Zastosowanie technik zgłębiania danych w przewidywaniu pozycji gracza NBA

Jan Iwaszkiewicz*

Uniwersytet Gdański

16 grudnia 2018

Streszczenie

W poniższym opracowaniu przedstawione zostają klasyfikatory oraz inne metody zgłębiania danych pozwalające na określanie pozycji gracza ligi NBA. Przy wykorzystaniu statystyk z danego sezonu sprawdzona zostaje skuteczność przedstawionych algorytmów. Dodatkowo zostaje opisany sposób doboru danych oraz ich potencjalny wpływ na wyniki badań.

I. Wstęp - opis bazy

NBA to amerykańska liga koszykówki mężczyzn, najbardziej znana i dochodowa na świecie. Swój suckes zawdzięcza ciągłemu rozwojowi i dostosowywaniu się do potrzeb fanów sportu. Idzie za tym rozwój samych graczy, którzy muszą co wieczór zapewniać najwyższy poziom rozrywki. Od sportu dla studentów i amatorów do światowego widowiska najwyższej klasy, od „białego" do „czarnego" sportu - tak można opisać rozwój koszykówki oraz NBA. Celem przeprowadzonego eksperymentu będzie sprawdzenie, czy w danym roku możliwe jest sklasyfikowanie pozycji gracza na boisku. Zgodnie z założeniami sportu niektóre pozycje jak rozgrywający (Point Guard) czy środkowy (Center), mają znacząco różniące się zadania podczas rozgrywki. Idące za tym różnice w statystykach graczy powinny jasno wyznaczać poszczególne pozycje jako klasy. Czy na pewno?

*indeks: 238215, kontakt: jiwaszkiewicz6@gmail.com

II. Przygotowanie danych

i. Czy sezon ma znaczenie?

By w pełni wykorzystać bazę danych konieczne jest przygotowanie jej pod wymagane przez algorytmy dane wejściowe. Na początku projektu założone zostało, że rok 1980 zostanie najstarszym możliwym do wybrania rokiem. Dlaczego? W tym roku dopiero wprowadzono linię rzutów za 3 punkty oraz rozpoczęto prowadzenie szczegółowych statystyk podczas sezonu. Wcześniej nie prowadzono takowych ze względów technicznych oraz słabego zainteresowania sportem. Zatem jaki rok wybrano do badań? Tutaj wybór padł na 2017. Oprócz dokładnych statystyk, ważnym czynnikiem jest stan ligi. Sport ciągle ewoluuje, a mniej więcej od roku 2015 NBA przeżyło ogromne zmiany. Zaczęto wykorzystywać skutecznie linię rzutu za 3 punkty oraz pojawiło się dużo graczy „hybrydowych", którzy nie są typowi dla swoich pozycji. Często są na tyle uniwersalni, że grają na pozycji, która aktualnie jest ważna dla danego zespołu lub dzięki temu uzysukją niesamowitą przewagę (przykładowo wzrostu) nad innymi graczami. Ostatecznie do badań wykorzystuje się jedynie jeden sezon. Dysproporcje między statystykami różnych sezonów

mogą być na tyle duże, że określenie klasy będzie wręcz niemożliwe. Przykładem może być rok 1980 oraz 2016, gdzie w pierwszym oddawnano rzuty za 3 punkty okazyjnie w ostateczności, natomiast w drugim stało się to jednym z podstawowych elementów gry.

ii. Statystyki danego gracza

Wybrane bazy połączono i wykonano na nich odpowiednie operacje, by uzyskać następujące statystki:

- **Player**
  Imię i nazwisko gracza, przycięte o niepotrzebne znaki.
- **Height**
  Wzrost gracza, podany w centymetrach.
- **Weight**
  Waga gracza, podana w kilogramach.
- **PTSavg**
  Średnia punktów z sezonu, ilość punktów ogółem podzielona na ilość rozegranych meczy.
- **TRBavg**
  Średnia zbiórek z sezonu, ilość zbiórek ogółem podzielona na ilość rozegranych meczy.
- **ASTavg**
  Średnia asyst z sezonu, ilość asyst ogółem podzielona na ilość rozegranych meczy.
- **STLavg**
  Średnia przechwytów z sezonu, ilość przechwytów ogółem podzielona na ilość rozegranych meczy.
- **BLKavg**
  Średnia bloków z sezonu, ilość bloków ogółem podzielona na ilość rozegranych meczy.
- **TOVavg**
  Średnia strat z sezonu, ilość strat ogółem podzielona na ilość rozegranych meczy.
- **TS.**
  True shooting precentage, statystka prowadzona by dokładniej określić skuteczność w oddawaniu rzutów.
- **FG.**
  Field goals precentage, skuteczność rzutów z gry.
- **X3P.**
  Skuteczność rzutów za 3 punkty.
- **X2P.**
  Skuteczność rzutów za 2 punkty.
- **eFG.**
  Effective field goal percentage, poprawiona statystyka z uwagą na „wartość” rzutu (3 oraz 2 punkty).
- **FT.**
  Skuteczność rzutów z linii rzutów osobistych.
- **Pos**
  Pozycja gracza na boisku, w wypadku wielu pozycji (gdy gracz w jednym sezonie przechodził między drużynami) wybrana zostaje pozycja na której rozegrał najwięcej spotkań.

Wszystkie statystki zostają dodatkowo wyselekcjonowane, bądź poddane innym operacjom w wypadku korzystania z poszczególnych algorytmów. Przy wystąpieniu wartości niemożliwych do zdefiniowania NA lub -INF, dane zastępowano średnimi ligowymi, bądź zerami w zależności od potrzeb.

iii. Pozycja na boisku jako klasa

By możliwe ułatwić pracę oraz dopasować się do założeń projektu kolumna z pozycją gracza upraszcza się do decyzji „yes” lub „no”. Pozycję którą chcemy poddać klasyfikacji ustalamy za pomocą zmiennej „desired position”. Dostępne prametry to „PG”, „SG”, „SF”, „PF” oraz „C”.

III. Klasyfikatory

i. Podział na zbiory

Do klasyfikatorów C4.5/ID3 (drzewo), Naive Bayes, kNN niezbędny jest podział danych na zbiory treningowe i testowe. Dokonano go za pomocą losowego wybrania rekordów przy prawdopodobieństawch 67% oraz 33%. Do metody kNN znormalizowane zostały również wszystkie kolumny z wartościami numerycznymi.

ii. Porównanie klasyfikatorów

Na przygotowanej bazie dokonano klasyfikacji za pomocą następujących metod.

- **C4.5/ID3 - drzewo decyzyjne**
- **Naive Bayes**
- **kNN - k najbliższych sąsiadów** Przy użyciu dwóch najbliższych sąsiadów.
- **Hierarchical Clustering** Metoda pokrewna do k-średnich. Jednak wykazująca większą skuteczność dla wybranej bazy.

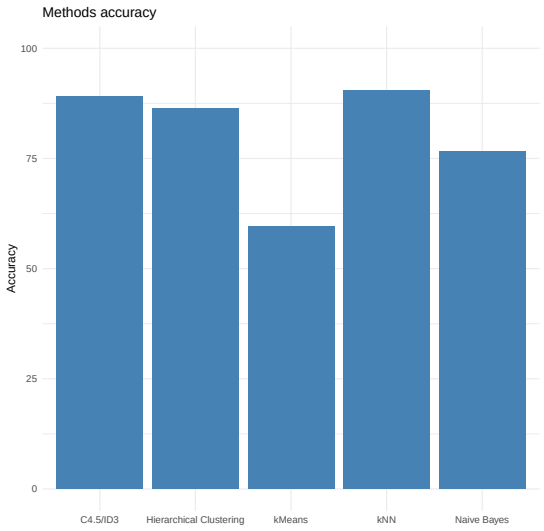
Oprócz wymienionych metod do wykresu przedstawiającego skuteczność klasyfikacji, zestawiona została również metoda k-Means. Pozwala to porównać ją z pochodzącym z tej samej rodziny sposobem Hierarchical Clustering. Wyniki widoczne są na Rysunku 1. Wyznaczone zastały również macierze błędu dla metod (Tablice 1, 2, 3 oraz 4). Przeprowadzone poza zakresem sprawozdania testy zwracały podobne wyniki dla większości sezonów, poza okresem (1996-2000) dla pozycji PG gdzie dochodziło do spadku skuteczności. Również pozycje „bardziej płynne" (SG, SF, PF), które często są wymienne i mają średnie statystki zbliżone do siebie, wykazywały się niższą skutecznością klasyfikacji. Szczególnie widoczne jest to w „nowoczesnej erze" NBA, rozpoczynającej się od połowy lat 90, aż do niedawno zanotowanego sezonu 2017.

Tablica 1: *Macierz błędu : C4.5/ID3 - drzewo decyzyjne*

Pred/Real	no	yes
no	118	9
yes	8	19

Tablica 2: *Macierz błędu : Naive Bayes*

Pred/Real	no	yes
no	97	7
yes	29	21



Rysunek 1: *Porównanie skuteczności w klasyfikowaniu graczy dla pozycji PG na podstawie danych z roku 2017.*

Tablica 3: *Macierz błędu : kNN - k najbliższych sąsiadów*

Pred/Real	no	yes
no	114	7
yes	7	19

Tablica 4: *Macierz błędu : Hierarchical Clustering*

Pred/Real	no	yes
no	342	18
yes	48	78

iii. TP, FP, TN, FN, THC, TPR, FPR, NBA, UG i inne skróty

Wartości TP, FP, TN, FN odpowiadają odpowiednio Prawdziwie Pozytywnemu, Fałszywie Pozytywnemu, Prawdziwie Negatywnemu oraz Fałszywie Negatywnemu zaklasyfikowaniu gracza na zadanej pozycji. Wyliczone zostały również wartości TPR, FPR, FNR oraz TNR (znane jako sensitivity, fall-out, miss rate oraz specificity). Zauważono również zależności między nimi:

$$FPR + TNR = 1$$

$$TPR + FNR = 1$$

Wydzielono dwa rodzaje błędu:

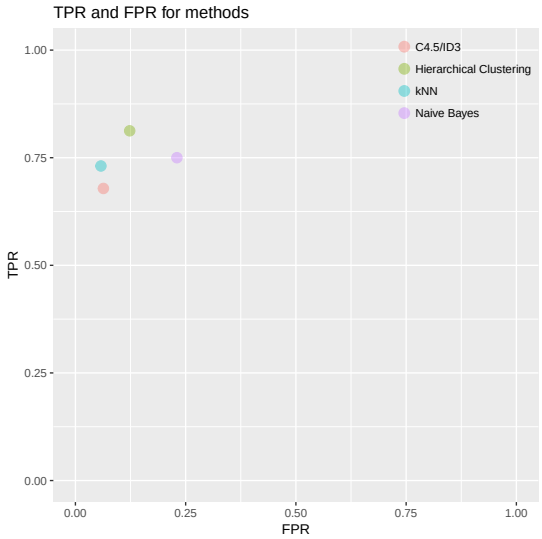
- **Pierwszego rodzaju**
  Czyli FP, błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości nie jest fałszywa.
- **Drugiego rodzaju**
  Oznaczony przez FN, błąd polegający na nieodrzuceniu hipotezy zerowej, która jest w rzeczywistości fałszywa.

Zakładamy, że szukamy rozgrywających. Zatem TP odpowiada poprawnemu przydzielniu klasy yes dla gracza będącego na pozycji rozgrywającego, FP przydzieleniu klasy yes dla gracza występującego przykładowo jako niski skrzydłowy, TN przypisaniu no dla rzucającego obrońcy oraz FN przydzieleniu no dla gracza będącego rozgrywającym.
Oba błędy mają wpływ na wynik czynników TPR, FPR, FNR i TNR. Obserwacje można wyciągnąć już ze wzorów (literatura „Sensitivity and specificity"). Poniżej jednak przedstawiona zostanie tabelka (Tablica 5) wyjaśniająca zależności:

Tablica 5: *Zależności wynikające z błędów pierwszego i drugiego rodzaju*

		Wartość			
		TPR	FPR	FNR	TNR
FP	↑	-	↑	-	↓
	↓	-	↓	-	↑
FN	↑	↓	-	↑	-
	↓	↑	-	↓	-

Osobistym przemyśleniem jest stwierdzenie, że błąd drugiego rodzaju (FN) jest bardziej „krzywdzący" dla obserwowanego zbioru danych. Zaklasyfikowany w ten sposób gracz (czyli w tym przypadku rozgrywający) nie jest najczęściej przygotowany do innej roli na boisku niż dowodzenia drużyną. Natomiast popularną metodą w dzisiejszej koszykówce jest wystawianie niskiego skrzydłowego jako rozgrywającego, który może w ten sposób rozciągnąć grę, jednocześnie posiadając podobny zestaw umiejętności co rozgrywający. Zatem błąd pierwszego rodzaju nie jest koniecznie krzywdzący, gdyby jakaś drużyna starała się dobierać gracza na podstawie wyników klasyfikatora.



Rysunek 2: *Porównanie sensitivity i false alarm dla klasyfikatorów.*

Na wykresie (Rysunek 2) przedstawione jest położenie klasyfikatorów w zależności od wyników FPR i TPR. Zakładając, że najlepszy klasyfikator nie popełnia błędów, idealna sytuacja zachodzi przy braku błędów. Znaczy to, że FP i FN są równe 0. Wtedy wartość TPR równa się 1, natomiast FPR wynosi 0. Zatem najlepszy klasyfikator znajduje się w punkcie (x = 0, y = 1), mówiąc potoczniej - w lewej górnej części wykresu. Odległości poszczególnych klasyfikatorów zostały obliczone i porównane (Tablica 6).

Tablica 6: *Odległość metody od współrzędnej (0,1) na Rysunku 2*

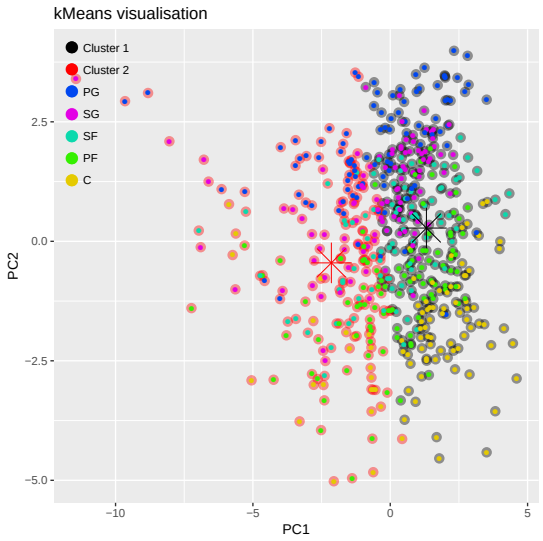
Metoda	Odległość
Hierarchical Clustering	0.2242859
kNN	0.2753761
C4.5/ID3	0.3276394
Naive Bayes	0.3620692

Jeżeli zwracamy uwagę jedynie na liczby najlepszym rozwiązaniem wydają się być Hierarchical Clustering oraz kNN, ponieważ znajdują się nabliżej założeń klasyfikatora idealnego. Jednak biorąc pod uwagę stabliność (szczególnie w innych latach) algorytm kNN lub C4.5/ID3 wydaje się być lepiej przystosowany do zadania.

IV. Grupowanie metodą k-średnich

i. Metoda nieskuteczna

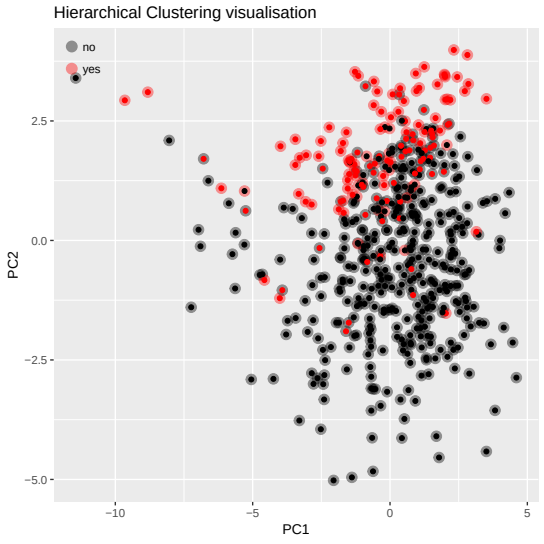
Ze względu na naturę bazy metoda nie potrafiła osiągnąć pożądanych rezultatów. Klasa zazwyczaj jest niewyraźnie wydzielona od reszty oraz niewielka w stosunku do ogółu (zwyczajowo w sezonie 20% zawodników jest odpowiednio przydzielonych do każdej z pozycji). Algorytm przedstawia się następująco (Rysunek 3). Widoczne centroidy dzielą ligę dokładnie w połowie. Inną możliwą przeszkodą, dla działania algorytmu, mogą być przeprowadzone operacje poprzedzające wprowadzenie danych wejściowych (log). Być może powodują zbyt duże znormalizowanie danych.



Rysunek 3: *Metoda k-średnich : klastry oraz nałożona na nią liga, sezon 2017.*

ii. Czy można to zrobić inaczej?

Obserwując wyniki metody k-means, postanowione zostało znaleźenie skutecznego rozwiązania dla problemu klastrowania. Wspomniane wcześniej Grupowanie Hierarchiczne (Hierarchical Clustering) opiera się na dzieleniu obserwacji na klastry bazując na podobieństwach między nimi. Nie jest tutaj konieczne określenie ilości klastrów na początku działania. Zatem bardzo możliwe wydaje się znalezienie dwóch grup, które z założeń projektu powinny być łatwe do wydzielenia ze zbioru obserwacji. Do obliczeń potrzebnych w algorytmie skorzystano z metody Połączeń Ward'a pod nazwą ward.D2. Korzysta ona z obserwacji, dzięki której nie jest wymagane podniesienie do kwadratu wszystkich odchyleń od punktów do centroidów. W procesie tworzą się pary zawierające pojedyncze obserwacje. Następnie pary klastrów są sukcesywnie łączone, aż do momentu gdy wszystkie klastry zostaną scalone w duży klaster zawierający podobne obiekty. Wyniki przedstawione zostały na Rysunku 4. Algorytm deprymuje pokrewną metodę k-średnich oraz prezentuje skuteczność zbliżoną do pozostałych metod (Rysunek 1).



Rysunek 4: *Metoda Hierarchical Clustering : klastry oraz nałożona na nią liga, sezon 2017, pozycja PG.*

V. Reguły asocjacyjne

Pomimo uproszczenia danych dwie wartości „above", „below" (powyżej i poniżej średnich ligowych) oraz wykluczeniu zbędnych lub powtarzających się, wyciągnięte z bazy reguły asocjacyjne można liczyć w setkach lub tysiącach. Wybrane zostały zatem tylko najbardziej interesujące z wysokim czynnikiem wiarygodności oraz wsparcia. Poniżej opisane kilka z nich:

- **Wynik: Pos=yes dla PG**
  Zgodnie z założenieniami rozgrywający nie zdobywa wielu punktów oraz nie gra pod koszem przez swój wzrost i wagę (poniżej średniej Height, Weight, TRBavg oraz BLKavg). Opiera swoją grę na asystach i przechwytach. Odznacza się również dobrym % z linii osobitych. Przykłady:
  - PTSavg=bellow
    ASTavg=above
    STLavg=above
    BLKavg=bellow
    X3P.=bellow
    FT.=above

  - Height=bellow
    ASTavg=above
    X3P.=bellow
    FT.=above
- **Wynik: Pos=no dla PG**
  Reguły dla pozostałych pozycji (najczęściej wyższych, silniejszych i grających bliżej kosza) również nie zaskoczyły podczas eksperymentu. Przykłady:
  - Height=above
    Weight=above
    TRBavg=above

  - TRBavg=above
    BLKavg=above
    X3P.=bellow
    FT.=above

VI. Podsumowanie badań

Po przeprowadzonych badaniach można zauważyć ewolucję koszykówki oraz graczy tworzących ligę NBA. Z biegiem czasu, różnice pomiędzy rolami na boisku coraz bardziej się zacierają. Nowe taktyki oraz wykorzystywanie statystyki w celu optymalizacji wyników drużyny, powodują drastyczne zmiany w założeniach koszykówki. W dniu dzisiejszym największą przeszkodą dla algorytmów jest uniwersalność obecnych w lidze zawodników. Jednak wciąż możliwe jest stosunkowo skuteczne sklasyfikowanie gracza, który odzacza się specjalną rolą na boisku (rozgrywający). Największym zaskoczeniem eksperymentu okazało się wykorzystanie innej niż przedstawionej podczas laboratoriów metody klasteryzacji. Pomimo swoich wad, przy odpowiednich danych, jest w stanie konkurować z innymi alogrytmami klasyfikującymi.

Literatura

[Zbiór danych] *NBA Players stats since 1950*, dostęp online: `https://www.kaggle.com/drgilermo/nba-players-stats`

[Zasoby online] *Sensitivity and specificity* , dostęp online: `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`

[Zasoby online] *How to understand the drawbacks of K-means* , dostęp online: `https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means`

[Zasoby online] *Hierarchical Clustering* , dostęp online: `https://en.wikipedia.org/wiki/Hierarchical_clustering`

[Fionn Murtagh, Pierre Legendre] *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?* , dostęp online: `hhttp://adn.biol.umontreal.ca/~numericalecology/Reprints/Murtagh_Legendre_J_Class_2014.pdf`

[Zasoby online] *Hierarchical Clustering documentation* , dostęp online: `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html`