Università degli studi di
Milano-Bicocca

F1801Q127

Modelli probabilistici per le decisioni

Predicting soccer results

Studenti:	*Matricole:*
Basso Matteo	807628
Ferri Marco	807130

Giugno 2019

Indice

1 Abstract

Il calcio è da anni lo sport più giocato e diffuso in svariati paesi del mondo. Moltissime persone seguono settimanalmente, con grande attenzione e fervore, ogni singola partita nella speranza di veder vincere la propria squadra del cuore.

In questo contesto si sono inevitabilmente sviluppati un gran numero di diversi business di enorme valore economico, fra cui ad esempio quello delle scommesse. Saper prevedere l'esito di una partita, sia in termini di vittoria che di goal fatti, rappresenta per moltissimi un argomento di particolare interesse. Durante questo studio, ci si è posti l'obiettivo di sviluppare un modello per la predizione del risultato delle partite di calcio in maniera automatica. Tale predizione deve essere basata sulla formazione delle due squadre considerate per ciascuna partita, nonché sulle caratteristiche dei giocatori coinvolti.

Il dominio di riferimento è stato modellato attraverso l'utilizzo di una Rete Bayesiana attraverso la quale fare le dovute inferenze e conseguentemente predire il risultato di una partita fra due squadre di calcio. Il medesimo approccio potrebbe essere riutilizzato in altri sport o giochi che prevedano l'interazione fra più squadre di cui si conoscono le statistiche di ciascun giocatore.

2 Introduzione

Viene qui presentata una visione generale del progetto, ovvero il dominio di riferimento e gli obiettivi che esso si pone, le scelte di design per la creazione del dataset ed eventuali ipotesi o assunzioni fatte durante lo sviluppo dell'elaborato.

2.1 Dominio di riferimento

Il calcio rappresenta, soprattutto negli ultimi anni, lo sport maggiormente diffuso in vari paesi del mondo. Molte persone seguono con grande attenzione tutte le partite cercando di capire preventivamente il vincitore, per piacere personale oppure per giocare nel mercato delle scommesse.

A seguito di questo fenomeno sono stati creati svariati portali che permettessero alle persone di entrare sempre di più in questo mondo. Un aspetto fondamentale per ogni piattaforma di scommesse è senz'altro la possibilità di predire il vincitore di un match attraverso la simulazione di partite di calcio con particolari formazioni.

2.2 Obiettivi dell'elaborato

Questo elaborato sarà suddiviso per argomenti, secondo un approccio di indagine incrementale e coerente con quanto praticamente svolto.

Verranno innanzitutto descritte le modalità di acquisizione dei dati dalla sorgente, a cui sarà associata anche una breve descrizione di quanto a disposizione. Quindi, si esploreranno i criteri e le assunzioni che hanno portato alla creazione del dataset su cui sviluppare il modello di predizione. Con questo, si intende dire che verranno inizialmente presentate alcune analisi qualitative sui dati e successivamente descritti i procedimenti svolti per l'integrazione e l'estrazione delle informazioni più rilevanti.

In seguito all'esportazione del dataset saranno descritte le modalità di suddivisione del dataset su cui fare inferenze, in particolare attraverso l'utilizzo delle Reti Bayesiane. Diversi modelli di generazione della rete verranno presentati e per ognuno saranno analizzati i risultati ed evidenziate le differenze. L'obiettivo del modello è quello di predire il team vincitore di un determinato match, utilizzando le informazioni dei giocatori presenti in campo.

Verrà infine presentata un'interfaccia utente per la simulazione di una partita di calcio, utile a mostrare l'effettivo impiego della rete in un prodotto di natura commerciale e pensato per l'utilizzo sul Web.

3 Descrizione del dataset originale

Di seguito viene presentata la descrizione del dataset originale, prelevato da Kaggle [2] e a sua volta creato attraverso l'unione di dati provenienti da diverse fonti, fra cui la più importante contenente le statistiche del famoso videogioco di calcio EA Sports FIFA. [1]. Il contenuto del dataset è rappresentato dalle principali squadre e giocatori di calcio in Europa, con le partite svolte in un periodo temporale di 6 anni.

3.1 Formato

Il dataset consiste di un database SQLite [5], ovvero un file con omonima estensione contenente un database relazionale di rapido utilizzo, in grado di essere condiviso e utilizzato facilmente su un qualsiasi dispositivo.

3.2 Giocatori

I giocatori sono espressi all'interno di due tabelle differenti nello schema del database considerato. La prima relazione denominata *Player* contiene informazioni di carattere generale ed è mostrata in tabella 1 a seguire.

Tabella 1: Player

Campo	Descrizione
id	Id del record
player_api_id	Identificativo univoco del giocatore
player_name	Nome completo del giocatore
player_fifa_api_id	Id utile per reperire le informazioni da fifa
birthday	Data di nascita
height	Altezza in centimetri
weight	Peso in libbre

La seconda tabella, denominata *PlayerAttributes* contiene invece i punteggi associati alle abilità personali di ciascun giocatore. Essa contiene inoltre un campo che indica a quale data tali punteggi si riferiscono, cioè quando sono state effettuate le rilevazioni. É possibile, infatti, che i dati associati ad un giocatore cambino nel tempo; per tale motivo, nel momento in cui si rivelerà necessario recuperare i dati dei giocatori facenti partecipanti ad una partita, sarà necessario selezionare i giusti record dalla tabella *Player* proprio in base alla data della partita.

Nella tabella 2 vengono meglio elencati i campi che descrivono un giocatore. La maggior parte di essi sono espressi su una scala intera da 1 a 100.

Tabella 2: PlayerAttributes

Campo	Descrizione
id	Id del record
player_api_id	Identificativo univoco del giocatore
player_fifa_api_id	Id utile per reperire le informazioni da FIFA
date	Data di riferimento del record
overall_rating	Valutazione generale
potential	Potenziale
preferred_foot	Piede preferito
attacking_work_rate	Punteggio di attacco
defensive_work_rate	Punteggio di difesa
crossing	Cross
finishing	Capacità di finalizzare (segnare un goal)
heading_accuracy	Tiro di testa
short_passing	Passaggio corto
volleys	Tiro al volo
dribbling	Dribbling
curve	Tiro con effetto curvilineo
free_kick_accuracy	Accuratezza dei calci piazzati
long_passing	Passaggio lungo
ball_control	Controllo di palla
acceleration	Accelerazione
sprint_speed	Velocità in scatto
agility	Agilità
reactions	Reattività
balance	Equilibrio
shot_power	Potenza di tiro
jumping	Salto
stamina	Resistenza
strength	Forza
long_shots	Tiri lunghi
aggression	Aggressività
interceptions	Intercettazione
positioning	Posizionamento
vision	Vista
penalties	Falli commessi
marking	Capacità di marcare
standing_tackle	Contrasto in piedi
sliding_tackle	Scivolata
gk_diving	Tuffo (portiere)
gk_handling	Presa (portiere)
gk_kicking	Rinvio (portiere)
gk_positioning	Posizionamento (portiere)
gk_reflexes	Riflessi (portiere)

3.3 Squadre

Le squadre sono descritte analogamente ai giocatori tramite 2 tabelle, *Team* 4 e *TeamAttributes* 3 riportate qui di seguito.

Tabella 3: Team

Campo	Descrizione
id	Id del record
team_api_id	Identificativo univoco della sqaudra
team_fifa_api_id	Id utile per reperire le informazioni da FIFA
team_name	Nome completo della squadra
date	Data di riferimento del record
buildUpPlaySpeed	Velocità di gioco
buildUpPlaySpeedClass	Discretizzazione velocità di gioco
buildUpPlayDribbling	Dribbling
buildUpPlayDribblingClass	Discretizzazione dribbling
buildUpPlayPassing	Passaggio
buildUpPlayPassingClass	Discretizzazione passaggio
buildUpPlayPositioningClass	Posizionamento della squadra, definito o libero
chanceCreationPassing	Creazione chance di passaggio
chanceCreationPassingClass	Discretizzazione creazione chance di passaggio
chanceCreationCrossing	Creazione chance di cross
chanceCreationCrossingClass	Discretizzazione creazione chance di cross
chanceCreationShooting	Creazione chance di tiro in porta
chanceCreationShootingClass	Discretizzazione creazione chance di tiro in porta
chanceCreationPositioningClass	Creazione chance di posizionamento
defencePressure	Pressing della difesa
defencePressureClass	Discretizzazione pressing della difesa
defenceAggression	Aggressività della difesa
defenceAggressionClass	Discretizzazione aggressività della difesa
defenceTeamWidth	Copertura della difesa
defenceTeamWidthClass	Discretizzazione copertura della difesa
defenceDefenderLineClass	Tipologia di difesa

Come osservabile, a differenza di quanto avvenga con i giocatori, per le squadre molti attributi appaiono già discretizzati attraverso una suddivisione in classi che descrive a parole la relativa scala da 1 a 100 di ciascun campo.

Tabella 4: Team

Campo	Descrizione
id	Id del record
team_api_id	Identificativo univoco della sqaudra
team_name	Nome completo della squadra
team_fifa_api_id	Id utile per reperire le informazioni da fifa
team_short_name	Nome abbreviato della squadra
team_long_name	Nome completo della squadra

3.4 Partite

La tabella *Match* contiene tutte le informazioni relative le squadre che hanno giocato una partita l'una contro l'altra, i giocatori coinvolti e il loro posizionamento in campo. I giocatori vengono numerati da 1 a 11 e il portiere viene sempre collocato in posizione [1;1]. Sono presenti inoltre dei valori il cui scopo è quello di indicare alcune specifiche tecniche sull'andamento della partita. Per brevità e mancanza di documentazione verranno semplicemente omessi, poiché poco rilevanti ai fini dello studio. Nella tabella 5 è presenta la descrizione dei vari campi.

Tabella 5: Match

Campo	Descrizione
id	Id del record
country_id	Id della nazione di riferimento
league_id	Id della lega di riferimento
season	Stagione calcistica
date	Data della partita
match_api_id	Id univoco della partita
home_team_api_id	Id del team in casa
away_team_api_id	Id del team in trasferta
home_team_goal	Goal del team in casa
away_team_goal	Goal del team in trasferta
home_player_i	Id dell'i-esimo giocatore in casa
away_player_i	Id dell'i-esimo giocatore in traferta
home_player_Xi	Posizione i-esimo gioc. (casa) sul lato corto del campo
home_player_Yi	Posizione i-esimo gioc. (casa) sul lato lungo del campo
away_player_Xi	Posizione i-esimo gioc. (trasferta) sul lato corto del campo
away_player_Yi	Posizione i-esimo gioc. (trasferta) sul lato lungo del campo

4 Creazione del dataset per le predizioni

Affinché i dati a disposizione possano essere utilizzati per la predizione del risultato di una partita su base statistica è necessario che vengano elaborati secondo un insieme di regole arbitrariamente scelte e su cui costruire i principi per portare a termine una specifica predizione. L'elaborazione stessa è perciò determinata sia attraverso delle assunzioni volte a identificare gli elementi più rilevanti nei dati, sia tramite l'effettiva esecuzione delle operazioni necessarie a trasformare il dataset originale in quello utilizzato per modellare la Rete Bayesiana.

4.1 Ipotesi e assunzioni

Al fine di sviluppare il progetto e procedere con la stesura del seguente elaborato, sono state fatte delle assunzioni che fungono da base portante.

Si assume che il risultato di una partita, a sua volta determinato dal numero di goal segnati da una squadra e dall'altra, sia influenzato dalla bravura dei giocatori che partecipano alla partita stessa. Nel gioco del calcio, come in molti altri, esiste una netta suddivisione fra i compiti assegnati a ciascun giocatore che vanno a determinare il ruolo di quest'ultimo: **portiere**, **difensore**, **centrocampista** e **attaccante**. É chiaro che ogni ruolo riveste la sua particolare importanza e influisce in maniera diversa sull'esito della partita. Per questo progetto, si è pensato di costruire un modello basato proprio sulle potenzialità delle due squadre sfidanti nei diversi ruoli assunti dai giocatori. Poiché tale informazione è mancante all'interno del dataset originale, il primo step cui si è fatto fronte è stata la determinazione dei ruoli di ciascun giocatore in campo. Tale obiettivo è stato raggiunto attraverso l'ipotesi, più che ragionevole, secondo la quale la posizione di una persona sul campo da gioco determina anche il suo ruolo all'interno della squadra e della partita.

In secondo luogo, si è scelto inoltre di supporre che le caratteristiche del team siano determinate solo dai giocatori che ne fanno parte. Non avendo a disposizione inoltre informazioni circa l'appartenenza di un giocatore ad un determinato team nel tempo, non è possibile ottenere una suddivisione dei giocatori in squadre ben formate. Nonostante quest'ultime si potrebbero ricavare direttamente dalle partite svolte, ciò comporterebbe la creazione di dati incoerenti col mondo reale poiché è risaputo che i giocatori di calcio sono soggetti ad un mercato di compra-vendita su base annuale. Per questo motivo, ai fini della predizione si è scelto di fatto di ignorare l'identità delle due squadre coinvolte in un match, ma semplicemente **limitarsi a considerare i 22 giocatori che hanno preso parte alla partita**, pesando adeguatamente le abilità che li contraddistinguono nel proprio ruolo.

4.2 Scelta del formato CSV

Per la creazione del dataset finale, utile per la modellazione della Rete Bayesiana, è stato scelto di trasformare il dataset in formato CSV, in quanto più semplice e pratico da utilizzare rispetto al database relazionale inizialmente fornito. Come si accennava, affinché possa essere creato il modello e su questo condotto delle previsioni, è necessario che i dati vengano elaborati e aggregati in un'unica tabella. Mentre SQLite consente agevolmente di lavorare sulle tabelle per effettuare *join*, selezioni e proiezioni sul modello relazionale, il DBMS non è particolarmente comodo per conservare i dati una volta terminata l'aggregazione. A tale scopo si è preferito quindi esportare successivamente il risultato ottenuto all'interno di un file CSV, decisamente più portabile e immediatamente leggibile rispetto ad un database.

4.3 Aggregazione dei dati

5 Rete Bayesiana

5.1 Strumenti e librerie

5.2 Assunzioni

5.3 Generazione della rete

5.3.1 Autogenerazione della struttura

5.3.2 Vincoli introdotti

5.3.3 Esperimenti

6 Performance della rete

6.1 Tempi

6.2 Cross Fold Validation

6.2.1 Accuracy

6.2.2 Precision e Recall

6.2.3 Loss

7 Web Demo

Al fine di dimostrare l'utilizzo della rete in un applicativo software, viene qui illustrato lo sviluppo e il funzionamento di un'applicazione web che permette la configurazione delle rose dei due team coinvolti in una partita e la conseguente predizione del vincitore.

7.1 Architettura

L'interfaccia web è stata sviluppata utilizzando l'architettura a 3 layer, separando frontend, backend e database.

Il database utilizzato risulta quello fornito inizialmente senza alcuna modifica. Esso consiste quindi in un file sqlite interrogabile e modificabile semplicemente tramite un web server. Esso risulta particolarmente utile per fornire i dettagli dei giocatori ed eventualmente dei team così che l'utente possa visualizzarli e sceglierli opportunamente.

Per lo sviluppo del backend è stato deciso di utilizzare l'engine javascript tramite il popolare progetto Node.js [3]. Esso è in grado di agire come middleware tra il frontend e il database, separando al meglio le logiche di manipolazione del dato. É inoltre incaricato di chiamare adeguatamente lo script R per la predizione del vincotore della partita e per svolgere l'inferenza.

Il fronted risulta sviluppato utilizzando la libreria javascript React.js [4]

7.2 Interfaccia grafica

Di seguito vengono mostrate brevemente le principali schermate dell'applicazione e il loro funzionamento.

All'avvio dell'app, utilizzando semplicemente un browser web, è possibile notare la schermata principale con 3 tab di selezione: *Home team*, *Away team* e *Results*. Mentre i primi 2 permettono di modificare le squadre, il terzo consente invece di visualizzare i risultati data la configurazione precedente.

La schermata di configurazione della squadra mostra sulla sinistra l'immagine di un campo da calcio su cui è possibile spostare i giocatori, semplicemente trascinandoli col mouse e opportunamente aggiunti tramite il pulsante sulla destra, creando così la formazione da utilizzare.

Una volta finita la configurazione cliccando sul tab dei risultati è possibile vedere in verde la squadra vincitrice e in rosso quella che invece è stata sconfitta.

Per non dover ricreare configurazioni dall'inizio ogni qualvolta si avvii l'applicazione inoltre, nella parte superiore destra dello schermo sono presenti dei tasti per salvare e caricarne di precedenti.

7.3 Predizione risultato

Per predire il vincitore della partita, come mostrato precedentemente, è stato necessario utilizzare la rete Bayesiana ottenuta a seguito degli esperimenti. Tale rete è stata infatti salvata su un file che è stato possibile poi ricaricare nell'ambiente R tramite l'apposito comando. Il web server javascript è dunque in grado di chiamare lo script R da riga di comando fornendogli in input una rappresentazione in formato JSON dei giocatori in campo con le loro caratteristiche. Lo script non deve far altro che manipolare i dati che gli sono stati forniti e applicare l'inferenza.

Per mostrare inoltre il puro funzionamento della rete è stato sviluppato un semplice form che permette l'inserimento delle evidenze e fornisce in output le distribuzioni di probabilità degli altri nodi.

8 Conclusioni

Lorem ipsum dolor sit amet.

Riferimenti bibliografici

[1] Ea sports fifa videogame.

[2] Kaggle dataset source.

[3] Node.js.

[4] React.

[5] Sqlite.