] D.O. Hebb, *The organization of behaviour*, New York, Wiley and Sons, 1949.

una coppia di neuroni viene eccitata simultaneamente, il valore sinaptico della loro connessione risulta aumentato. Questo aumento del valore della connessione fu associato all'apprendimento. Hebb aveva inoltre ipotizzato che l'aggiustamento dei pesi tra le connessioni neuronali non bastava per imitare la memoria a lungo termine, e che essa necessitasse quindi di una modifica alla struttura della rete neurale in questione.

Tra la fine degli anni Cinquanta e i primi anni Sessanta Frank Rosenblatt[39] pubblicò il suo lavoro sul percettrone, un progetto di computazione neurale riguardante un modello di riconoscimento ottico. Ogni percettrone consisteva in un singolo neurone in grado di ricevere input da una serie di recettori di luminosità, sui quali veniva proiettato un determinato pattern di input che il percettrone stesso imparava a riconoscere. La modifica dei pesi delle connessioni sinaptiche avveniva automaticamente, quando il percettrone forniva un output scorretto rispetto al pattern di input. Rosenblatt dimostrò inoltre che il percettrone era in grado di classificare le figure anche in presenza di *data noise*, in quanto la distribuzione veniva distribuita in tutta la rete neurale. Pur avendo diversi limiti, il lavoro di Rosenblatt fu fondamentale per gli sviluppi successivi poiché dimostrò che una rete neurale artificiale poteva essere addestrata a classificare input attraverso l'induzione da esempi, anziché utilizzare il tradizionale algoritmo top-down delle istruzioni programmate.

Nel 1960 Widrow sviluppò presso la Stanford University ADALINE (*adaptive linear neuron*), un sistema neurale formato da uno strato dotato di un sistema di apprendimento simile a quello del percettrone. Nello stesso anno, grazie alla collaborazione di Hoff[40], sviluppò un nuovo algoritmo per l'apprendimento, quello che oggi chiamiamo "regola Delta": alla base di esso vi è l'intuizione di modificare i valori delle connessioni sinaptiche proporzionalmente alla differenza tra il valore corretto e quello fornito dal neurone.

La curiosità e l'entusiasmo formate dagli ultimi studi furono notevolmente smorzate nel 1969 da Minsky e Papert, che nella loro opera misero a nudo tutti i limiti del percettrone[41]: la rete di Rosenblatt era in grado di risolvere solo problemi linearmente separabili, mentre era incapace di separare le classi correlate con relazioni non lineari. A riguardo, dimostrarono matematicamente che il percettrone non era in grado di risolvere il problema dello XOR. Le critiche sollevate da Minsky e Papert ebbero un effetto frenante sulla materia, che perse sia credibilità agli occhi della comunità scientifica che i

[39] F. Rosenblatt, "*The Perceptron: A Probalistic Model For Information Storage And Organization In The Brain*", Psychological Review, 1958.

[40] B. Widrow e M.E. Hoff, *Adaptive switching circuits*, in IRE WECON Convention Record, Part IV, 1960.

[41] M. Minsky e S. Papert, *Perceptrons: an introduction to computational geometry,* The MIT press, Cambridge MA, 1969.

finanziamenti governativi degli Stati Uniti.

Il periodo buio degli studi di computazione neurale si protasse fino ai primi anni Ottanta. Nel 1982, il premio Nobel per la fisica John Hopfield bocciò l'idea di monodirezionalità del percettrone, che strideva con la struttura biologica del neurone. A lui, anche grazie ai nuovi finanziamenti erogati dalla DARPA per i progetti neurocomputazionali, è stata riconosciuta l'importanza di aver riacceso l'interesse verso le reti neurali. Gli studi successivi si focalizzarono quindi nella creazioni di modelli multi-strato con funzioni di attivazione non lineare, dotate di un processo di apprendimento adeguato alla loro complessità.

Questi sforzi diedero vita al PDP (*Parallel Distribuited Processing*), un modello in grado di elaborare le informazioni in modo parallelo come avviene nel cervello, presentato da Rumelhart e McClelland nel 1986[42]. Lo stesso Rumelhart, riprendendo gli studi di Werbos[43] propose l'algoritmo di retropropagazione dell'errore, il modello di apprendimento ancora oggi più comunemente utilizzato. Questo algoritmo, che si basa sulla "regola Delta", propone un potente metodo ricorsivo per la modifica dei pesi sinaptici di una qualsiasi rete neurale, indipendentemente dal numero di strati presenti e dal numero di neuroni che compongono ogni strato.

Verso il finire degli anni Ottanta la diffusione dell'algoritmo di retropogazione dell'errore e lo sviluppo di computer sufficientemente potenti segnarono il definitivo rilancio delle reti neurali, che vennero applicate diffusamente anche in campo commerciale[44].

[42] D. E. Rumelhart e J. L. McClelland, *PDP. Microstruttura dei processi cognitivi*, Il Mulino, Bologna, 1991, trad. it di R. Luccio e M. Ricucci.

[43] P. Werbos, *Beyond regression: new tools for prediction and analysis of behavioral sciences*, tesi di dottorato, Harvard University, 1974.

[44] D. Floreano e C. Mattiussi, *cit.*, pag 29.

I era	Eventi significativi
1943	McCulloch and Pitts, formalizzazione del neurone artificiale
1949	D. Hebb e l'apprendimento per auto-organizzazione
1956	"Dartmouth Summer Research Project on AI" con (Minsky, McCarty, Rochester, Shannon)
1960	Widrow: ADALINE
1962	Il perceptron di Rosenblatt
1969	"Perceptrons", Minsky & Papert
70s	Periodo "buio": degni di nota gli associatori di Anderson, i modelli per apprendimento senza supervisione di Kohonen, gli studi di Grossberg
II era	**Eventi significativi**
1982	Reti di Hopfield: memorie associative e soluzione di problemi
1986	PDP e diffusione di Backpropagation
1987	La prima conferenza significativa dell'IEEE a San Diego (II era)
1989	I chip neurali si affacciano sul mercato: *Analog VLSI and Neural Systems*
1990	J. Pollack e le reti neurali che elaborano strutture dati
1994	Prima Conferenza Mondiale sull'Intelligenza Computazionale (Orlando)
1994	Nasce il progetto NeuroCOLT (*Computational Learning Theory*)
2001	L'IEEE approva la creazione della "*Neural Networks Society*"

Figure 2.1: Eventi significativi nell'evoluzione delle reti neurali

2.2 Il neurone biologico

Le reti neurali artificiali, come visto precedentemente, prendono ispirazione dal funzionamento dei sistemi nervosi biologici. Prima di passare ai circuiti artificiali, verranno descritti brevemente come sono strutturati questi sistemi.

Il neurone può essere considerato l'unità computazionale elementare del cervello. L'insieme di migliaia di neuroni interconnessi tra loro garantisce le funzionalità dell'encefalo: questa rete, inoltre, è in grado di variare e adattarsi in risposta delle sollecitazioni provenienti dall'esterno. Questa capacità di adattamento è riconosciuta come apprendimento ed è ciò che la comunità scientifica cerca con successo di riprodurre.

Ogni neurone è composto dal corpo cellulare, chiamato soma, che contiene il nucleo cellulare. Da esso si ramifica un gran numero di fibre, i dendriti, ed una singola libra lunga chiamata assone: i dendriti si ramificano a forma di rete attorno alla cellula e l'assone si allunga in genere di circa un centimetro. Verso l'estremità quest'ultimo si

suddivide in ramificazioni che si legano ai dendriti ed ai corpi cellulari di altri neuroni, attraverso una giunzione che prende il nome di sinapsi. Ogni neurone forma sinapsi con altri neuroni.

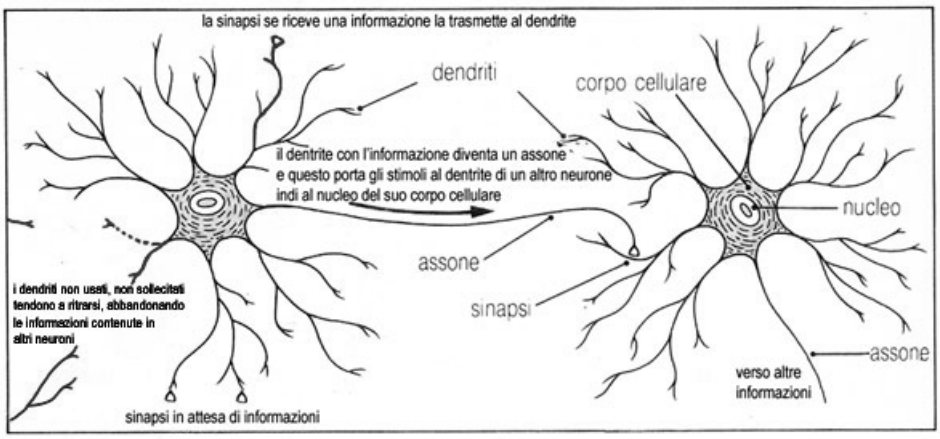

Figure 2.2: Struttura biologica di un neurone

I neuroni trasmettono un segnale elettrico lungo gli assoni. Tra il terminale di un assone e la cellula che riceverà il segnale sussiste uno spazio, la fessura sinaptica, che viene superato dai segnali grazie a sostanze chimiche dette neurotrasmettitori. Semplificando, la quantità di neurotrasmettitore rilasciato determina la conduttività della sinapsi (cioè quanto attenui o esalti il segnale elettrico proveniente dall'assone). Il neurone a valle della fessura sinaptica è dotato di recettori in grado di intercettare il neuromodulatore: si generano così correnti locali nei pressi della sinapsi, che possono sommarsi nello spazio e nel tempo in prossimità dei dendriti e del neurone. Quando la somma delle correnti che arriva alla base dell'assone si rivela superiore ad una certa soglia, viene generato un impulso (*spike*) di corrente di breve durata, da 2 a 5 millisecondi. Per uno stimolo con adeguata intensità, il neurone risponde con uno *spike* o non risponde: non c'è nessun tipo di risposta intermedia. Lo *spike* si propaga attraverso l'assone verso le sinapsi e, quando le raggiunge, queste rilasciano i neurotrasmettitori, dando il via alla ripetizione dell'intero processo per i neuroni collegati a valle.

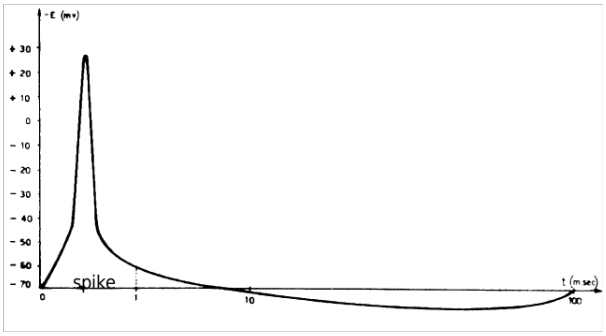

Figure 2.3: Funzione di sparo del neurone

2.3 Il neurone artificiale e le ANN

Come nei sistemi biologici appena esami, anche nelle ANN l'unità base viene chiamata neurone. Il neurone artificiale opera esattamente come quello biologico, ricevendo input dall'ambiente esterno o da altre unità neuronali che elabora e rilascia successivamente verso l'esterno. Ad ognuno di essi è associato un valore di attivazione numerico, determinato dai segnali ricevuti in ingresso dalle connessioni sinaptiche e dalla funzione di attivazione. Il segnale trasmesso come output dipende dal peso sinaptico associato alla connessione tra l'unità e quella posta più a valle. La potenza del segnale viene è anch'essa in funzione di un numero, positivo se le sinapsi sono eccitatorie, negativo in caso di sinapsi inibitorie[45].

2.3.1 Struttura del neurone artificiale[46]

Dalla struttura biologica si può quindi modellizzare i neuroni come delle funzioni non lineari che trasformano degli input $(x_1, x_2, x_3, ..., x_n)$ in un output y. Come nelle cellule biologiche a ogni input è associato un peso $(w_1, w_2, w_3, ..., w_n\ )$, positivo se la sinapsi è considerata eccitatoria, negativo nel caso opposto. Nell'equazione si è soliti aggiungere un altro parametro w_0 detto *bias*, il quale rappresenta il peso dell'input x_o che viene posto costante a 1. Viene ora definito il potenziale di attivazione del neurone, ovvero il segnale con cui il neurone trasmette la sua attività all'esterno, come la somma pesata dei vari input:

$$a_i = \sum_{i=0}^{N} w_i x_i \tag{2.1}$$

[45] D. Floreano e S. Nolfi, *Reti neurali: algoritmi di apprendimento, ambiente di apprendimento, architettura*, in Giornale Italiano di Psicologia, a. XX, febbraio 1993.
[46] D. Floreano e C. Mattiussi, *cit.*, pag. 36-37-38

Alla funzione soglia si è soliti togliere il valore della soglia ϑ_i del neurone:

$$a_i = \sum_{i=0}^{N} w_i x_i - \vartheta_i \tag{2.2}$$

La risposta del neurone y_i viene calcolata sottoponendo il potenziale di attivazione così ottenuto all'azione di una funzione di attivazione $F(a_i)$:

$$y_1 = F(a_i) = F\left(\sum_{i=0}^{N} w_i x_i - \vartheta_i\right) \tag{2.3}$$

Data la complessità di una rete neurale, in cui i diversi neuroni che la compongono ricevano una o più connessioni sinaptiche, risulta più semplice analizzare il sistema in notazione vettoriale. Poiché il potenziale di attivazione di un neurone è una funzione lineare dei segnali che riceve in ingresso, il potenziale di un intero strato di neuroni $A^T = \{A_1, A_2, ..., A_3\}$ si può riscrivere come il prodotto tra il vettore dei segnali d'ingresso $x^T = \{x_1, x_2, ..., x_n\}$ e la matrice $W = \{w_{11}, w_{12}, ..., w_{1n}, w_{21}, w_{22}, ..., w_{mn}\}$ di connessioni sinaptiche in cui le righe m corrispondono ai neuroni riceventi e le colonne n ai segnali di ingresso:

$$A = W \cdot x \tag{2.4}$$

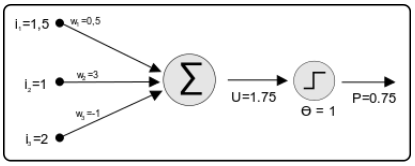

Figure 2.4: Schematizzazione di un neurone artificiale

2.3.2 Funzioni di attivazione

La funzione di attivazione descritta determina il tipo di risposta che un neurone è in grado di emettere.

Nel neurone di McCulloch e Pitts la soglia veniva mantenuta fuori dal potenziale di attivazione, facendo risultare la risposta dell'unità data come una funzione a gradino:

$$F(A) = \begin{cases} 1 & se\ A > \vartheta \\ 0 & altrimenti \end{cases} \tag{2.5}$$

dove ϑ rappresenta la soglia del neurone. Quest'ultimo, dunque, può essere solo attivo o viceversa inattivo, trasmettendo solo un bit d'informazione.

Un'informazione maggiore può essere trasmessa se si utilizza un funzione continua lineare del tipo:

$$F(A) = kA \tag{2.6}$$

in cui k è una costante. Questa funzione viene spesso limitata entro un certo intervallo (solitamente, $[0, 1]$) per contenere l'attivazione del neurone. Le funzioni di attivazione continue vengono utilizzate in quanto permettono al neurone di trasmettere segnali graduati secondo intensità, una proprietà che simula la frequenza di scarica delle unità biologiche.

L'ultima famiglia di funzioni di attivazione è formata da funzioni continue non lineari. La più utilizzata da esse è la funzione sigmoloide:

$$F(A) = \frac{1}{1 + e^{-kA}} \tag{2.7}$$

con una costante k che rappresenta il grado di inclinazione della funzione (all'aumentare di k, la funzione approssima la funzione a gradino. Le rette $y = 0$ e $y = 1$sono gli asintoti orizzontali.

Nel modello che verrà implementato, tutti i neuroni della rete - esclusi i neuroni di ingresso - utilizzano la stessa funzione di attivazione per calcolare il segnale in uscita.

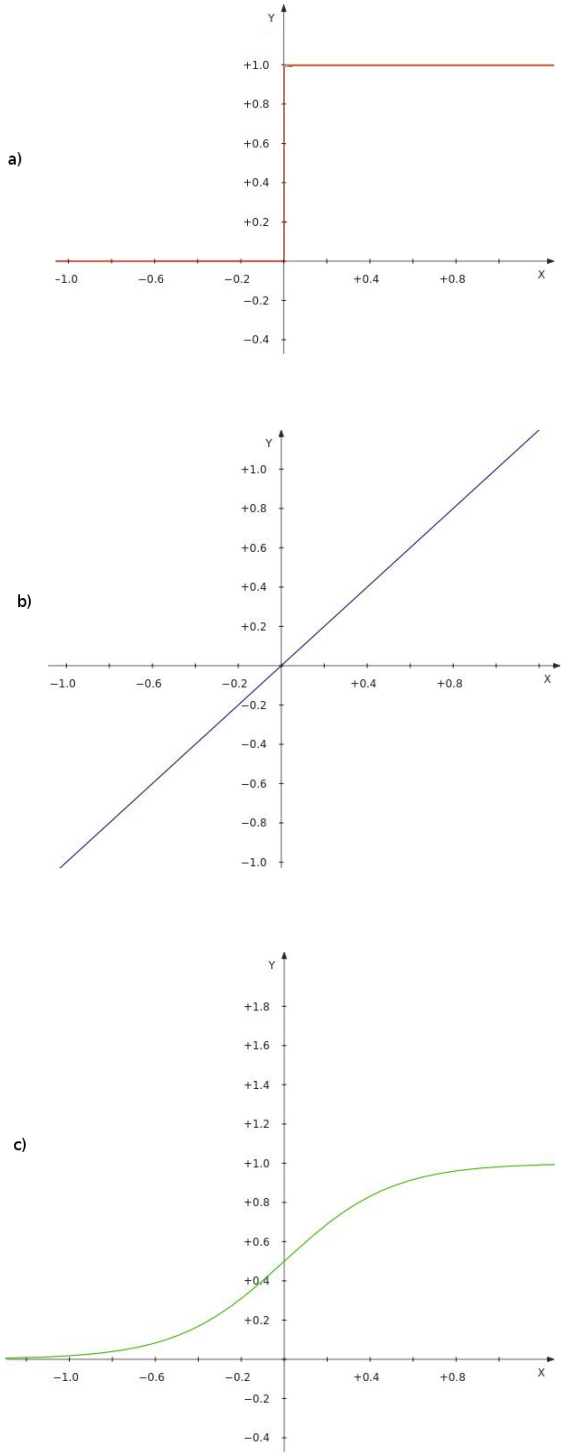

Figure 2.5: Funzioni di attivazione: a) funzione a gradino con $\vartheta = 0$; b) funzione lineare continua; c) funzione sigmoide.

2.3.3 Architetture e Modelli

L'architettura di una rete neurale è caratterizzata dalla distinzione tra neuroni di ingresso e d'uscita, da quanti strati di neuroni sono presenti nella rete e dalla presenza o meno di connessioni di retroazione.

Quanto alla prima caratteristica, possiamo distinguere reti etero-associative e reti auto-associative: nelle prime i nodi che ricevono gli input dall'ambiente esterno sono distinti da quelli che forniscono l'output della rete; le reti auto-associative possiedono viceversa uno strato unico di unità connesse reciprocamente, e sono quindi in grado di fornire risposte che variano nel tempo in presenza di input esterni costanti.

Le reti etero-associative si dividono a loro volta in: reti *feedforward* ad uno strato, reti *feedforward* multistrato, reti *feedback.* Il termine *feedforward* viene utilizzato per indicare le architetture in cui ciascun nodo della rete riceve connessioni unicamente dai nodi presenti negli strati inferiori, facendo procedere il flusso di informazioni in modo unidirezionale, dai neuroni di ingresso a quelli di uscita.

Nelle reti ad uno strato, ogni neurone di ingresso è collegato con tutti i neuroni presenti nello strato successivo, ma non sono presenti connessioni tra unità dello stesso strato. Il numero di unità presenti nella rete varia in base al risultato richiesto alla rete stessa. I neuroni di ingresso (*input layer*) non hanno funzione di elaborazione: essi trasmettono ai neuroni di uscita (*output layer*) le informazioni ottenute dall'ambiente esterno.

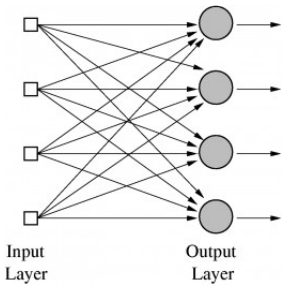

Figure 2.6: Rete neurale ad uno strato

In molte applicazioni un singolo strato di neuroni non è sufficiente alla rete per apprendere l'associazione presente tra pattern di ingresso e pattern di uscita. In questi casi si utilizzano reti multistrato, in cui vi è uno o più strati interni di neuroni chiamati strati nascosti (*hidden layer)*: essi non comunicano direttamente con l'ambiente esterno, ma producono rappresentazione interne degli stimoli che ricevono come input, agevolando il calcolo della rete. La risposta di output è ottenuta calcolando l'attivazione graduale degli strati di neuroni, procedendo dai nodi interni verso quelli di uscita.

Matematicamente, queste reti sono rappresentate attraverso matrici, i cui componenti sono costituiti dai pesi delle connessioni tra le coppie di neuroni di due strati adiacenti.

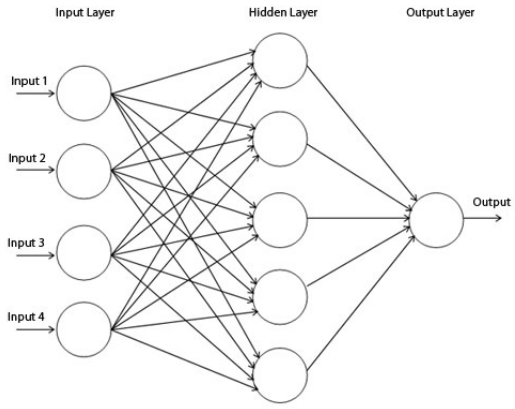

Figure 2.7: Rete neurale multistrato

I parametri di maggiore importanza in una rete neurale sono il numero di strati nascosti e di neuroni presenti in ogni strato: se la rete è strutturata in modo troppo semplice potrebbe non essere in grado di riprodurre tutte le particolarità presenti nei dati, viceversa una rete troppo complessa potrebbe riprodurre in modo soddisfacente i dati forniti, ma essere troppo specializzata e non essere in grado di riprodurne altri futuri. Il problema appena evidenziato, dovuto ad un eccesso di gradi di libertà, è molto comune nell'applicazione delle reti neurali e viene chiamato *overfitting*.

Un particolare architettura è riservata alle reti-feedback. Queste reti sono caratterizzate dalla presenza di connessioni tra neuroni dello stesso strato: ogni nodo, quindi, è in grado di ricevere stimoli sia da unità di altri strati, sia da quelle componenti il suo medesimo strato. In questi casi, si parla di nodi completamente connessi.

Il metodo di calcolo di queste reti è definito ricorrente, poiché prevede numerosi cicli di apprendimento prima di ottenere un output stabile. La numerosità dei cicli non è purtroppo definibile a priori.

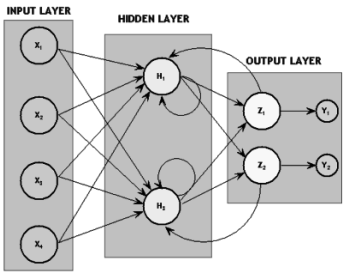

Figure 2.8: Rete neurale feed-back

2.3.4 Le regole Hebbiane

L'unità di base su cui si fondano le architetture neurali appena descritte è il neurone. Il primo a studiare le interazioni tra di essi fu Donald O. Hebb nel 1949, anno in cui formulò la Sua regola: se due neuroni collegati fra loro sono contemporaneamente attivi, l'efficacia sinaptica della connessione viene rinforzata[47]. In particolare, la sola attivazione del nodo presinaptico è sufficiente ad attivare il nodo postsinapitco: questa associazione viene ulteriormente rafforzata ogni qualvolta i due neuroni sono attivi simultaneamente.

Le reti neurali costruite con la regola di Hebb sono in grado di apprendere solo associazioni di pattern ortogonali, ovvero pattern in cui la somma dei prodotti dei singoli componenti è zero: non sono quindi in grado di imparare a conoscere associazioni che presentano elementi di input in comune ma che richiedo output diversi. Due pattern di input che è possiedono elementi in comune, infatti, causano l'attivazione di neuroni di uscita corrispondenti a entrambi i pattern, generando una risposta mista. La produzione di risposte miste dovute alla sovrapposizione di pattern di ingresso viene definita interferenza.

Alcune limitazioni furono superate dalla regola postsinaptica, tratta dagli studi di Stent[48] e Singer[49] applicati ai neuroni biologici. Questa regola prevede che il valore della connessione sinaptica sia incrementato in presenza dell'attivazione simultanea dell'unità presinaptica e di quella postsinaptica, ma che venga ridotto nei casi in cui il nodo postsinaptico sia attivo e quello presinaptico risulti inattivo. La regola postsinaptica ha permesso la parziale riduzione del fenomeno di interferenza presente nelle reti hebbiane.

La situazione simmetricamente opposta è prevista dalla regola presinaptica, in cui il valore del peso sinaptico è aumentato nel caso in cui i nodi siano attivi simultaneamente,

[47] D. Floreano e C. Mattiussi, *cit.*, pag.60-63

[48] G. S. Stent, *A physiological machanism for Hebb's postulate of learning*, in Proceedings of the National Academy of Sciences, vol.70.

[49] W. Singer, *Activity-Dependant self organisation of synaptic connections as a substrate of Learning*, in The neural and molecular bases of learning, a cura di J.P. Changeaux e M. Konishi, London, Wiley

mentre si decrementa ogni qualvolta l'unità presinaptica è attiva e quella postsinaptica inattiva. Essa è migliore della regola postsinaptica nelle reti dove molti pattern di input diversi e sovrapposti devono essere associati allo stesso pattern di input.

La combinazione delle precedenti regole viene definita regola della covarianza. Quest'ultima, introdotta inizialmente da Hopfield, prevede che la connessione sia rafforzata quando le unità siano nello stesso stato (attive o inattive), viceversa indebolita nei casi in cui i nodi siano in stati diversi. La regola della covarianza è in grado di svolgere compiti di associazione e classificazione complessi.

Unità presinaptica	+	+	-	-
Unità postsinaptica	+	-	+	-
Hebb	+			
Postsinaptica	+		-	
Presinaptica	+	-		
Covarianza	+	-	-	+

Table 2.1: Funzionamento delle regole hebbiane: variazione dei pesi sinaptici in funzione dell'attività pre e post sinaptica.

2.3.5 Paradigmi di apprendimento

Nei sistemi nervosi biologici l'apprendimento è graduale e stimolato dall'esperienza. Allo stesso modo, le reti neurali artificiali sono in grado di apprendere modificando gradualmente i valori sinaptici mediante la presentazione iterata di una serie di esempi. Le informazioni acquisite vengono memorizzate nei pesi presenti tra le connessioni sinaptiche, quindi l'apprendimento si ottiene grazie alla definizione precisa dei valori da associare ai pesi tra i diversi nodi della rete neurale. Affinché il processo vada a buon fine, è necessario che le forme già apprese vengano conservate man mano che la rete ne apprende di nuove.

Tradizionalmente, si classificano due tipologie di apprendimento:

1. *apprendimento supervisionato*, in cui la modifica dei valori sinaptici avviene mediante la misura della differenza tra la risposta fornita dalla rete neurale e la risposta desiderata per ogni vettore di input. Questo paradigma include anche gli algoritmi definiti apprendimento per rinforzo (*reinforcement learning*), i quali richiedono una misura della bontà della risposta della rete anziché la specificazione della risposta esatta per ogni pattern d'addestramento.

2. *apprendimento non supervisionato*: non viene imposta alla rete una risposta esatta

esterna, ma le vengono date alcune semplici regole di elasticità sinaptica. Grazie ad esse, la rete è in grado di auto-organizzarsi gradualmente durante la fase di codifica dei pattern di input.

Qualunque sia il paradigma di apprendimento, esistono alcuni elementi generali comuni:

- i valori iniziali dei pesi sinaptici sono assegnati in modo casuale entro un piccolo e determinato intervallo di valori oppure sono posti uguali a zero;
- l'apprendimento consiste nella presentazione ripetuta di una serie di vettori chiamati pattern di addestramento;
- la velocità di apprendimento è regolata da una costante η chiamata tasso di apprendimento, che può assumere valori compresi tra zero e uno.
- completata la fase di apprendimento, i valori sinaptici vengono registrati per studiare la risposta della rete su vettori di test: questi nuovi pattern sono presentati alla rete ingresso, che procede nel calcolo di attivazione delle unità senza modificarne i pesi sinaptici, affinché sia possibile studiare le capacità della stessa di generalizzazione a nuovi stimoli.

Verranno ora visionati nel dettaglio i diversi paradigmi di apprendimento.

Apprendimento supervisionato

Nell'apprendimento supervisionato ogni pattern è composto da una coppia di informazioni, il vettore di ingresso e il vettore contenente la risposta desiderata. La rete viene quindi addestrata presentando di volta in volta sia l'input che l'output ricercato.

Inizialmente, ai pesi sinaptici sono associati a valori casuali. Nella successiva fase di addestramento si presenta alla rete il training-set, e per ogni coppia di input-output si calcola l'errore, ovvero la differenza tra il risultato desiderato e la risposta effettiva della rete. In base all'errore, i pesi vengono infine modificati per ottenere il risultato corretto. Questo processo viene quindi ripetuto in modo ciclico, finché l'errore non raggiunge un punto di minimo locale.

Successivamente, la rete viene collaudata attraverso un test-set: viene fornito alla rete un diverso insieme di input, con il fine di valutare le sue capacità di generalizzazione.

Se il processo di addestramento è andato a buon fine, la rete sarà in grado di generare risultati anche con input di ingresso processati per la prima volta.

Le reti che utilizzano questo paradigma di apprendimento usufruiscono di una particolare regola per stabilire i pesi tra le connessioni dei nodi, la cosiddetta «regola delta». Questa regola, conosciuta anche come «regola di Widrow-Hoff» dal nome degli autori che la scoprirono[50], misura la distanza tra la risposta fornita dalla rete e quella desiderata.

Consideriamo il vettore $x = (x_1, x_2, ..., x_i)$l'input fornito al neurone, t l'output desiderato e y la risposta d'uscita del neurone. L'errore δ sarà quindi:

$$\delta = t - y \tag{2.8}$$

Il segnale di errore δ attua un meccanismo di controllo con l'obiettivo di applicare una sequenza di aggiustamenti ai pesi sinaptici del neurone al fine di avvicinare la risposta ottenuta a quella desiderata. Secondo la «regola delta», la variazione del peso sinaptico generico Δw_i è data dall'espressione:

$$\Delta w_i = \eta \delta x_i \tag{2.9}$$

in cui η è definito come il tasso di apprendimento (*learning rate*), che assume valori compresi tra zero e uno. Esso rappresenta la velocità con cui il nodo neurale è in grado di apprendere. La scelta del valore da associare al tasso deve essere oculata, in quanto influenza la convergenza e la stabilità del processo di apprendimento.

Il funzionamento di questa regola consiste nella minimizzazione dei pesi sinaptici che hanno contribuito alla formazione dell'errore, attraverso una serie ripetuta di presentazione dei pattern di addestramento. L'apprendimento può avvenire per cicli o per epoche: nel primo caso la modifica viene calcolata e addizionata ai pesi sinaptici per ogni coppia di addestramento; nel secondo caso tutte le coppie sono presentate alla rete, le modifiche vengono addizionate progressivamente e la somma totale così ottenuta addizionata ai pesi alla fine dell'epoca[51].

Con questo paradigma, i pesi assumono questo valore:

$$w_i = w_i + \Delta w_i \tag{2.10}$$

Caso particolare dell'apprendimento supervisionato è il paradigma di *apprendimento per rinforzo*, in cui i pesi vengono mutuati in base a un'informazione di tipo qualitativo

[50] B. Widrow e M.E. Hoff, *Adaptive switching circuits*, in IRE WESCON Convention Record, Part IV, 1960.

[51] D. Floreano e C. Mattiussi, *cit.*, pag. 74-78.

che indica la bontà della risposta fornita dalla rete. Esso è una forma di apprendimento supervisionato in quanto è presente un segnale proveniente dall'ambiente, ma questo segnale non individua la risposta corretta, valuta solamente la risposta del sistema. L'aggiustamento dei pesi avviene mediante meccanismi di rinforzo: l'apprendimento consiste quindi nel modificare i parametri interni della rete in funzione di aumentare la probabilità di emettere risposte che ricevono rinforzi positivi e diminuire quelle che ricevono rinforzi negativi.

Semplificando, nel caso di risposte considerate corrette l'algoritmo aumenta i pesi dei nodi attivi, rinforzandone l'azione prodotta. Viceversa, risposte scorrette causano una diminuzione dei pesi e un indebolimento di potenza dell'azione prodotta. L'apprendimento viene guidato da un indice di valutazione della risposta della rete neurale.

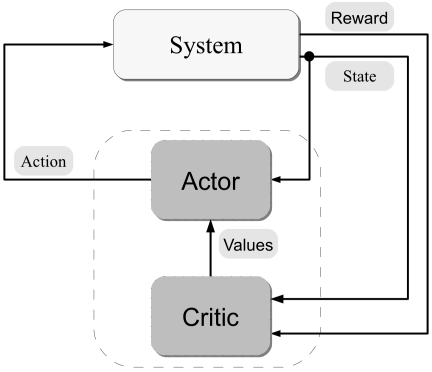

Figure 2.9: Apprendimento con rinforzo

Apprendimento non supervisionato

I sistemi nervosi biologici sono in grado di apprendere senza la guida di un insegnante esterno: sono in grado di auto-organizzarsi in base alle caratteristiche dei segnali di input. I paradigmi di apprendimento non supervisionato hanno due proprietà generali:

1. Le regole di apprendimento e le dinamiche interne della rete tendono ad essere più semplici di quelle presenti nei paradigmi supervisionati: le regole di apprendimento sono di tipo hebbiano, le funzioni di attivazione delle unità sono lineari o a gradino e la rete è composta da un solo strato di sinapsi.

2. L'apprendimento consiste nell'estrazione delle informazione che descrive nel modo migliore i pattern di input: quest'ultima è rappresentativa di caratteristiche in comune tra i pattern che permettono al modello di elaborare classificazioni senza

l'ausilio di un supervisore esterno. I pattern di ingresso devo essere ridondanti affinché l'algoritmo sia in grado di riconoscere la struttura rilevante al di sotto di essi.

Questo tipo di algoritmi di apprendimento sono utilizzati in problematiche di ricerca di eventuali strutture nascoste in dati non classificati: non conoscendo a priori i dati da analizzare, infatti, non è possibile basarsi su una funzione di errore per valutare una potenziale soluzione, ma si procede confrontando i pattern di input e valutando le caratteristiche distintive o comuni tra di essi. Ciò permette al modello di eseguire classificazioni o rappresentazioni compatte senza l'ausilio di una guida esterna.

Apprendimento competitivo

Un particolare tipo di apprendimento non supervisionato è chiamato apprendimento competitivo. Attraverso di esso, la rete impara a riconoscere le regolarità statistiche della sequenza di pattern che le vengono rappresentate.

L'architettura della rete è generalmente a strati, con connessioni forward eccitatorie tra nodi facenti parte di strati diversi e con connessione feedback inibitorie tra i neuroni adiacenti appartenenti al medesimo strato. Le connessioni inibitorie laterali permettono ai neuroni di entrare in competizione tra loro per il diritto a rispondere ad un dato sottoinsieme di input cosicché un solo neurone o uno solo per gruppo è attivo in un certo momento. Il neurone che emette il segnale di uscita non nullo è chiamato *winner-takes-all*.

In questo modo i neuroni tendono a specializzarsi su un insieme di input simili divenendo riconoscitori di caratteristiche per differenti classi di input.

Il neurone che vince la competizione è quello con input netto v_k più alto per un dato input x. Il segnale di output y_k del neurone vittorioso è settato a 1, mentre quello degli altri viene settato a 0:

$$y_k = \begin{cases} 1 & se\, v_k > v_j,\ \forall\, j \neq k \\ 0 & altrimenti \end{cases} \tag{2.11}$$

dove v_k è la combinazione lineare di tutti gli input forward e feedback.

Il neurone che vince la competizione apprende spostando i pesi sinaptici dagli input inattivi agli input attivi.

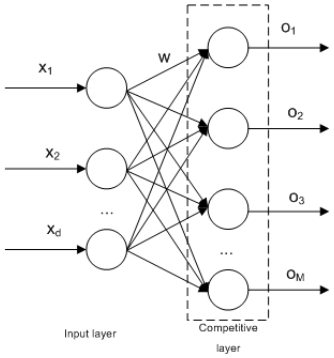

Figure 2.10: Struttura di una semplice rete di apprendimento competitivo.

2.3.6 Algoritmi di apprendimento: la retropropagazione dell'errore

La scelta dell'architettura della rete e del paradigma di apprendimento condiziona in modo determinante la capacità di apprendimento della rete. Ad oggi, esistono numerosi algoritmi di apprendimento, ognuno con specifiche caratteristiche, che consentono loro di adattarsi sia alle diverse architetture, sia alle differenti tipologie di pattern di input e variabili. Ognuno di questi algoritmi ha comunque lo stesso obiettivo: trovare la combinazione di pesi sinaptici in grado di approssimare nel modo migliore la risposta della rete all'output desiderato.

Sarà qui approfondito l'algoritmo di retropropagazione dell'errore, che verrà utilizzato nell'analisi pratica. Questo algoritmo è stato proposto per la prima volta da Rumelhart, Hinton e Williams nel 1986, in risposta al problema posto da Minsky e Papert di determinare i pesi sinaptici delle unità nascoste delle reti feed-forward multistrato. Esso prevede di calcolare l'errore commesso da un neurone dell'ultimo strato nascosto propagando all'indietro l'errore calcolato sui nodi di uscita collegati a tale neurone. Lo stesso procedimento è poi ripetuto per tutti i neuroni del penultimo strato nascosto, e così via.

Il principio generale di funzionamento dell'algoritmo è la discesa del gradiente dell'errore tramite la modifica dei singoli pesi sinaptici. La funzione di attivazione dei nodi è in genere quella sigmoidale, mentre i pesi sono inizializzati con valori casuali.

L'algoritmo utilizza una versione generalizzata della «regola delta», in grado di calcolare i pesi anche delle unità appartenenti agli strati nascosti: attraverso un metodo ricorsivo, l'errore dei neuroni nascosti è considerato l'errore dello strato in uscita, viene quindi retropropagato dalle connessioni sinaptiche e moltiplicato per la derivata prima della funzione di output di ogni nodo nascosto. La retropropagazione assicura la diminuzione dell'errore complessivo durante ogni ciclo di apprendimento, ma non è certo il raggiungimento del minimo globale a causa di diversi fattori di disturbo, quali la non lin-

earità delle funzioni, la complessità dello spazio multidimensionale dei pesi e la presenza di punti di minimo locali.

Si possono identificare due fasi all'interno dell'algoritmo: la prima prevede che, presentati gli input(in genere il training set), i segnali viaggino dall'ingresso verso l'uscita al fine di calcolare la risposta della rete (*forward phase*); è poi presente una seconda fase durante la quale i segnali di errore vengono propagati all'indietro, sulle stesse connessioni su cui nella prima fase hanno viaggiato gli ingressi, ma in senso contrario, dall'uscita verso l'ingresso (*backward phase*). Durante questa seconda fase vengono modificati i pesi intermedi. Esistono due modalità di applicazione dell'algoritmo di backpropagation: nella modalità *batch* i pesi sono aggiornati dopo aver presentato alla rete tutti gli esempi del training set, nella modalità *on-line* (o incrementale) i pesi sono aggiornati dopo la presentazione di ogni esempio.

2.4 Campi di applicazione delle ANN

I campi di applicazione delle reti neurali sono diversi ed eterogenei: trovano applicazione nell'informatica, in particolare nei processi di compressione e decompressione dei dati; possono essere utilizzate su dati finanziari, per determinare il prezzo, il rischio o la volatilità di titoli; offrono un efficace supporto nella valutazione del merito creditizio dei clienti degli istituti di credito e nel calcolo del rating dei titoli di credito.

Grazie alle loro particolari caratteristiche, possono fornire un valido appoggio nei reparti di produzione e controllo qualità nelle aziende manifatturiere, per prevenire la difettosità dei prodotti.

Numerose sono le applicazione in campo medico, in particolare nei reparti di diagnostica per classificare il paziente in base ai sintomi accusati; nel reparto di ricerca e sviluppo, nella sezione di ricerca genetica e studio della struttura del DNA; in medicina generale.

Altre applicazioni riguardano il riconoscimento facciale, di immagini, di testi scritti su supporto cartaceo e la previsione di fenomeni complessi, come ad esempio le previsioni meteorologiche, applicazioni ingegneristiche e biologiche.

Un campo in cui le reti neurali stanno prendendo piede negli ultimi anni riguarda il mondo del *betting*, per predire il risultato finale degli incontri sportivi selezionati.

Anche Siri, il sistema di riconoscimento vocale dell'IPhone di Apple, è basato su una rete neurale, così come Akinator, un'applicazione per smartphone in grado di indovinare la persona (reale o immaginaria) a cui si sta pensando.

3 Cluster Analysis

La classificazione può essere definita come la separazione e denominazione delle apparenze[52]. L'attività umana è stata da sempre caratterizzata da tecniche di classificazione: nella Bibbia, il primo compito affidato da Dio ad Abramo è di trovare un nome a «tutti gli animali dei campi e agli uccelli del cielo» che popolano il giardino dell'Eden[53]. Nel corso della storia umana, l'opera di classificazione è stata una costante per tutte le civiltà che si sono succedute nel corso dei secoli: si pensi ai trattati, alle enciclopedie, ai cataloghi, alle collezioni. In questi raccoglitori di conoscenze si possono individuare molti esempi significativi di classificazioni e queste constatazioni mostrano che esse svolgono ruoli di grande importanza nella organizzazione e nella gestione della conoscenza. Nel tentativo di semplificare la complessità del mondo, l'uomo tende ad organizzare ogni aspetto dell'ambiente che lo circonda, dividendo gli oggetti o gli eventi in categorie basate sui loro usi e sulle loro caratteristiche.

In statistica, si definisce classificazione quell'attività che, mediante un algoritmo statistico, individua una rappresentazione di alcune caratteristiche di una entità da classificare (oggetto o nozione) e le associa ad una etichetta classificatoria[54].

Come descritto nell'introduzione del precedente capitolo, tra i metodi statistici di classificazione si distingue tra metodi supervisionati e metodi non supervisionati.

L'analisi di raggruppamento (in inglese *clustering* o *cluster analysis*) è una tecnica di *data mining*, facente parte dei metodi non supervisionati, in cui non è presente una variabile target da prevedere. Questo tipo di analisi consiste nella suddivisione di un set di dati complesso in una serie di subset più semplici, chiamati appunto *cluster*, ovvero di individuare delle regioni distinte nelle quali raggruppare dati in modo tale che i dati all'interno di una stessa regione siano simili in base ad un criterio scelto inizialmente, mentre quelli in regioni differenti siano dissimili tra loro. In base alla realizzazione di questi raggruppamenti, si è in grado di dar loro, grazie ad una o più tecniche di valutazione, un significato preciso.

[52] K. Jajuga, A. Sokolowski e H.H Bock, *Classification, clustering and data analysis: recent advances and applications*, Springer, Berlin, 2002.
[53] *La Bibbia*, Genesi 2,4-24
[54] `it.wikipedia.com`

3.1 Cenni storici

Le prime forme di classificazione basate su procedure matematiche per organizzare gli oggetti in base alle similarità osservate furono introdotte da Tryon nel 1939[55] e da Cattel nel 1944[56]. La *cluster analysis* è poi emersa nella comunità scientifica come topic rilevante durante gli anni Sessanta e Settanta dello scorso millennio. I pioneri a motivare la ricerca a livello globale sui metodi di clustering furono Sokal e Sneath nel 1963, all'interno del loro scritto *Principles of numerical taxonomy*[57]. Al loro lavoro, seguirono molte altre pubblicazioni[58], che contribuirono a definire i problemi di base e le principali metodologie di analisi di raggruppamento, oltre a diffondere tale tecnica di analisi nella comunità scientifica.

3.2 Concetti chiave

La *cluster analysis* è una tecnica statistica multivariata di tipo esplorativo, capace di scomporre una realtà complessa di osservazioni in tipologie specifiche. L'obiettivo che si pone questa analisi è sostanzialmente quello di riunire unità tra loro eterogenee in più sottoinsiemi tendenzialmente omogenei e mutuamente esaustivi: semplificando, queste unità vengono suddivise in un certo numero di gruppi a seconda del loro livello di "similarità", valutata a partire dai valori che una serie di variabili prescelte assume in ciascuna unità.

Questo tipo di analisi perviene ai seguenti risultati[59]:

- la generazione di ipotesi di ricerca, infatti per effettuare una analisi di raggruppamento non è necessario avere in mente alcun modello interpretativo;
- la riduzione dei dati in forma tale da rendere facile la lettura delle informazioni rilevate e parsimoniosa la presentazione dei risultati;
- ricerca tipologica per individuare gruppi di unità statistiche con caratteristiche distintive che facciano risaltare la fisionomia del sistema osservato;
- la costruzioni di sistemi di classificazione automatica[60];

[55] R. Tryon, *Cluster analysis*, McGraw Hill, New York, 1939.
[56] R.B. Cattel, *A note on correlation clusters and cluster search methods*, in Psychometrica vol 9, pag 169-184.
[57] R.R. Sokal e P.H. Sneath, *Principles of numerical taxonomy*, Freeman, San Francisco - London, 1963.
[58] Per maggiori approfondimenti, si veda H.H. Bock, *Origins and extensions of the k-means algorithm in cluster analysis*, Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany.
[59] L. Fabbris, *Analisi esplorativa di dati multidimensionali*, Cleup editore, 1983.
[60] N. Jardine e R. Sibson, *Mathematical taxonomy*, Wiley,London, 1971.

- la ricerca di classi omogenee, dentro le quali si può supporre che i membri siano mutuamente surrogabili[61].

Riprendendo il primo punto, si nota che il clustering, a differenza delle altri tipi di analisi multivariata, non compie nessun tipo di assunzione "a priori" sulle tipologie fondamentali esistenti che possono caratterizzano il collettivo studiato: questa tecnica ha un ruolo di ricerca e identificazione di strutture latenti, allo scopo di classificazione delle partizioni.

3.3 Fasi di applicazione

Le tecniche di clustering si compongono di diverse fasi di applicazione:

1. Una prima fase riguarda la definizione e scelta delle variabili di classificazione: questa fase richiede un alto grado di soggettività, in quanto le variabili sono scelte dal ricercatore in base alle sue convinzioni personali o alle sue intuizioni riguardo alla struttura dati in questione. Le variabili quantitative utilizzate nell'analisi debbono essere espresse nella stessa unità di misura: se sono espresse in unità di misura diverse o hanno ordini di grandezza diversi, è opportuno procedere a una standardizzazione delle stesse;

2. Il secondo step è la definizione della misura di dissimilarità esistente fra le unità statistiche. I caratteri rilevati possono essere espressi in quattro scale di misura distinte: nominali, ordinali, per intervalli e per rapporti. I dati qualitativi possono essere misurati solo con riferimento alle prime due, mentre le variabili quantitative ammettono tutte le quattro scale. Nel caso di quest'ultimi possono essere utilizzati vari tipi di indici di distanza[62]:

    - la distanza euclidea, che corrisponde al concetto geometrico di distanza nello spazio multidimensionale:

$$d_{hk} = \sqrt{\sum_{v=1}^{p} w_v \left(x_{hv} - x_{kv}\right)^2} \tag{3.1}$$

[61] P.E. Green, R.E. Frank., P.J. Robinson, *Cluster Analysis in text market selection*, Management science, 1967.
[62] J.A. Hartigan, *Clustering Algorithms*, Wiley, 1975.

dove x_{hv},x_{kv} sono rispettivamente le coordinate geometriche dei punti P_he P_k e w_v il peso da associare alla variabile;

- il quadrato della distanza euclidea, se si desidera dare un peso maggiore agli oggetti che stanno ad una certa distanza;
- la distanza assoluta, che rappresenta la distanza media tra le dimensioni:

$$d_{hk} = \sum \mid x_{hv} - x_{kv} \mid w_v \tag{3.2}$$

  questa distanza è consigliata quando le variabili di classificazione sono in scala ordinale;
- la distanza di Chebychev, che può essere appropriata in casi in cui si voglia definire due oggetti come "differenti" se essi sono diversi in ciascuna delle dimensioni:

$$d_{hk} = max \mid x_{hv} - x_{kv} \mid \tag{3.3}$$

- la distanza di Mahalanobis, che tiene conto anche delle interdipendenze esistenti tra le variabili utilizzate attraverso un ridimensionando del peso delle variabili portatrici di informazioni eccedenti, in quanto già fornite da altre. Quando le variabili originarie sono correlate tra loro è improprio utilizzare la distanza euclidea, mentre è pertinente l'uso della statistica proposta da Mahalanobis data dalla forma quadratica:

$$D^2_{hk} = (\chi_h - \chi_k)' \, W^{-1} \, (\chi_h - \chi_k) \qquad con\ h \neq k = 1, ..., n \tag{3.4}$$

  dove χ_h e $\chi_k$sono i vettori con le osservazioni su he k e W la matrice di varianze-covarianze tra le variabili osservate. Nell'uso di questo tipo di distanza è necessario usare cautela per la possibile presenza di collinearità o quasi-collinearità tra le variabili: in questi casi, la matrice non è invertibile o comunque si producono errori di misura o di calcolo che determinano notevoli distorsioni sui risultati dell'analisi.

3. Dopo aver scelto la misura di dissomiglianza, è necessario procedere al raggruppamento delle unità osservate. Per fare ciò, si è soliti distinguere in due categorie di algoritmi di raggruppamento:

a) metodi gerarchici che conducono ad un insieme di gruppi ordinabili secondo livelli crescenti, con un numero di gruppi da n ad 1;

b) metodi non gerarchici. forniscono un'unica partizione delle n unità in g gruppi, e g deve essere specificato a priori.

4. Successivamente, si valuta sia la partizione ottenuta nella fase precedente sia la scelta presa sul numero ottimale dei gruppi;

5. Infine, si interpretano i risultati ottenuti attraverso l'analisi.

3.4 Algoritmi di raggruppamento

3.4.1 Metodi gerarchici

I metodi gerarchici producono una struttura gerarchica dei dati e della loro organizzazione in gruppi, attraverso l'associazione di una struttura ad albero binario all'insieme di punti: le foglie dell'albero rappresentano le singole n unità, mentre i nodi corrispondono ai sottoinsiemi dei punti; è presente, per le caratteristiche intrinseche di un albero binario, una gerarchia nei sottoinsiemi associati ai rami.

I metodi gerarchici possono essere agglomerativi o divisivi. Descriviamo nel dettaglio i metodi agglomerativi, i più largamente utilizzati. Questi metodi gerarchici aggregano in successione gruppi con bassa dissimilarità, partendo da una situazione iniziale in cui $K = n$, ovvero in cui ogni unità forma un gruppo a sé stante. Le agglomerazioni vengono reiterate fino alla situazione finale $K = 1$, in cui tutti gli individui appartengono ad un unico gruppo. L'albero binario che rappresenta la struttura gerarchica così formata è chiamato dendrogramma, nel quale sull'asse delle ordinate viene riportato il livello di distanza, mentre sull'asse delle ascisse vengono riportate le singole unità. Ogni ramo del diagramma corrisponde ad un grappolo. La linea di congiunzione di due o più rami individua il livello di distanza al quale i grappoli si fondono.

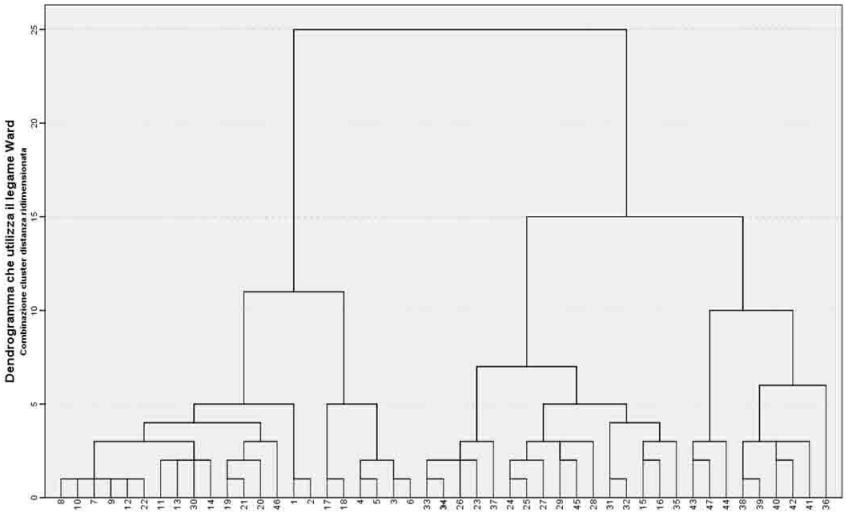

Figure 3.1: Esempio di dendrogramma

Questi metodi di aggregazione necessitano di una misura di dissimilarità tra i diversi gruppi per essere operativi; i differenti algoritmi gerarchici agglomerativi sono determinati a seconda del criterio utilizzato:

1. Metodo del legame singolo. Presi due gruppi G e G', la loro distanza è definita come il minimo tra tutte le $n_1 n_2$ distanze che si possono calcolare tra ciascuna unità i di G e ciascuna unità j di G' :

$$d_S(G, G') = min \quad d(i, j) \qquad \forall\, i \in G, \forall\, j \in G' \tag{3.5}$$

   Questo algoritmo privilegia la differenza tra i gruppi piuttosto che l'omogeneità degli elementi di ogni gruppo, valorizzando le similarità tra gli elementi;

2. Metodo del legame completo. Si considera la maggiore delle distanze istituibili a due a due tra tutti gli elementi dei due gruppi:

$$d_C(G, G') = max \quad d(i, j) \qquad \forall\, i \in G, \forall\, j \in G'$$

   si uniscono i gruppi aventi la distanza così definita minore. Questo algoritmo privilegia l'omogeneità degli elementi di ciascun gruppo a scapito della differenziazione netta tra essi;

3. Metodo del legame medio. Si calcola del valore medio aritmetico di tutte le distanze

tra gli elementi:

$$d_M(G, G') = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=2}^{n_2} d(i, j) \qquad \forall\, i \in G, \forall\, j \in G' \tag{3.6}$$

si uniscono i due gruppi che presentano la più piccola distanza così definita. Essendo basato sulla media delle distanze, i risultati sono più attendibili e i gruppi risultano più omogenei e ben differenziati tra di loro.

4. Metodo del centroide. Si determinano i vettori contenenti i valori medi delle p variabili in tutti gruppi (i cosiddetti centroidi), e le distanze tra i gruppi viene assunta pari alla distanza tra i rispettivi centroidi. Se $\overline{X}_G$ e $\overline{X}_{G'}$ sono i centroidi avremo:

$$d_X(G, G') = d(\overline{X}_G, \overline{X}_{G'}) \tag{3.7}$$

questo metodo è utilizzato in presenza di *outliers*[63] nel campione, in quanto non ne viene influenzato.

A seconda della misura adottata, cambiano quindi i raggruppamenti formati.

All'interno dei dendrogrammi è possibile incorporare l'informazione relativa al grado di dissimilarità tra i diversi gruppi, rendendo proporzionale l'altezza del segmento verticale che collega due ramificazioni successive alla caduta di dissimilarità ottenuta passando da K a $K + 1$ gruppi. Questa peculiarità aiuta nella scelta del numero di gruppi nei casi in cui non sia noto a priori, causa la mancanza di regole oggettive e univoche per la determinazione della numerosità dei gruppi.

L'approccio contrario rispetto ai metodi agglomerativi è costituito dai metodi divisivi: quest'ultimi seguono una logica analoga a quella appena descritta, ma rovesciata: vi è un gruppo iniziale contenente tutte le unità, per poi procedere secondo suddivisioni successive. Tali metodi si basano sulla partizione di un insieme in due sottoinsiemi, e sulla suddivisione delle classi precedentemente ottenute, sempre e soltanto in ulteriori bipartizioni. La divisione di un gruppo viene valutata (attraverso gli stessi algoritmi utilizzati per i metodi agglomerativi) secondo il grado di dissimilarità tra i possibili sottogruppi che si possono formare dalla sua scissione.

La caratteristica principale che distingue gli algoritmi di tipo gerarchico dai metodi non gerarchici è l'irrevocabilità nell'assegnazione di un oggetto ad un cluster: una volta

[63] Outlier è il termine utilizzato in statistica per definire, in un insieme di osservazioni, un valore anomalo e aberrante, ovvero distante dalle altre osservazioni disponibili.

che un'unità entra a far parte di un cluster, non può essere successivamente rimossa e assegnata ad un gruppo diverso.

3.4.2 Metodi non gerarchici

I metodi non gerarchici hanno la peculiarità di ripartire direttamente le n unità in g gruppi, fornendo come prodotto finale una sola partizione delle n unità[64]. Il primo passo è definire a priori il numero dei gruppi in cui si vuole ripartire l'insieme di partenza. Successivamente, queste procedure si articolano in due fasi:

1. Determinazione di una prima partizione delle n unità in un numero g di gruppi;
2. Successivamente avviene lo spostamento delle unità all'interno dei gruppi g, allo scopo di ottenere una partizione migliore in base ai criteri di omogeneità interna ai gruppi ed eterogeneità tra di essi.

Individuare la partizione ottimale comporta però l'esame di tutte le possibili assegnazioni delle n unità nei g gruppi, un'operazione molto onerosa in termini di calcoli. Per risolvere tale problema, i metodi non gerarchici operano attraverso una strategia di raggruppamento basata sulla valutazione di un numero ristretto e accettabili di possibili partizioni alternative: scelta la partizione iniziale, vengono allocate le unità in funzione della massimizzazione della funzione obiettivo.

L'idea di fondo è individuare dei centroidi attorno ai quali costituire i gruppi: le osservazioni sono associate al centroide a cui sono più vicine. La posizione di ogni centroide non è fissa, ma viene aggiornata man mano che l'algoritmo procede nel raggruppamento.

Supponendo di dividere, secondo un criterio generico, le osservazioni in K gruppi, la dissimilarità totale può essere vista come:

$$\sum_{i,i'} d\left(i,i'\right) = \sum_{k=1}^{K} \sum_{G(i)=k} \left(\sum_{G(i')=k} d(i,i') + \sum_{G(i')\neq k} d(i,i') \right) = D_{entro} + D_{tra} \tag{3.8}$$

con $G(i)$ il gruppo a cui è assegnato l'individuo i-esimo,

$D_{entro} = \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')=k} d(i,i')$ la dissimilarità complessiva all'interno dei gruppi,

[64] M. Andemberg, *Cluster analysis for applications*, New York Academic Press, 1973.

$D_{tra} = \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')\neq k} d(i,i')$ la dissimilarità complessiva tra i diversi gruppi. L'obiettivo prefissato è scegliere i gruppi in modo da minimizzare la dissimilarità all'interno di essi: D_{entro} deve quindi essere minimizzata, mentre va massimizzata D_{tra}.

In conseguenza del fatto che il numero di raggruppamenti che si possono creare per un certo valore di K è finito, la minimizzazione può essere realizzata per un numero finito di operazioni, scandendo tutte le possibili scelte[65]. Considerando la crescita esponenziale del numero di raggruppamenti possibili all'aumentare di n, questa opzione non è attuabile: si ricorre quindi ad algoritmi sub-ottimi.

Il più utilizzato di questi algoritmi è il *K-means*, basato sul criterio di somma dei quadrati (SSQ). Questo algoritmo è stato, in una prima fase proposto molti autori in diverse forme e sotto differenti assunzioni; successivamente, sono state investigati gli aspetti teorici di fondo dell'algoritmo e possibili modifiche al metodo stesso. Il *K-means* divenne così ben presto una procedura standard all'interno dei metodi di raggruppamento, anche se in alcuni contesti era conosciuto con denominazioni diverse, tra cui *dynamic cluster method*, *iterated minimum-distance partition method* e *nearest centroid sorting*[66]. Il nome K-means fu invece adottato per la prima volta da MacQueen[67] per il suo sequenziale algoritmo per la minimizzazione del criterio di somma dei quadrati SSQ. Infine nel 1979, in un'opera di Diday[68], a cui contribuirono 22 co-autori, fu dato un notevole livello di generalizzazione delle idee base dell'algoritmo e si stabilì il suo utilizzo per il *clustering* basato su modelli. Il *K-means* divenne così ben presto una procedura standard all'interno dei metodi di raggruppamento.

Questo algoritmo utilizza la distanza euclidea per formare la dissimilarità tra variabili quantitative:

$$D_{entro} = \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')=k} \| \tilde{x}_i - \tilde{x}_{i'} \|^2 = 2 \sum_{k=1}^{K} \sum_{G(i)=k} \| \tilde{x}_i - m_k \|^2 \tag{3.9}$$

dove m_k è il vettore formato dalle medie aritmetiche di ciascuna variabile del k-esimo gruppo.

Il *K-means* si propone di minimizzare D_{entro} una volta fissato il numero K di gruppi e la posizione iniziale dei centroidi x_k: l'algoritmo procede iterativamente, raggruppando

[65] A. Azzalini e B. Scarpa, *Analisi dei dati e Data Mining*, Springer, Milano, 2004.
[66] H.H. Bock, *cit.*
[67] J. MacQueen, *Some methoods for classification and analysis of mulivariate observations*, in: L.M. LeCam, J. Neyman (eds.): Proc. 5th Berkely Symp. Math. Statist. Probab. 1965/66. Univ. of California Press, Berkely 1967, vol. I.
[68] E. Diday et al, *Optimisation en classification automatique.* Vol. I, II. Institut National der Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France 1979.

in successione gli individui intorno ai centroidi, anch'essi soggetti all'aggiornamento iterativo, fino al raggiungimento di un punto di stazionarietà. La convergenza è sempre raggiunta, ma non è certo che corrisponda ad un punto di minimo assoluto della funzione obiettivo[69]: spesso i valori convergono ad un punto di minimo locale.

I limiti che si possono annoverare nell'utilizzo di algoritmi non gerarchici sono:

- la classificazione finale può essere influenzata dalla scelta iniziale dei poli: occorre porre attenzione all'ordine delle unità;
- si possono ottenere soluzioni instabili in presenza di valori anomali, numerosità insufficiente o qualora non sussista una struttura in gruppi nei dati;
- la sua applicabilità esclusivamente a variabili quantitative. Questo limite può essere superato sostituendo alla distanza euclidea un altro criterio di dissimilarità ed introducendo il concetto di medoide, ovvero un'unità presa a rappresentanza del gruppo tale da minimizzare la dissimilarità all'interno di esso.

3.5 Determinazione del numero ottimale dei gruppi

Stabilito l'algoritmo di raggruppamento de utilizzare, è necessario scegliere quanti *cluster* utilizzare. Per definizione, la maggior parte delle procedure di *cluster analysis* continuano senza sosta finché tutti gli oggetti del dataset iniziale sono assegnati ad un gruppo. Come abbiamo evidenziato nei paragrafi precedenti, nel caso di algoritmi di tipo gerarchico l'analisi si arresta nel momento in cui tutti gli oggetti sono contenuti all'interno di un singolo *cluster*. Questa soluzione è di scarso interesse per la ricerca applicata. Fortunatamente, esistono alcuni criteri che possono assisterci nella determinazione del numero ottimale dei gruppi[70]:

- I criteri esterni fanno affidamento su informazioni provenienti dall'esterno per valutare la partizione dei dati. Molti di questi criteri sono stati formulati, ma sono utilizzati raramente in quanto la vera struttura dei gruppi non è quasi mai conosciuta a priori;

[69] A. Azzalini e B. Scarpa, *cit.*

[70] P.A. Gore, *Cluster Analysis*, in Handbook of Applied Multivariate Statistics and Mathematical Modeling, Southern Illinois University Academic Press, Carbondale Illinois, 2000.

- I criteri interni, chiamati anche *stopping rules*, utilizzano informazioni inerenti alla soluzione proposta dall'analisi di raggruppamento per determinare il grado di somiglianza tra la partizione ottenuta e il data set da cui è stata creata la partizione. Il gran numero di criteri interni preclude una dettagliata descrizione degli stessi, ma i risultati di alcuni studi empirici rivelano che essi variano in validità a seconda del caso concreto, e sono per questo di difficile implementazione;

Nonostante molte delle *stopping rules* proposte siano state integrate nei programmi statistici come R, SAS o SPSS, chi analizza i risultati ha bisogno di fare comunque affidamento sulla sua razionalità teorica, sulla sua valutazione soggettiva o in ulteriori ed addizionali computazioni statistiche per determinare il numero ottimale di *cluster* da utilizzare. Tra di esse, le maggiormente utilizzate sono la rappresentazione grafica con ordinata il numero dei gruppi e in ascissa i valori corrispondenti della misura della dissomiglianza, in cui il numero ottimale è il valore in corrispondenza del quale si osserva un forte appiattimento della spezzata; il metodo delle differenze tra misure di dissomiglianza a passi successivi, in cui data la formula $\Delta g = d_{g-1} - d_g$il numero di gruppi ottimo coincide con quello per cui Δ_g è massimo; l'utilizzo della Pseudo F, in cui si prende g in corrispondenza di F massimo, con $F = \frac{GSS/(k-1)}{WSS/(n-k)}$, GSS la somma dei quadrati tra i gruppi e WSS la somma dei quadrati all'interno dei gruppi: questo indice è valido per qualunque algoritmo gerarchico.

3.6 Validazione dei risultati ottenuti e presentazione dei risultati

Lo step finale dello sviluppo di un sistena di classificazione è valutare la validità della struttura di *cluster* ottenuta al di fuori del dataset utilizzato nell'analisi di raggruppamento. Si testa cioè l'utilità predittiva e descrittiva della struttura e si valuta la sua coerenza confrontando i risultati con quelli ottenuti utilizzando altri campioni, indipendenti da quello iniziale, facenti parte della stessa popolazione. Se si desidera fare inferenza oltre la popolazione utilizzata nello studio, la soluzione ottenuta dall'analisi deve necessariamente essere validata su popolazioni diverse.

La soluzione prospettata da un'analisi di raggruppamento guadagna forza e credibilità nel caso in cui si è capaci di replicare la soluzione ottenuta utilizzando diverse osservazioni.

Concluso il processo di validazione, si procede con la presentazione dei risultati ottenuti

nell'analisi. Aldenderfer e Blashfield[71] hanno stilato alcune linee guida a riguardo:

- spiegare e descrivere la struttura teorica ed empirica utilizzata nello studio sui gruppi e come oggetti e osservazioni sono stati selezionate per includerle nel caso;
- fornire una descrizione soddisfacente della similarità o del metodo di distanza utilizzato nell'analisi, e spiegare per quale motivo è stata fatta quella scelta precisa;
- infine è necessario fornire una chiara descrizione di come i gruppi sono stati selezionati e evidenziare la validità della struttura ottenuta.

[71] M.S. Aldenderfer e R.K. Blashfield, *Cluster analysis*, Sage Publications, Thousand Oaks, CA, 1984.

4 Analisi empirica: una cluster analysis per le performance economiche e finanziarie delle società di Serie A

In questa sezione del lavoro verrà sviluppata un'analisi sui bilanci delle società partecipanti alla Serie A italiana durante le stagioni 2010/11, 2011/12, 2012/13. L'intento è quello di individuare, attraverso una *cluster analysis*, se vi sono squadre caratterizzate da una situazione economico-finanziaria simile. I risultati ottenuti saranno poi valutati attraverso un confronto con i risultati sportivi conseguiti durante lo svolgimento degli stessi campionati.

In particolare, presenteremo:

- il dataset utilizzato per l'analisi;
- il modello scelto e le sue caratteristiche;
- il risultato dell'analisi.

La finalità del modello è classificare, in base agli indici precedentemente dedotti dalla riclassificazione di bilancio, le società calcistiche in base a criteri di dissomiglianza. Inoltre, l'analisi verrà condotta considerando le tre stagioni come un unico dataset, per valutare possibili trend migliorativi o peggiorativi della performance economica e se le società che hanno conseguito gli stessi risultati sportivi in annate diverse erano dotate di una struttura economico-finanziaria simile o meno.

4.1 Variabili utilizzate nel modello ed analisi descrittiva del dataset

La definizione delle variabili è avvenuta mediante una riclassificazione del bilancio[72] di esercizio delle società disputanti il campionato calcistico di Serie A per tre stagioni sportive

[72] Questa operazione consiste nell'aggregare i molteplici valori del bilancio per interpretare meglio l'andamento dell'impresa.

consecutive: 2010/2011, 2011/2012, 2012/2013. Al fine di valutare la bontà strutturale delle società sportive interessate dall'analisi, sono stati utilizzati alcuni indici di bilancio, già descritti nella sezione 1.4 del primo capitolo:

- **Totale Attività**, in cui sono raccolte tutte le voci presenti nell'attivo patrimoniale del bilancio d'esercizio per ogni singola società;
- **EBITDA**, che indica l'utile ottenuto dall'azienda al netto degli interessi passivi, delle imposte, degli ammortamenti e delle svalutazioni. Un suo valore negativo denota problemi di redditività per la società;
- **EBIT**, una misura della resa del capitale investito nella società;
- **Utile (perdita)**: risultato finale della gestione, che può essere positivo o negativo;
- **ROA**, che esprime il rendimento delle attività impiegate nell'attività d'impresa;
- **ROS**: rappresenta la redditività della società in relazione alla capacità remunerativa del flusso di ricavi;
- **EBITDA Margin**, indice della redditività operativa della società: è legato alle condizioni di mercato e ai vantaggi competitivi specifici dell'impresa;
- **Tasso di Indebitamento**, una misura relativa del valore complessivo delle passività che esprime un giudizio sul livello di rischio finanziario associato alla società;
- **Indice di solidità patrimoniale**, che consente di valutare la proporzione esistente tra i Mezzi di terzi, al netto delle attività finanziarie disponibili, ed il Patrimonio netto al fine di coprire il fabbisogno generato dal capitale investito netto.

Le nove variabili elencate sono state raccolte per le 20 società disputanti la serie A per le stagioni considerate, formando un dataset di 60 unità:

TEAM	EBITDA	EBIT	RIS. NETTO	TOT. ATTIVITA'	ROS	ROA	EBITDA M.	T. IND.	I. SOL. PATR.
Bari (10/11)	2.008.420	-5.633.201	14.175.598	83.527.237	-0,10	-0,07	0,03	94,93	10,83
Bologna (10/11)	16.656.807	93.262	-4.166.419	106.854.597	0,00	0,00	0,27	4,62	0,18
Brescia (10/11)	6.872.964	821.124	23.527	51.737.622	0,02	0,02	0,15	4,06	1,59
Cagliari (10/11)	1.428.617	-5.561.423	-1.807.384	85.406.910	-0,12	-0,07	0,03	1,89	-0,04
Catania (10/11)	19.583.149	10.658.605	6.449.511	110.899.895	0,19	0,10	0,36	3,33	1,04
Cesena (10/11)	4.546.383	-288.186	-2.365.004	75.294.748	-0,01	0,00	0,10	148,95	28,79
Chievo (10/11)	10.935.947	2.339.890	-257.661	75.245.180	0,05	0,03	0,26	109,73	35,61
Fiorentina (10/11)	317.285	-32.269.202	-9.604.353	178.314.365	-0,39	-0,18	0,00	1,14	-0,02
Genoa (10/11)	23.174.645	-14.194.427	-16.964.706	222.218.890	-0,14	-0,06	0,23	207,02	73,36
Inter (10/11)	-30.512.076	-92.143.928	-86.813.786	455.690.888	-0,35	-0,20	-0,11	-19,84	-6,88
Juventus (10/11)	-31.596.582	-92.154.792	-95.414.019	334.040.001	-0,54	-0,28	-0,18	-68,46	-24,46
Lazio (10/11)	34.317.142	22.138.078	9.982.408	165.245.840	0,24	0,13	0,37	14,73	0,77
Lecce (10/11)	8.205.930	476.421	-747.825	33.576.897	0,01	0,01	0,21	14,97	6,59
Milan (10/11)	-48.361.638	-104.650.358	-64.803.893	522.486.758	-0,49	-0,20	-0,23	70,94	49,13
Napoli (10/11)	44.750.379	10.212.596	4.197.829	110.053.332	0,08	0,09	0,34	2,75	0,13
Palermo (10/11)	36.159.744	12.544.314	7.769.812	107.810.852	0,12	0,12	0,36	2,57	0,87
Parma (10/11)	26.789.404	3.905.508	658.122	141.993.787	0,04	0,03	0,29	3,90	1,51
Roma (10/11)	-561.000	-24.781.000	-30.778.000	102.560.000	-0,16	-0,24	0,00	-3,33	-1,25
Sampdoria (10/11)	403.353	-14.668.963	-12.109.456	98.868.266	-0,23	-0,15	0,01	5,97	0,67
Udinese (10/11)	27.862.672	8.366.394	2.897.161	161.022.474	0,08	0,05	0,28	2,93	0,00

Table 4.1: Dataset utilizzato per l'analisi - stagione 2010/11

TEAM	EBITDA	EBIT	RIS. NETTO	TOT. ATTIVITA'	ROS	ROA	EBITDA M.	T. IND.	I. SOL. PATR.
Atalanta (11/12)	-1.340.149	-13.807.902	-10.918.220	70.204.775	-0,35	-0,20	-0,03	13,24	1,50
Bologna (11/12)	18.978.483	-3.588.606	-6.189.530	106.800.892	-0,05	-0,03	0,27	7,34	-0,17
Cagliari (11/12)	22.355.929	8.474.089	2.508.267	111.007.619	0,13	0,08	0,33	2,70	0,32
Catania (11/12)	18.629.180	8.253.889	4.292.614	106.037.068	0,15	0,08	0,34	2,42	0,76
Cesena (11/12)	15.206.188	6.550.980	2.070.233	112.129.605	0,12	0,06	0,27	42,59	6,14
Chievo (11/12)	17.827.207	4.235.453	616.383	83.744.747	0,08	0,05	0,33	63,62	17,55
Fiorentina (11/12)	-10.106.744	-43.786.473	-32.474.084	156.972.324	-0,60	-0,28	-0,14	2,10	0,33
Genoa (11/12)	49.967.680	5.692.955	-67.494	304.389.285	0,04	0,02	0,38	303,16	104,96
Inter (11/12)	-17.272.391	-76.881.910	-77.147.926	478.194.761	-0,32	-0,16	-0,07	-23,42	-8,78
Juventus (11/12)	7.467.731	-41.188.373	-48.654.550	427.780.347	-0,19	-0,10	0,03	5,62	1,96
Lazio (11/12)	10.730.596	-10.985.332	580.492	263.697.029	-0,13	-0,04	0,13	2,16	0,25
Lecce (11/12)	5.589.971	-9.140.616	-9.202.085	15.152.969	-0,28	-0,60	0,17	-3,13	-1,00
Milan (11/12)	-44.880.070	-103.085.887	-75.540.069	523.141.399	-0,45	-0,20	-0,20	26,85	19,55
Napoli (11/12)	57.776.448	26.175.580	14.720.757	138.168.981	0,17	0,19	0,37	2,14	-0,24
Novara (11/12)	3.508.749	-637.567	-1.306.809	29.594.064	-0,02	-0,02	0,12	28,63	3,37
Palermo (11/12)	18.809.422	-4.539.737	-4.011.700	146.460.534	-0,05	-0,03	0,19	4,61	-0,10
Parma (11/12)	26.849.740	4.215.596	-2.467.709	172.585.520	0,04	0,02	0,25	5,52	1,96
Roma (11/12)	-19.616.000	-50.615.000	-58.474.000	188.397.000	-0,37	-0,27	-0,14	-4,59	-1,08
Siena (11/12)	-722.320	-17.401.641	1.817.249	92.690.294	-0,33	-0,19	-0,01	38,98	3,82
Udinese (11/12)	39.296.761	19.023.727	8.782.162	172.517.862	0,15	0,11	0,32	2,49	-0,10

Table 4.2: Dataset utilizzato per l'analisi - stagione 2011/12

TEAM	EBITDA	EBIT	RIS. NETTO	TOT. ATTIVITA'	ROS	ROA	EBITDA M.	T. IND.	I. SOL. PATR.
Atalanta (12/13)	10.239.553	-1.648.462	-2.155.308	75.917.973	-0,03	-0,02	0,17	26,18	7,16
Bologna (12/13)	16.787.839	-2.321.295	-3.976.046	88.599.161	-0,03	-0,03	0,23	8,48	0,94
Cagliari (12/13)	6.477.314	-3.466.107	-985.459	93.881.169	-0,07	-0,04	0,14	2,24	0,13
Catania (12/13)	13.768.860	391.183	91.713	89.149.473	0,01	0,00	0,28	2,06	0,91
Chievo (12/13)	17.022.622	3.575.246	1.530.557	110.838.146	0,06	0,03	0,28	35,87	10,04
Fiorentina (12/13)	36.532.929	3.860.990	1.155.691	182.081.302	0,03	0,02	0,33	1,41	0,09
Genoa (12/13)	36.250.369	-6.656.835	-14.846.953	308.407.534	-0,05	-0,02	0,27	266,30	79,57
Inter (12/13)	-14.623.808	-75.480.499	-79.881.808	447.519.240	-0,37	-0,17	-0,07	-69,09	-24,17
Juventus (12/13)	56.711.196	-3.806.006	-15.910.649	443.366.100	-0,01	-0,01	0,20	8,12	3,28
Lazio (12/13)	17.511.888	-3.723.046	-5.894.288	169.728.461	-0,03	-0,02	0,16	18,48	0,63
Milan (12/13)	61.928.000	4.649.000	-6.857.000	334.284.000	0,01	0,01	0,19	-7,08	-4,55
Napoli (12/13)	51.156.282	11.288.212	8.073.447	136.748.114	0,07	0,08	0,33	1,62	-0,35
Palermo (12/13)	6.216.799	-18.962.747	-19.388.536	107.172.901	-0,27	-0,18	0,09	6,95	0,94
Parma (12/13)	10.234.096	-20.018.405	-3.223.792	212.539.072	-0,23	-0,09	0,12	8,14	3,33
Pescara (12/13)	9.907.268	2.693.256	1.260.478	37.200.953	0,06	0,07	0,22	7,74	1,34
Roma (12/13)	946.000	-31.369.000	-40.130.000	173.966.000	-0,20	-0,18	0,01	-3,64	-1,36
Sampdoria (12/13)	-38.995.798	-51.749.296	-38.128.752	84.505.250	-1,50	-0,61	-1,13	7,66	1,83
Siena (12/13)	14.546.689	-859.448	-1.772.507	102.263.027	-0,01	-0,01	0,22	186,40	24,56
Torino (12/13)	-5.522.450	-16.404.925	-10.991.484	62.954.770	-0,50	-0,26	-0,17	81,84	2,68
Udinese (12/13)	68.390.453	41.309.172	32.265.947	250.172.834	0,25	0,17	0,42	2,69	0,00

Table 4.3: Dataset utilizzato per l'analisi - stagione 2012/13

All'interno del campione, che non presenta dati mancanti, vi sono 15 società presenti per tutte le stagioni considerate, mentre altre, a seconda dei meccanismi di retrocessione/promozione nella massima serie[73], ne fanno parte solo per uno o due campionati. Nello specifico, il Bologna, il Cagliari, il Catania, il Chievo Verona, la Fiorentina, il Genoa, L'Internazionale, la Juventus, la Lazio, il Milan, il Napoli, il Palermo, il Parma, la Roma e l'Udinese hanno disputato tutte le stagioni considerate; Atalanta, Cesena, Lecce, Sampdoria e Siena hanno partecipato a due campionati; mentre il Bari, il Brescia, il Novara, il Pescara e il Torino hanno preso parte ad una sola stagione.

Saranno ora analizzate brevemente le statistiche descrittive delle variabili utilizzate nel modello:

variabili	2010/2011	
	totale	*media*
EBITDA	152.981.545,00	7.649.077,25
EBIT	-314.789.288,00	-15.739.464,40
RIS. NETTO	-279.678.538,00	-13.983.926,90
TOT. ATTIVITÀ	3.222.848.539,00	161.142.426,95
ROA	-9,77%	
ROS	-16,08%	
EBITDA MARG.	7,81%	
T. INDEBIT.	11,36	
SOLID. PATR.	3,80	

Table 4.4: Statistiche descrittive delle variabili per la stagione 2010/11

[73]Nelle stagioni considerate, le tre società che occupavano alla 38° giornata le ultime tre posizioni della classifica, venivano retrocesse in serie B.

variabili	2011/12	
	totale	*media*
EBITDA	219.056.411,00	10.952.820,55
EBIT	-293.036.775,00	-14.651.838,75
RIS. NETTO	-291.066.019,00	-14.553.300,95
TOT. ATTIVITÀ	3.699.667.075,00	184.983.353,75
ROA	-7,92%	
ROS	-14,31%	
EBITDA MARG.	10,70%	
T. INDEBIT.	9,01	
SOLID. PATR.	2,73	

Table 4.5: Statistiche descrittive delle variabili per la stagione 2011/12

variabili	2012/13	
	totale	*media*
EBITDA	375.486.101,00	18.774.305,05
EBIT	-168.699.012,00	-8.434.950,60
RIS. NETTO	-199.764.749,00	-9.988.237,45
TOT. ATTIVITÀ	3.511.295.480,00	175.564.774,00
ROA	-4,80%	
ROS	-7,43%	
EBITDA MARG.	16,54%	
T. INDEBIT.	12,94	
SOLID. PATR.	3,82	

Table 4.6: Statistiche descrittive delle variabili per la stagione 2012/13

variabili	*tot. aggregato*	*media aggr.*
EBITDA	747.524.057,00	12.458.734,28
EBIT	-776.525.075,00	-12.942.084,58
RIS. NETTO	-770.509.306,00	-12.841.821,77
TOT. ATTIVITÀ	10.433.811.094,00	173.896.851,57
ROA	-7,44%	
ROS	-12,37%	
EBITDA MARG.	11,91%	
T. INDEBIT.	10,83	
SOLID. PATR.	3,36	

Table 4.7: Statistiche descrittive delle variabili - valori aggregati

In particolare, si nota:

- L'andamento crescente del valore dell'indice EBITDA nel periodo considerato: il trend è imputabile all'aumento del valore della produzione rispetto a costi della produzione tendenzialmente costanti (fig. 4.1);

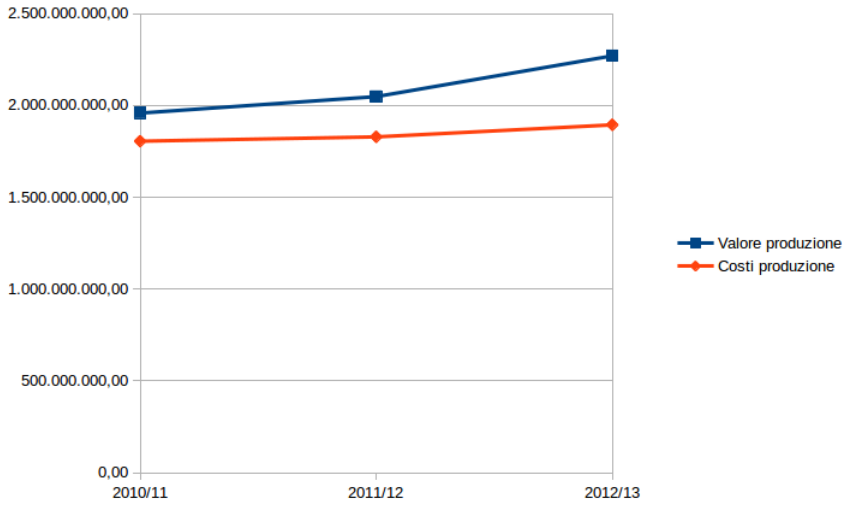

Figure 4.1: Trend del valore della produzione e dei rispettivi costi

- L'andamento negativo e crescente dell'EBIT (nonostante la crescita del valore degli ammortamenti e degli accantonamenti) determinato dal trend di crescita dell'EBITDA;
- il conseguente trend crescente del risultato netto, anche se le perdite aggregate restano consistenti (-199.764.749 € nella stagione 2012/2013);
- il valore del ROA e del ROS, più che dimezzati nel corso del triennio ma associati a valori negativi;
- il trend di crescita dell'*EBITDA margin*, che assume valori positivi, figlio del costante aumento delle voci di ricavo;

4.2 Il Software R

Su questo campione si è deciso di operare una *cluster analysis* al fine di individuare gruppi di società con valori simili dal punto di vista economico-finanziario. Il software di analisi statistica utilizzato è R, distribuito tramite licenza GNU GPL e disponibile

per diversi sistemi operativi. Attualmente, è al 18-esimo posto del TIOBE Programming Community Index[74].

R è un linguaggio e un ambiente di sviluppo per il calcolo statistico e per la grafica[75]. Si tratta di un progetto *open source* simile al linguaggio di programmazione S, di cui può essere considerato come una diversa implementazione. Il linguaggio è orientato agli oggetti ed è stato sviluppato inizialmente da Robert Gentleman e Ross Inaka del dipartimento di statistica dell'università di Auckland.

Il software offre una grande varietà di tecniche grafiche e statistiche: test statistici tradizionali, tecniche di modellazione lineare e non lineare, analisi sulle serie storiche, classificazione e *clustering*. Inoltre, l'ambiente di R può essere facilmente esteso tramite estensioni chiamate pacchetti, scaricabili da un'apposita categoria di siti Internet chiamata CRAN (*Comprehesive R Archive Network*).

4.3 I risultati dell'analisi

Al fine di individuare un possibile e adeguato numero di gruppi in cui suddividere il campione, il dataset è stato analizzato con degli algoritmi gerarchici. In particolare, sono stati utilizzati dei algoritmi agglomerativi, secondo i quali si procede a raggruppare le varie unità statistiche attraverso aggregazioni successive partendo da un numero di *cluster* uguale al numero degli individui (in questo caso, quindi, 20).

Un'operazione preliminare è stata la trasformazione della matrice iniziale dei dati $(n \times p)$ in una matrice $(n \times n)$ di dissimilarità tra le coppie di individui. Successivamente:

- vengono uniti gli elementi più "vicini" della matrice di dissimilarità creata, formando un nuovo cluster composto da questi due individui;
- si definisce una nuova matrice di dissimilarità per calcolare la distanza tra il gruppo appena formato e i gruppi già esistenti. Esistono diverse modalità di calcolo della dissimilarità tra i gruppi, come spiegato nella sezione 3.4.1;
- si uniscono i due nuovi individui (o *cluster*) più "vicini";
- Si reitera il procedimento, ridefinendo ad ogni passo la matrice di dissimilarità.

Per calcolare il grado di dissimilarità tra i gruppi è stato utilizzato il metodo del legame singolo (la distanza tra due conglomerati è la minima distanza tra tutte le possibili

[74] la classifica è consultabile al link `www.tiobe.com/index.php/content/paperinfo/tpci/index.html`
[75] Per ulteriori informazioni, si visiti il sito `www.r-project.org`

coppie di individui), il metodo del legame completo (la distanza tra due conglomerati è la massima distanza tra tutte le possibili coppie di individui), il metodo del legame medio (la distanza tra due conglomerati è la distanza media tra tutte le possibili coppie di individui) e il metodo del centroide (la distanza tra due conglomerati è la distanza tra i centroidi di ogni *cluster*). I rispettivi dendrogramma associati al procedimento sopra descritto sono qui riportati.

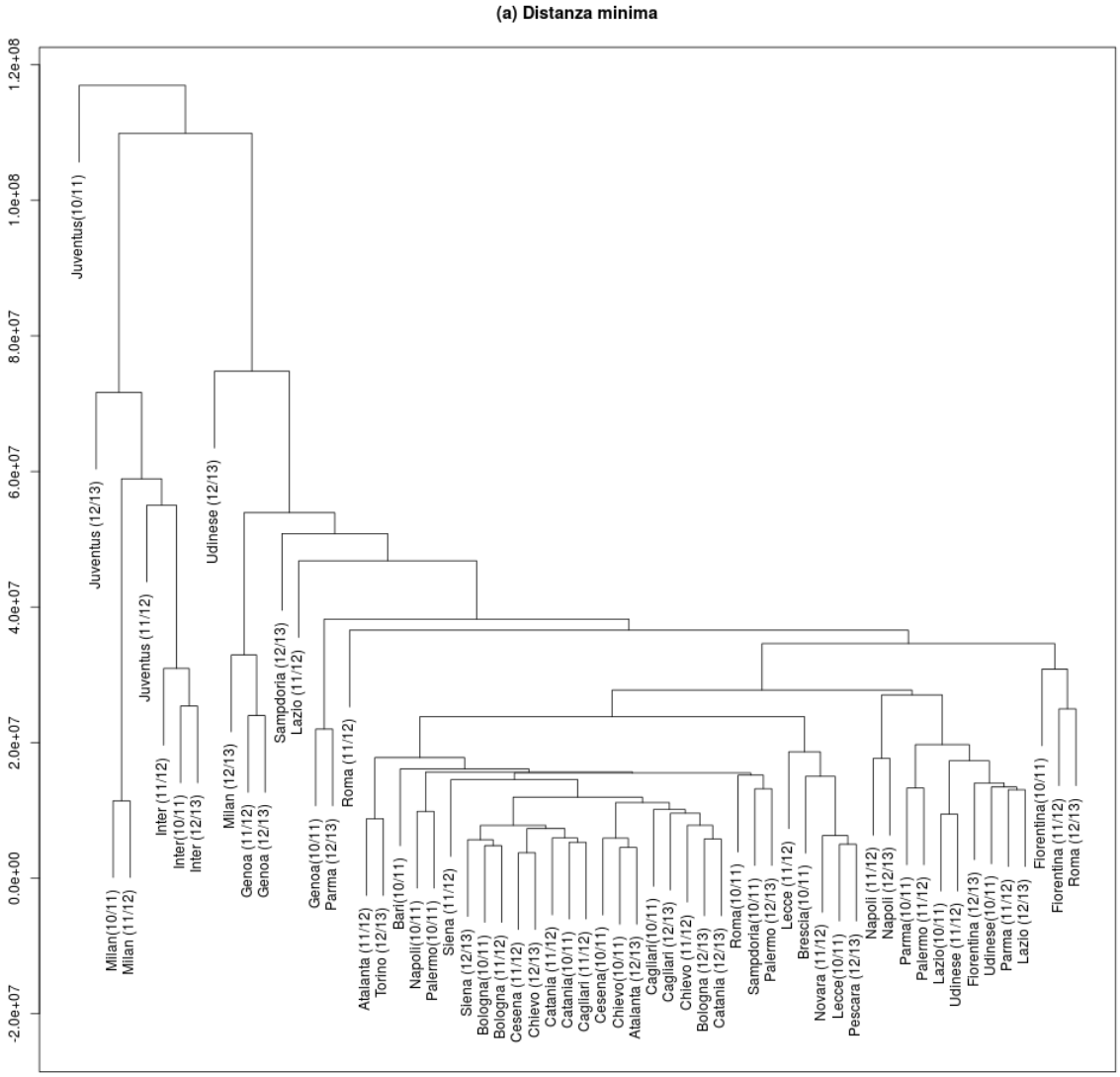

Figure 4.2: Dendrogramma derivato dal metodo del legame singolo

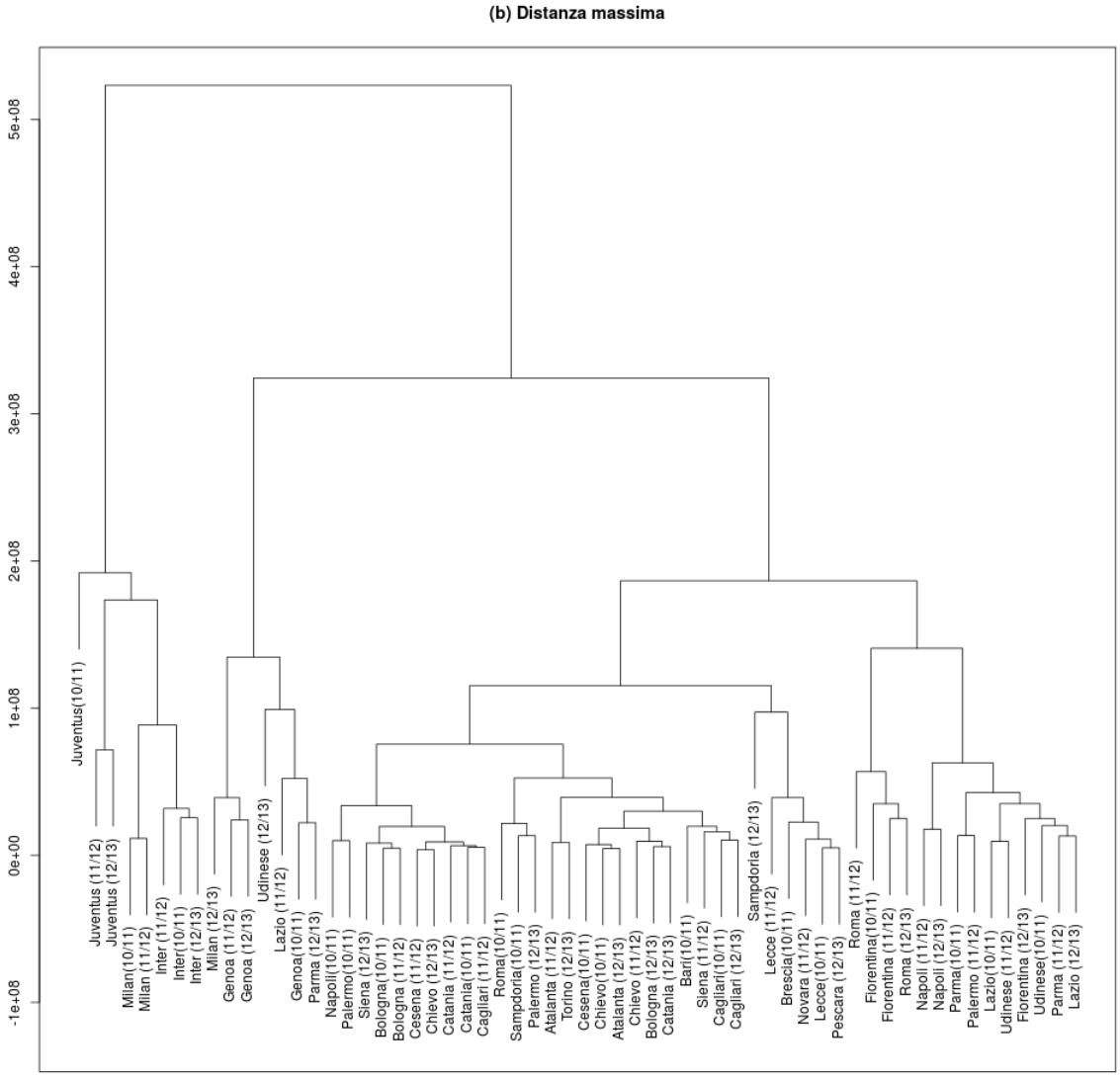

Figure 4.3: Dendrogramma derivato dal metodo del legame completo

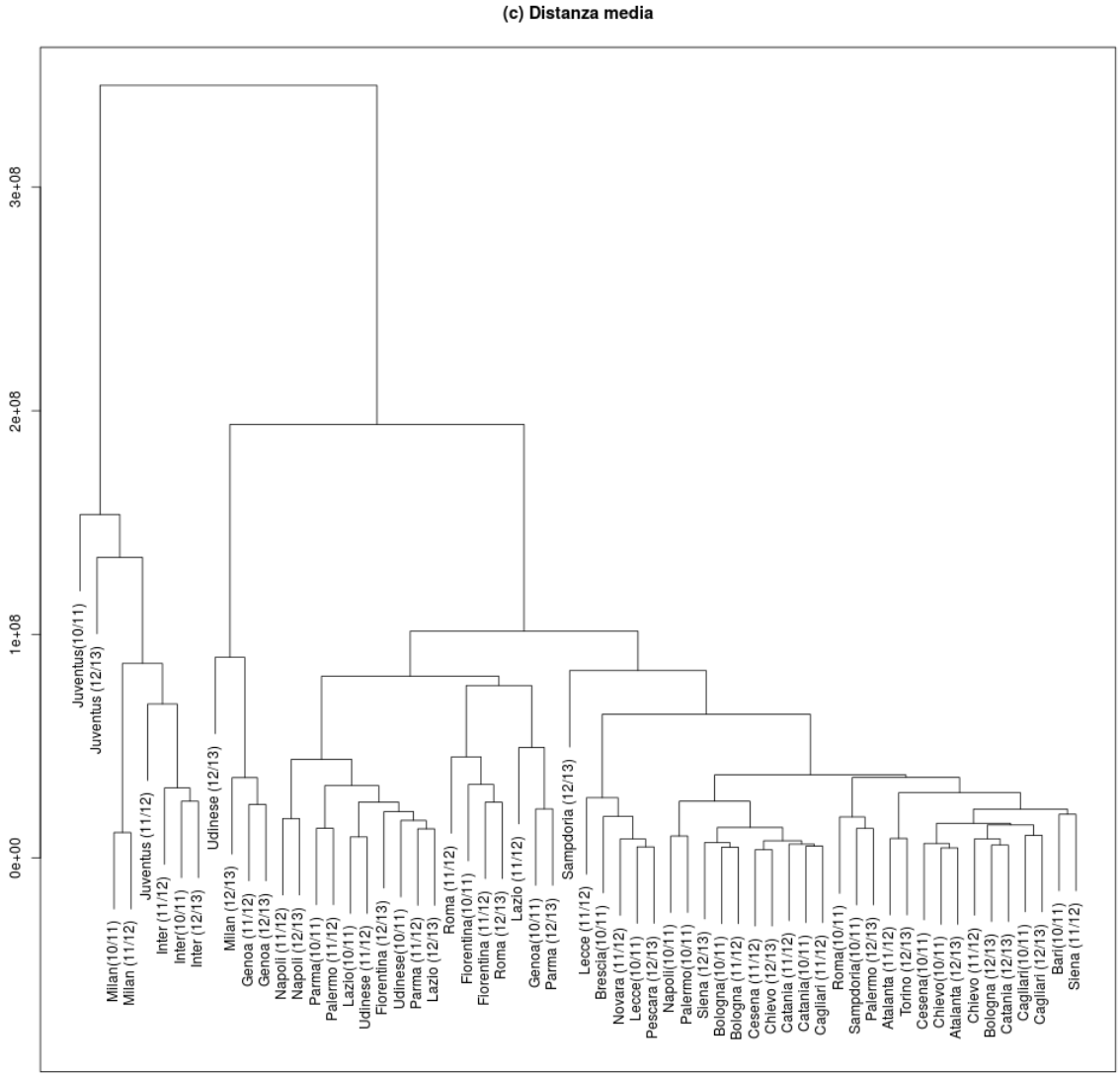

Figure 4.4: Dendrogramma derivato dal metodo del legame medio

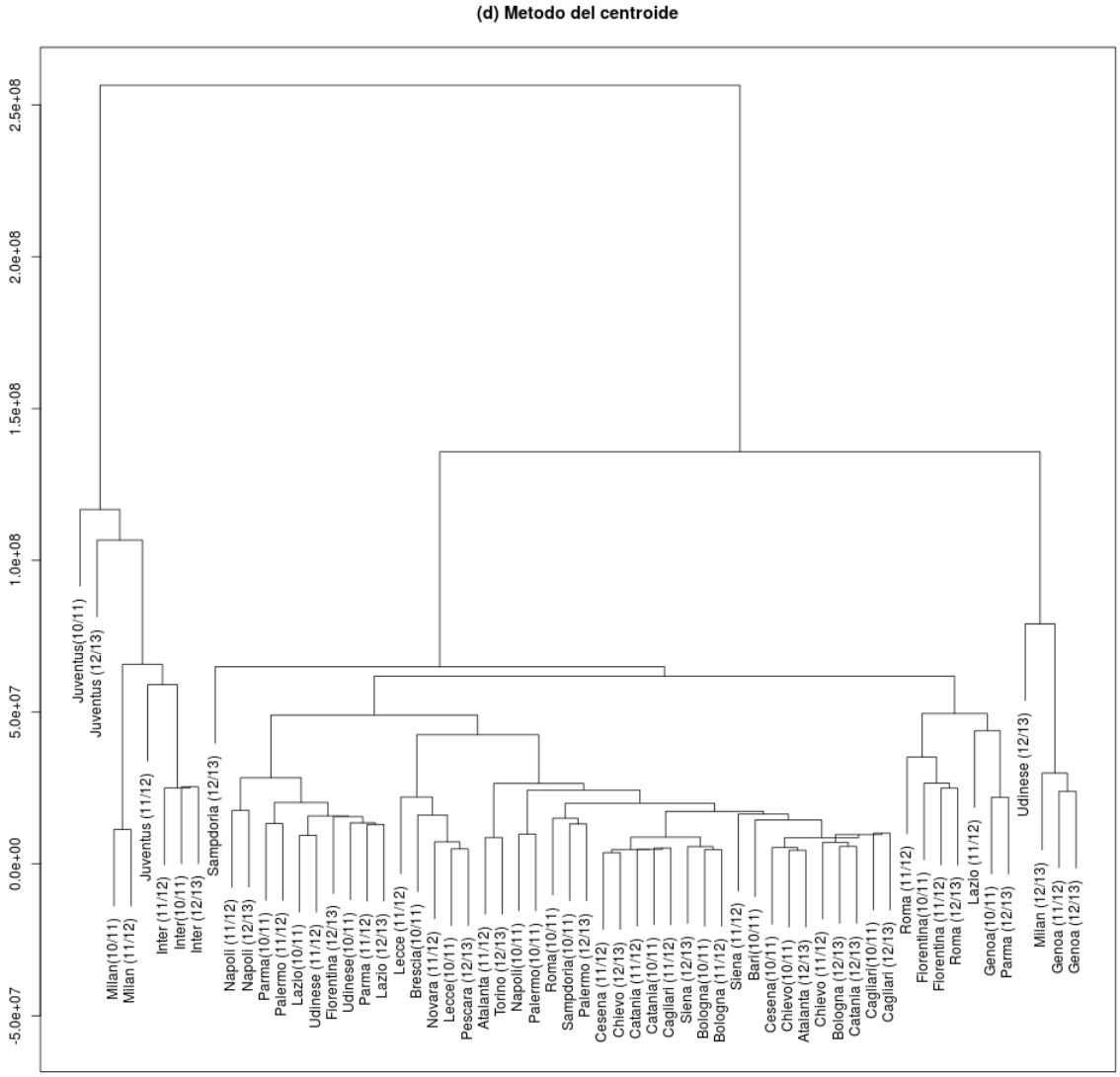

Figure 4.5: Dendrogramma derivato dal metodo del centroide

Confrontando i vari grafici, si nota in tutti un numero elevato di *cluster*, alcuni formati da un singolo individuo. La scelta è ricaduta sul metodo del legame completo: tagliando ad un'altezza pari a $(7,4)\,e^7$ i *cluster* trovati sono 13 (fig. 4.6)

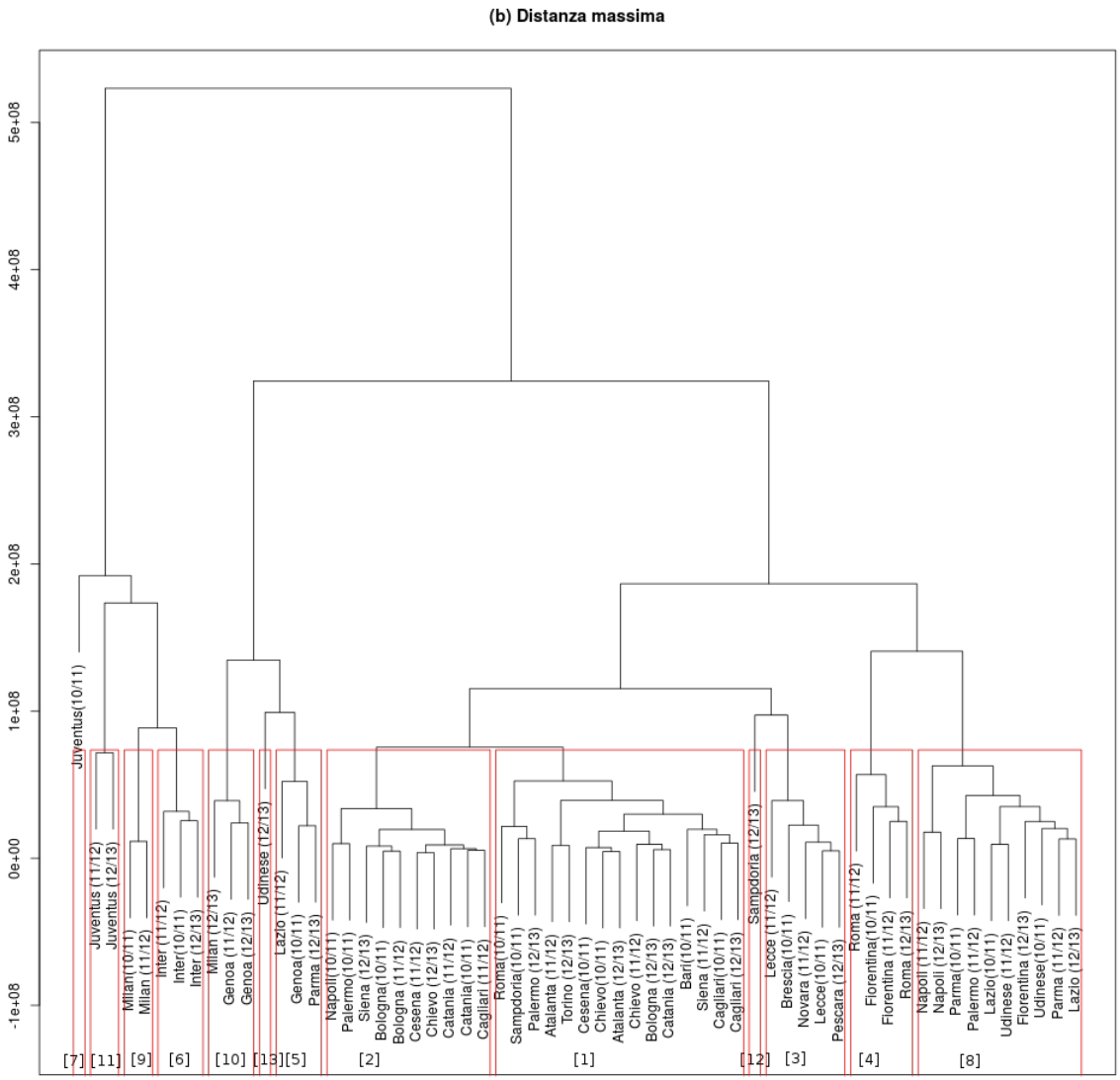

Figure 4.6: Clusters ottenuti dal dendrogramma derivato dal metodo completo

Tre società (Juventus 10/11, Udinese 12/13, Sampdoria 12/13) formano dei gruppi singoli, mentre il *cluster* più numeroso conta 15 individui.

Calcolando i valori medi per ogni *cluster*, si nota (fig. 4.8):

- Le società "sane", ovvero quelle con i conti in regola, che rispettano la regola UEFA del pareggio di bilancio (producono utile) sono divise in 3 *cluster* diversi: nel numero [2], nel [8] e nel [13] (quest'ultimo composto dalla sola Udinese 12/13). Le

unità componenti i *clusters* citati sono state divise e non raggruppate in un unico gruppo per i diversi valore del tasso di indebitamento e del tasso di solidità patrimoniale: per il *cluster* [2] essi sono rispettivamente $29,06$ e $4,39$, mentre per il numero [8] $5,72$ e $0,42$. L'Udinese 12/13 forma un gruppo a sé stante per il valore del totale delle attività, decisamente maggiore rispetto a quello medio degli altri due;

- i *cluster* [1], [3], [5] e [10] sono caratterizzati da società che producono contenute perdite di esercizio. Le differenze tra questi gruppi vanno ricercate nel diverso valore:
  - del totale attivo, rispettivamente nell'ordine di 85 milioni di euro, 33 milioni, 233 milioni e 315 milioni;
  - dell'indice di indebitamento, che cresce dal contenuto $10,45$ per il gruppo [3] fino al $187,47$ del [10]. Stessa riflessione può essere condotta riguardo i valore dell'indice di solidità patrimoniale, che assume valore $2,38$ per il *cluster* [3] mentre per il [10] segna $59,99$;
  - del ROA e del ROS, negativi per tutti i *clusters* tranne che per il numero [10], dove assumono valore 0;
- le società inserite nei *clustess* [4], [11] e [12] (quest'ultimo formato dalla sola Sampdoria 12/13) sono caratterizzate da perdite di esercizio superiori ai 30 milioni di euro: in particolare, le squadre appertenenti a [4] e [12] hanno il valore dell'EBITDA negativo, che denota preoccupanti problemi di redditività. Anche in questo caso le società sono state inserite in gruppi diversi a seconda del loro valore dell'attività totale. Interessante infine notare che le società inserite nel *cluster* [4] presentano valori negativi per tasso di indebitamento e indice di solidità patrimoniale, sintomo di necessità di ricapitalizzazione;
- i gruppi [6], [7], [9] contengono società con situazioni economico-finanziarie disastrose: EBIT, EBITDA e risultato netto assumono valori profondamente negativi, con perdite di esercizio superiori ai 70 milioni di euro. La negatività degli indici ROA e ROS per queste squadre indica il rendimento negativo delle risorse investite nell'impresa, con i ricavi caratteristici incapaci di coprire anche i soli costi di gestione. La distinzione in 3 *clusters* diversi è imputabile al diverso valore dell'attivo patrimoniale delle società in questione.

CLUSTERS	EBITDA	EBIT	RIS. NETTO	TOT. ATTIVITÀ	ROS	ROA	EBITDA M.	T. INDEBIT.	I. SOL. PATR
[1]	5.499.624,87	-7.865.288,40	-5.268.774,33	85.681.173,60	-0,14	-0,09	0,10	40,12	7,35
[2]	22.388.917,00	5.591.492,70	1.669.036,70	108.469.503,30	0,08	0,05	0,30	29,06	4,39
[3]	6.816.976,40	-1.157.476,40	-1.994.542,80	33.452.501,00	-0,04	-0,10	0,17	10,45	2,38
[4]	-7.114.864,75	-39.509.918,75	-35.170.609,25	174.412.422,25	-0,39	-0,17	-0,07	-1,25	-0,53
[5]	14.713.112,33	-15.066.054,67	-6.536.002,00	232.818.330,33	-0,17	-0,06	0,16	72,44	25,65
[6]	-20.802.758,33	-81.502.112,33	-81.281.173,33	460.468.296,33	-0,35	-0,18	-0,08	-37,45	-13,28
[7]	-31.596.582,00	-92.154.792,00	-95.414.019,00	334.040.001,00	-0,54	-0,28	-0,18	-68,46	-24,46
[8]	33.690.268,80	9.071.130,20	3.389.605,10	158.655.287,50	0,07	0,06	0,29	5,78	0,42
[9]	-46.620.854,00	-103.868.122,50	-70.171.981,00	522.814.078,50	-0,47	-0,20	-0,22	48,90	34,34
[10]	49.382.016,33	1.228.373,33	-7.257.149,00	315.693.606,33	0,00	0,00	0,28	187,46	59,99
[11]	32.089.463,50	-22.497.189,50	-32.282.599,50	435.573.223,50	-0,10	-0,06	0,12	6,87	2,62
[12]	-38.995.798,00	-51.749.296,00	-38.128.752,00	84.505.250,00	-1,50	-0,61	-1,13	7,66	1,83
[13]	68.390.453,00	41.309.172,00	32.265.947,00	250.172.834,00	0,25	0,17	0,42	2,69	0,00

Table 4.8: Valori medi degli indici per i cluster trovati

Verrà ora valutata, in base al *cluster* occupato nel corso delle diverse stagioni, l'evoluzione della situazione economico-finanzaria delle società calcistiche:

- la Juventus ha migliorato la sua condizione, passando dal *cluster* [7] occupato nella stagione 2010/12 al *cluster* [11] per le stagioni successive. Stesso dicasi del Milan, che si sposta dal gruppo [9] al gruppo [10];
- l'Internazionale è l'unica società che staziona nello stesso gruppo (il numero [6]) per tutti e tre gli esercizi, evidenziando una condizione economico-finanziaria critica e preoccupante;
- la Fiorentina ha migliorato i suoi risultati economici, passando dal *cluster* [4] (occupato nelle stagioni 2010/11 e 2011/12) al numero [8] nel 2012/13;
- la Roma ha invece peggiorato la sua condizione, passando dal [1] al [4]: il motivo è forse da ricercare nel cambio proprietario e nei conseguenti investimenti effettuati dalla nuova proprietà. Anche il Parma ha acuito la sua posizione, lasciando il *cluster* [8] occupato nelle stagioni 2010/11 e 2011/12 per il numero [5];
- percorso inverso del Parma lo ha compiuto la Lazio, migliorando la sua condizione nell'ultima stagione considerata;
- Udinese e Napoli hanno leggermente migliorato la rispettiva condizione economica, comunque di gran lunga migliore tra tutte le altre società considerate: la prima, ha lasciato il gruppo [8] per formare un *cluster* a sé stante nel 2012/13, mentre il Napoli occupa il numero [8] dalla stagione 2011/12.

Un'ulteriore valutazione può essere effettuata confrontando i risultati della *cluster analysis* con i risultati sportivi conseguiti nelle stagioni considerate:

- Le società vincitrici del titolo (Milan nel 2010/11, Juventus nel 2011/12 e nel 2012/13) appartengono a gruppi (il [9] e [11]) caratterizzati da posizioni economico-finanziarie negative, con notevoli perdite nei rispettivi esercizi: ciò sembra avvalorare la tesi secondo cui per vincere il campionato è necessario un esborso economico importante da parte delle società;
- L'accesso alle posizioni valide per le competizioni europee non sembra legato a nessuna regola economica: nelle stagioni considerate, hanno ottenuto il pass per disputare la Champions League e l'Europa League sia società sane dal punto di vista economico (come Udinese, Napoli, Lazio) sia società che invece presentano situazioni di bilancio profondamentente negative (Inter e Roma);

- le 9 squadre retrocesse in serie B nelle stagioni analizzate appartengono tutte ai *cluster* [1], [2], [3]. Non sembra esserci correlazione tra i risultati sportivi e i risultati economici: il Cesena nel 2011/12 e il Siena nel 2012/13, pur presentando una buona situazione economico-finanziaria, sono retrocesse dalla massima serie in quella cadetta.

5 Analisi empirica: una rete neurale per le performance sportive delle società di Serie A

Nella seguente sezione della tesi si porrà l'obiettivo di sviluppare e testare un modello previsionale dei risultati delle squadre calcistiche partecipanti al campionato di serie A nelle ultime dieci stagioni disputate. A tal fine, verrà implementata una rete neurale artificiale che, prese come input alcune variabili di gioco raccolte da Opta, verificherà se tali variabili siano in grado di fornire previsioni adeguate riguardo al risultato sportivo delle società calcistiche considerate.

Verrà ivi presentato:

- il dataset utilizzato per l'analisi, il *training set* e il *test-set;*
- il modello di rete neurale artificiale scelto e le sue caratteristiche principali;
- il risultato dell'analisi.

Il modello stimerà, per ogni società considerata, un output numerico rappresentante il numero di punti conquistati in una stagione a seconda dei dati di input considerati. La fase predittiva sarà preceduta da una fase di addestramento in cui un *training set* composto da combinazioni di input e output consentirà alla rete di apprendere la relazione esistente tra di essi. Apprese le associazioni tra valori di ingresso e valore d'uscita, il modello sarà in grado di elaborare generalizzazioni previsionali anche su dati non ancora processati.

Successivamente, sarà valutata la potenzialità della rete come strumento di supporto decisionale in fase di formazione della rosa di calciatori: fissato un punteggio obiettivo per una determinata società, si determinerà se la squadra sia in grado di raggiungere la quota punti prefissata. In caso contrario, verranno modificati i valori di ingresso della società, inserendo al loro interno input di calciatori che la società possa permettersi in termini economici e sottraendo ai valori della rosa aggregata quelli corrispondenti dei

calciatori che perdono il posto a discapito dei nuovi acquisti. Verrà quindi stimato il punteggio che la rosa così modificata è in grado di ottenere.

5.1 Variabili utilizzate nel modello

All'interno del database Opta sono state scelte, secondo criteri di significatività, 24 variabili da inserire come output nella rete neurale artificiale, dettagliatamente descritte nel paragrafo 1.5:

- Goal: reti realizzate dalla squadra nel corso del campionato;
- Goal to shot ratio: la percentuale dei goal segnati sulla totalità di tiri effettuati;
- Shooting Accuracy: rapporto tra i tiri nello specchio di porta e i tiri complessivi;
- Penalty Success Rate: percentuale di realizzazione dei rigori concessi a favore;
- Assist: passaggi volontari che permettono ad un compagno di siglare una rete senza dover dribblare un avversario, eccezion fatta per il portiere;
- Key Passes: sono tutti quegli assist mancati, ovvero quelli che non hanno portato alla realizzazione di un goal;
- Passing Accuracy Second Half: percentuale di passaggi riusciti all'interno della metà campo avversaria;
- Passing Accuracy Final Third: percentuale di passaggi riusciti all'interno degli ultimi 25 metri della metà campo avversaria;
- Crosses & Corners Accuracy: percentuale dei cross e dei calci d'angoli riusciti sul totale calciati;
- Dribble Success Rate: percentuale dei dribbling riusciti sul totale tentati;
- Tackles Success Rate: percentuale di azioni di contrasto riuscite su quelle complessivamente tentate;
- Clearances, Block & Interceptions: raggruppa tutte le azioni difensive diverse dai *tackle;*
- Fouls Conceded in the Danger Area (inc Pens): indica il numero totale di falli concessi all'interno dell'ultima tre quarti campo, inclusa l'area di rigore;

- Fouls Won in the Danger Area (inc Pens): tutti i falli vinti all'interno della tre quarti campo avversaria, area di rigore inclusa;
- Penalties Conceded: calci di rigore concessi alla squadra avversaria;
- Red Cards: numero di cartellini rossi estratti dal direttore di gara verso giocatori della squadra;
- Goals Conceded: totale di goal concessi agli avversari nell'arco del campionato;
- Saves Made: azioni del portiere della squadra che hanno evitato la realizzazione di una rete da parte degli avversari;
- Saves to Shot Ratio: percentuale di parate effettuate sui tiri totali nello specchio della squadra avversaria;
- Clean Sheets: numero di volte in cui la squadra ha terminato l'incontro senza subire reti;
- Total Shot Conceded: numero totale di tiri concessi alle formazioni avversarie;
- Duels Won %: percentuale di duelli vinti dai giocatori della squadra nel corso del torneo;
- Opponent 2nd Yellow: sono tutti i giocatori avversari fronteggiati ed espulsi dall'arbitro per somma di ammonizioni nel corso del campionato;
- Opponent Reds: avversari espulsi del direttore di gara per rosso diretto nel corso dei match disputati.

Queste 24 variabili sono state raccolte per le società disputanti il campionato calcistico di Serie A per le ultime 10 stagioni (dalla stagione 2005/06 alla stagione 2013/14).

Il dataset non presenta dati mancanti. In particolare, Cagliari, Fiorentina, Inter, Lazio, Milan, Roma ed Udinese hanno disputato tutti i campionati analizzati; Chievo Verona,Juventus, Palermo,Parma e Sampdoria hanno partecipato a 9 stagioni in massima serie; Atalanta, Catania e Siena ad 8; Bologna, Genoa e Napoli hanno presenziato a 7 edizioni; Livorno a 6; Lecce, Reggina e Torino a 5; Empoli e Messina a 3 campionati; Ascoli, Bari, Brescia e Cesena a 2; Hellas Verona, Novara, Pescara, Sassuolo, e Treviso hanno disputato un solo campionato di serie A negli ultimi 10 anni.

L'output finale della di ogni società restituisce un valore numerico rappresentate il numero di punti conquistati un una determinata stagione.

Dopo aver descritto i dati, il passo successivo è la suddivisione del dataset originale in due parti:

- il *training set*, necessario alla fase di addestramento e apprendimento della rete;
- il *test-set*, nel quale saranno testate le capacità predittive apprese dalla rete durante la fase di apprendimento.

Ci si aspetta che, in caso di soddisfacente apprendimento della relazione esistente tra gli input, il risultato atteso della previsione sul *test-set* non si discosti di molto dai dati reali.

5.2 Caratteristiche della rete neurale artificiale implementata

La rete neurale è stata implementata, attraverso il sofware Visual Basic, con determinate caratteristiche riguardanti l'architettura, la tipologia di apprendimento e l'algoritmo di addestramento:

5.3 Il software Visual Basic

Visual Basic è un linguaggio di programmazione *event driven*[76] rilasciato per la prima volta da Miscrosoft nel 1991. Giunto alla sesta e definitiva edizione nel 1998, la sua sintassi deriva dal linguaggio BASIC[77].

Visual Basic è stato progettato per essere un linguaggio facile da imparare ad utilizzare e successivamente implementare. Alcune delle sue peculiarità sono:

- rapido sviluppo di applicazioni (RAD);
- realizzazione di interfacce GUI anche complesse;
- pratico accesso alle banche dati;
- creazione di controlli e oggetti Active X;
- il codice sorgente di base è liberamente accessibile;

[76] La programmazione ad eventi è un paradigma di programmazione dell'informatica in cui il flusso del programma è largamente determinato dal verificarsi di eventi esterni anziché eseguire istruzioni con percorsi fissati.

[77] Linguaggio di programmazione ad alto livello sviluppato presso l'Università di Dartmouth nel 1964.

- elabora in molte situazioni alla stessa velocità di altri linguaggi di programmazioni più formali, quali C e C++;
- Possibilità di eseguire un'applicazione senza effettuare una compilazione completa; in questo modo è possibile cambiare il codice e continuare l'esecuzione direttamente in fase di debug.

Attualmente, è al decimo posto del TIOBE Programming Community Index[78], indice che misura la popolarità dei linguaggi di programmazione.

5.4 I risultati dell'analisi

[78]la classifica è consultabile al link `www.tiobe.com/index.php/content/paperinfo/tpci/index.html`

Conclusioni

Nella presente tesi è stato affrontato un percorso storico che ha permesso di delineare lo scenario attuale in cui è inserita la massima competizione calcistica italiana.

Dai primissimi anni del Novecento, caratterizzati dalla nascita delle prime associazioni sportive nazionali, si è giunti a fine secolo alla formazione, grazie ai numerosi interventi a livello legislativo sia nazionale che comunitario, di società sportive costituite come società di capitali aventi scopo di lucro.

La situazione economico-finanziaria delle società, nonostante i tentativi di salvaguardia del legislatore e il boom dei ricavi da diritti televisivi, non ha subito miglioramenti significativi: il valore aggregato delle perdite di esercizio della Serie A ammontava nel 2013 a circa 200 milioni di euro.

In questo contesto, sono stati sviluppati due modelli: il primo misura la performance economica delle società sportive attraverso la costruzione di una *cluster analysis;* il secondo valuta i risultati sportivi delle squadre di serie A attraverso l'implementazione di una rete neurale artificiale.

Il modello di analisi economico-finanziaria suddivide le 60 società considerate in 13 *cluster* diversi. Analizzando i valori medi contenuti nei gruppi, si nota che 3 di essi contengono società (per un totale di 21) con valori di bilancio sani e in grado di produrre un risultato economico positivo; 26 società (divise in 4 *cluster* diversi) sono caratterizzate da perdite di bilancio lievi (nell'ordine degli 8 milioni di euro); 3 gruppi (7 squadre) hanno valori degli indici di redditività prossimi allo zero e perdite per circa 30 milioni di euro; gli ultimi 3 *cluster* denotano squadre in grave dissesto economico-finanziario, caratterizzate da valori di bilancio negativi e perdite superiori ai 70 milioni di euro.

Dai stessi risultati dell'analisi di raggruppamento sono state fatte alcune considerazione sull'andamento economico-finanziario delle società nel corso delle stagioni considerate, anche allo scopo di valutare l'impatto del fair play finanziario sul campionato di serie A. Degno di nota è il trend migliorativo di quasi la totalità delle grandi squadre.

Riferimenti bibliografici

Aldenderfer M.S. e Blashfield R.K., *Cluster analysis*, Sage Publications, Thousand Oaks, CA, 1984.

Andemberg M., *Cluster analysis for applications*, New York Academic Press, 1973.

Azzalini A. e Scarpa B., *Analisi dei dati e Data Mining*, Springer, Milano, 2004.

Bernoldi A., Sottoriva C., *La disciplina della redazione del bilancio di esercizio delle società di calcio. Confronto con l'esperienza internazionale e impatto del cd. «Financial fair play»*, in Rivista di diritto ed economia dello sport, Vol VII, Fasc 1, 2011.

Bianchi L. A. e Corrado D., *Bilanci delle società di calcio. Le ragioni di una crisi.* Egea, Milano, 2004.

Bodie Z, Kane A. e Marcus A. J., *Essentials of Investments*, McGraw Hill Irwin, 2004.

Bock H. H., *Origins and extensions of the k-means algorithm in cluster analysis*, Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany.

deMartini A., *La disciplina dei diritti televisivi nello spor*t, in Rivista di Economia dello Sport, Vol. VII, fasc. 2, 2011.

Diday E. et al, *Optimisation en classification automatique.* Vol. I, II. Institut National der Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France 1979.

Fabbris L., *Analisi esplorativa di dati multidimensionali*, Cleup editore, 1983

Floreano D. e Mattiussi C., *Manuale sulle reti neurali*, Il Mulino, Bologna, 2004.

Floreano D. e Nolfi S., *Reti neurali: algoritmi di apprendimento, ambiente di apprendimento, architettura*, in Giornale Italiano di Psicologia, a. XX, febbraio 1993.

Gore P. A., *Cluster Analysis*, in Handbook of Applied Multivariate Statistics and Mathematical Modeling, Southern Illinois University Academic Press, Carbondale Illinois, 2000.

Green P. E., Frank. R. E., Robinson P.J., *Cluster Analysis in text market selection*, Management science, 1967.

Hartigan J. A., *Clustering Algorithms*, Wiley, 1975.

Hebb D. O., *The organization of behaviour*, New York, Wiley and Sons, 1949.

Jajuga K., Sokolowski A. e Bock H. H., *Classification, clustering and data analysis: recent advances and applications*, Springer, Berlin, 2002.

Jardine N. e Sibson R., *Mathematical taxonomy*, Wiley,London, 1971.

Lago U., Baroncelli A. e Szimanski S., *Il business del Calcio. Successi sportivi e rovesci finanziari*, Egea, Milano, 2004.

MacQueen J., *Some methoods for classification and analysis of mulivariate observations*, in: L.M. LeCam, J. Neyman (eds.): Proc. 5th Berkely Symp. Math. Statist. Probab. 1965/66. Univ. of California Press, Berkely 1967, vol. I.

Maiuri G., *Un modo diverso di pensare calcio: L'approccio Sistemico e la Periodizzazione Tattica*, Youcanprint Self-Publishing editore.

Mancin M., *Il bilancio delle società sportive professionistiche. Normativa civilistica, principi contabili nazionali e internazionali (IAS/IFRS)*, CEDAM, Venezia, 2009.

McCulloch W. e Pitts W, *A logical calculus of the ideas immanent in nervous activity,* in Bulletin of Mathematical Biophysics, vol. 5.

Minsky M. e Papert S., *Perceptrons: an introduction to computational geometry,* The MIT press, Cambridge MA, 1969.

Romanato M., F*rancesco Gabrielli (1857-1899). Le origini del calcio in Italia: dalla ginnastica allo sport*, Antilia, Treviso, 2008.

Rosenblatt F., "*The Perceptron: A Probalistic Model For Information Storage And Organization In The Brain*", Psychological Review, 1958.

Rumelhart D. E. e McClelland J. L., *PDP. Microstruttura dei processi cognitivi*, Il Mulino, Bologna, 1991, trad. it di R. Luccio e M. Ricucci.

Singer W., *Activity-Dependant self organisation of synaptic connections as a substrate of Learning*, in The neural and molecular bases of learning, a cura di J.P. Changeaux e M. Konishi, Wiley, London.

Sokal R. R. e Sneath P. H., *Principles of numerical taxonomy*, Freeman, San Francisco - London, 1963.

Sostero U., Ferrarese O., Mancin M., Marcon C., *Elementi di bilancio e di analisi economico-finanziaria*, Libreria Editrice Cafoscarina, Venezia, 2011.

Stent G. S., *A physiological machanism for Hebb's postulate of learning*, in Proceedings of the National Academy of Sciences, vol.70.

Tryon R., *Cluster analysis*, McGraw Hill, New York, 1939.

Valeri M., *Standard IAS/IFRS e nuove esigenze di disclosure nel bilancio delle società di calcio*, Giappichelli editore, Torino, 2008.

Werbos P., *Beyond regression: new tools for prediction and analysis of behavioral sciences*, tesi di dottorato, Harvard University, 1974.

Widrow B. e Hoff M.E., *Adaptive switching circuits*, in IRE WECON Convention Record, Part IV, 1960.

Zazzaro G., *Data Mining: esplorando le miniere alla ricerca della conoscenza nascosta*, in Matematicamente.it, numero 9, maggio 2009.