Big Data Analysis on NBA & NCAA

Colleges' impact on professionistic games

Marco Faretra, Gabriele Marini

12 luglio 2017

RIASSUNTO

Lo scopo di questo progetto è quello di produrre un ranking dei college americani sulla base delle performance dei primi anni di carriera dei giocatori professionisti. Allo scopo si sono analizzate tutte le statistiche dei giocatori NBA dai primi anni '50 fino ad oggi e si è semplificato il gioco fino ad individuare 6 categorie di giocatori, non mutuamente esclusive. Ad ogni giocatore viene assegnato un punteggio a seconda della particolare categoria che si sta analizzando, il punteggio di un college è la somma di tutti i punteggi dei giocatori provenienti da quel college, tenendo conto della particolare categoria analizzata. Lo scopo dell'analisi è quello di comprendere come alcuni college possano puntare su particolari aspetti del gioco, scelta che si dovrebbe riflettere nelle statistiche dei primi anni di carriera di un giocatore professionistico.

1 INTRODUZIONE

La lega di pallacanestro professionistica americana, meglio conosciuta come NBA, è stata una delle prime realtà a fruire dei dati dei suoi giocatori allo scopo di perfezionare questi ultimi e rendere le franchigie partecipanti sempre più competitive. Questo flusso di dati permette oggigiorno di avere a disposizione una quantità di dati immensa, delle tipologie più disparate, dalle statistiche dei singoli giocatori o delle singole partite, fino ad arrivare alle statistiche più dettagliate play-by-play [1].

Quello che non tutti sanno è che esiste un mondo dietro a quello della lega professionistica, altrettanto vasto, ovvero quello dei college e della NCAA. Prima di rendersi disponibili per il draft i giocatori tendono a svolgere qualche anno di preparazione in uno dei molti college americani. Questi, oltre che a fornire borse di studio ai giocatori, permettono loro di affinare gli aspetti legati al gioco, oltre che a dare loro una discreta visibilità agli occhi degli scout NBA.

Per questo progetto è di interesse proprio la correlazione tra gli insegnamenti del college e l'effettiva applicazione di questi nella sfera professionistica. Il nostro scopo è quello di analizzare, utilizzando tecniche Big Data, le statistiche di tutti i giocatori NBA dai primi anni '50 fino ad oggi, al fine di stilare una classifica dei college americani che hanno avuto nella loro storia almeno un giocatore che è riuscito a fare il salto di categoria nella lega professionistica.

A tale scopo si sono individuate 6 categorie di giocatori: tiratori (divisa a sua volta in tiratori da 2 e tiratori da 3), rimbalzisti, all-around, +/- guys, difensori, attaccanti. Per ognuna di queste categorie si è calcolato una score per ogni giocatore sulla base di alcune particolari statistiche tra le molte disponibili[2]. Ogni giocatore contribuisce, per la particolare categoria scelta, al punteggio totale del college di appartenenza. In questo modo college che curano di più una particolare caratteristica avranno un punteggio più alto se analizziamo la categoria legata a quella caratteristica. Una volta calcolato lo score per ogni categoria per ogni college è possibile stilare un ranking dei college sulla base delle categorie.

2 SETTING

Non avendo a disposizione un dataset già pronto si sono utilizzati i dati messi a disposizione da BasketBall-Reference, utilizzando delle procedure standard di estrazione dati da web si è ricavato un sottoinsieme dei dati offerti dal dominio. Una volta ottenuti i dati, questi sono stati riversati all'interno sia di un document store che all'interno di un key-value store, con le dovute differenze di modellazione.

3 APPROCCIO

Una volta estratti i dati, è stato scelto un document store, per salvare tutti i dati estratti. Il document

[1] https://www.bigdataball.com/nba-historical-playbyplay-dataset

[2] per i tiratori, ad esempio, è di fondamentale importanza la % del tiro, mentre le statistiche relative ai rimbalzi sono state ignorate

store è stato affiancato ad un key-value store per il salvataggio dei profili storici dei giocatori.

Una volta preparati i due stores, si può iniziare l'elaborazione dei dati vera e propria utilizzando Spark. Il software eseguibile prevede due modalità di esecuzione: locale e distribuita. Per brevità si analizzerà solamente la modalità locale [3].

Ad ogni giocatore viene associato uno score, questa operazione viene eseguita per ogni categoria definita. Prima i giocatori vengono filtrati utilizzando delle soglie, queste sono adattive, a seconda dell'andamento della stagione del giocatore correntemente analizzata possono variare ed essere più o meno permissive. Se il giocatore passa il filtraggio allora a seconda della categoria si tengono in considerazione solo alcune statistiche, queste vengono pesate e sommate tra loro in modo tale da ottenere il risultato finale per il giocatore:

$$\sum_{stats} player\_stat_i * w_i \pm b_i$$

Per alcune categorie si sono definiti anche alcuni bonus/malus per rendere lo score più particolareggiato.

Non vengono analizzate tutte le stagioni di un giocatore ma solamente le prime 4[4], questo perché dopo un certo periodo di permanenza nella lega professionistica alcuni giocatori acquisiscono alcune abilità tipiche dalla squadra per cui giocano, di fatto "eliminando" o comunque rendendo non più distinguibile la linea tra abilità collegiali e abilità acquisite successivamente.

4 SOLUZIONE TECNOLOGICA

Il linguaggio di riferimento scelto per orchestrare la computazione è Python, grazie alla sua estrema flessibilità e alla possibilità di prototipare una soluzione in breve tempo. Il motore di elaborazione è costituito principalmente da PySpark e Spark MLlib, sono presenti a contorno alcune librerie per il calcolo matematico come NumPy.

MongoDB è stato scelto per salvare i dati estratti, permette un veloce inserimento dei documenti all'interno del database, soprattutto se questi non hanno necessità di essere messi in relazione, oltre a queste motivazioni, l'ampia diffusione di questo sistema porta con se anche numerose librerie e connettori di supporto. Per quanto riguarda il salvataggio dei dati nel key-value store, il sistema scelto è stato Redis, fondamentalmente a causa della larga diffusione che porta con se documentazione e librerie di supporto. Questo sistema è stato utilizzato per salvare i profili dei giocatori, legando il profilo storico di un giocatore al suo identificatore.

A seguito della preparazione è necessario un ultimo passaggio per completare il setup: con l'aiuto di Spark MLlib si calcolano media, varianza, massimo e minimo per ogni statistica in una particolare stagione; il procedimento viene ripetuto per ogni stagione disponibile. Per alcune statistiche questo procedimento può essere più complicato: alcuni aspetti del gioco sono variati nel corso degli anni, ad esempio dal 1979 è stato introdotto il tiro da 3 punti, oppure alcune statistiche non sono rese disponibili prima di una certa stagione. Queste anomalie nei dati costringono a ripensare la gestione delle statistiche, in particolar modo al fine di poter confrontare giocatori moderni con quelli del passato.

Per parallelizzare i dati verso i worker si possono utilizzare entrambi gli stores, tuttavia MongoDB è fornito di connettore apposito per Hadoop, mentre redis non utilizza un connettore ma parallelizza i record utilizzando più volte la primitiva "parallelize()". Si può definire inoltre l'indirizzo ip della macchina su cui risiedono gli stores, in maniera da cambiare la posizione del server principale secondo bisogno.

Se si è scelto MongoDB come fornitore dei dati, i worker prendono i dati dal server principale utilizzando il connettore Hadoop-MongoDB, realizzando un RDD che rappresenta una collezione di documenti; se invece si è scelto di utilizzare Redis verranno creati vari RDD e poi uniti a mano a mano fino a formare un singolo RDD. La computazione avviene come descritto in precedenza su ogni singolo RDD.

Alla fine della computazione, se si è scelto di analizzare i giocatori, si ottengono una serie di coppie rappresentanti al primo membro gli identificatori dei giocatori ed al secondo membro lo score associato alla categoria scelta prima di lanciare l'analisi. Se invece si è scelto di analizzare i college si otterranno una serie di triple, in cui il primo membro è il nome del college, il secondo è lo score associato alla particolare categoria scelta, mentre il terzo è il numero di giocatori che hanno frequentato quel college.

All'interno del progetto sono disponibili dei banali strumenti per eseguire alcune aggregazioni di base una volta calcolati tutti gli score. Inoltre è disponibile, in allegato al progetto, una visualizzazione realizzata a partire dai risultati ottenuti al termine delle varie esecuzioni: questa permette di vedere i dati ad un diverso livello di aggregazione stato -¿ college -¿ giocatore, oltre alle precedenti aggregazioni è possi-

[3] La modalità distribuita è molto simile, l'unica differenza è la distribuzione delle dipendenze ai vari worker. Le altre variazioni vengono gestite mediante delle opzioni passate allo script principale.

[4] Poiché la carriera professionistica di un giocatore dura mediamente 4.8 anni

bile anche esplorare la carriera di un singolo giocatore in forma compatta, analizzando le statistiche per partita di ogni stagione giocata.

5 RISULTATI

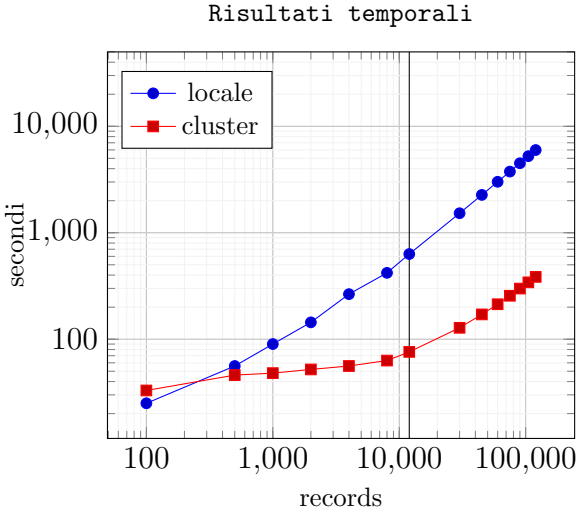
In calce è possibile trovare i risultati dei primi 10 giocatori e dei primi 10 college per ogni categoria.

Di seguito invece i risultati temporali ottenuti facendo partire un'analisi dei college sulla categoria "attaccanti", utilizzando 1 master e 19 nodi (task).[5] La macchina utilizzata per i test in locale è la seguente:

- MacBook Pro mid 2012
- CPU: 2.5GHz Intel Core i5
- 12 GB Ram
- 500 GB HDD

La tipologia di macchine utilizzate per eseguire i test sul cluster è la seguente:

- m3.xl
- 4vCPU
- 15GB RAM
- 40GB x 2 SSD



I tempi sono stati ottenuti analizzando fino a 12000 giocatori ovvero 407M di statistiche atomiche[6], i valori superiori a questa soglia sono stati previsti utilizzando i valori precedentemente ottenuti e una funzione di regressione lineare. Già dall'analisi di poche centinaia di record il tempo di esecuzione del cluster risulta inferire al tempo di esecuzione in locale.

6 CONCLUSIONI

Dai risultati possiamo evincere che esistono effettivamente dei college migliori di altri dal punto di vista della qualità dei giocatori che escono da questi ultimi, l'analisi storica eseguita dimostra che i college "migliori", sono stati sempre i più affidabili se si vuole eseguire il salto di qualità nella sfera professionistica.

Alcuni giocatori riconosciuti attualmente come ottimi giocatori risultano avere una posizione abbastanza bassa in classifica, questo è dovuto molto probabilmente al fatto che alcuni di questi giocatori hanno migliorato le loro caratteristiche di gioco più avanti nella carriera e quindi risultano avere uno score basso se si analizzano solo i primi anni di carriera.

[5] Tutte le categorie hanno una modalità di calcolo molto simile, le differenze temporali non sono apprezzabili cambiando categoria analizzata.

[6] Questo limite è rappresentato dalla linea verticale presente nel grafico all'ascissa 12000

Giocatore	*Plus Minus Score*
Draymond Green	100.0
Tayshaun Prince	98.3
Manu Ginobili	97.16
Josh Howard	94.34
Kawhi Leonard	93.89
Tony Parker	91.71
Rajon Rondo	90.33
Dwyane Wade	88.48
Lance Stephenson	87.92
Klay Thompson	86.26

Giocatore	*Rebounds Score*
Wilt Chamberlain	100.0
Bill Russell	89.19
Jerry Lucas	74.47
Wes Unseld	66.6
Maurice Stokes	66.28
Elgin Baylor	65.52
Nate Thurmond	65.07
Walt Bellamy	64.3
Elvin Hayes	62.57
Moses Malone	60.67

Giocatore	*3 point shooters Score*
Jason Kapono	100.0
Stephen Curry	98.3
Wally Szczerbiak	97.92
Mark Price	97.84
Drazen Petrovic	96.77
Kyle Korver	96.24
Kelenna Azubuike	96.12
Trent Tucker	95.7
Bobby Simmons	95.1
Jose Calderon	95.08

Giocatore	*Total Score*
Wilt Chamberlain	455.33
Karl-Anthony Towns	444.24
Larry Bird	439.99
Shawn Marion	439.79
Bill Walton	427.93
Moses Malone	427.66
Charles Barkley	427.62
Kevin Love	420.92
LeBron James	415.09
Kawhi Leonard	415.07

Giocatore	*All around Score*
Wilt Chamberlain	100.0
Bill Russell	89.58
Jerry Lucas	77.01
Moses Malone	71.45
David Robinson	70.52
Bill Walton	69.84
Wes Unseld	68.81
Elgin Baylor	68.08
Maurice Stokes	68.08
Walt Bellamy	66.91

Giocatore	*2-point shooters Score*
Charles Barkley	100.0
John Stockton	99.76
Karl-Anthony Towns	98.09
Jose Calderon	96.38
Kiki Vandeweghe	95.59
Reggie Miller	94.7
Cedric Maxwell	94.67
Arron Afflalo	94.26
Otto Porter	94.07
Kawhi Leonard	93.51

Giocatore	*Attackers Score*
Wilt Chamberlain	100.0
Elgin Baylor	74.58
Kareem Abdul-Jabbar	73.49
Michael Jordan	71.99
Rick Barry	71.7
Oscar Robertson	70.81
Bob McAdoo	66.87
George Gervin	66.21
Walt Bellamy	65.2
Elvin Hayes	64.25

Giocatore	*Defenders Score*
Bill Walton	100.0
Moses Malone	74.84
David Robinson	71.91
Larry Bird	69.71
Artis Gilmore	69.35
Tim Duncan	67.11
Shawn Marion	64.31
Kevin Love	64.15
John Shumate	63.96
Larry Kenon	63.74

College	*Plus Minus Score*
Marquette University	100.0
Wake Forest University	99.97
San Diego State University	96.85
University of Illinois at Urbana-Champaign	96.41
University of Oklahoma	94.84
Washington State University	94.17
Virginia Commonwealth University	93.82
Augsburg College	92.27
Louisiana Tech University	92.17
University of California	91.66

College	*Rebounds Score*
University of California, Los Angeles	100.0
University of Kentucky	96.67
University of North Carolina	77.4
Duke University	68.72
University of Kansas	66.13
Ohio State University	54.33
Indiana University	50.6
University of Louisville	50.35
University of Michigan	46.79
University of Arizona	45.74

College	*3 point shooters Score*
University of California, Los Angeles	100.0
University of North Carolina	93.1
University of Kentucky	80.48
University of Arizona	74.37
Duke University	70.07
Ohio State University	57.75
Georgia Institute of Technology	54.15
Indiana University	51.82
University of Michigan	51.63
University of Connecticut	51.33

College	*Defenders Score*
University of California, Los Angeles	100.0
University of Kentucky	95.16
University of North Carolina	85.63
Duke University	81.23
University of Arizona	61.42
University of Kansas	53.6
University of Michigan	49.1
University of Connecticut	45.57
Georgetown University	43.82
Indiana University	43.27

College	*All around Score*
University of California, Los Angeles	100.0
University of Kentucky	99.43
University of North Carolina	81.02
Duke University	73.44
University of Kansas	64.36
Indiana University	52.18
Ohio State University	51.0
University of Arizona	50.5
University of Louisville	49.94
University of Michigan	48.35

College	*2-point shooters Score*
University of California, Los Angeles	100.0
University of North Carolina	81.61
University of Kentucky	70.93
Indiana University	62.81
Duke University	53.56
Ohio State University	51.94
University of Arizona	51.06
University of Michigan	46.01
University of Kansas	42.31
St. John's University	42.06

College	*Attackers Score*
University of Kentucky	100.0
University of California, Los Angeles	97.22
University of North Carolina	95.33
Duke University	73.22
Indiana University	60.11
University of Kansas	58.42
Ohio State University	54.94
University of Michigan	50.88
University of Arizona	46.2
Syracuse University	45.52

College	*Total Score*
University of California, Los Angeles	663.6
University of North Carolina	573.92
University of Kentucky	573.66
Duke University	420.24
University of Kansas	392.67
Indiana University	385.18
University of Arizona	375.97
Ohio State University	369.96
University of Michigan	338.17
University of Notre Dame	310.25