Uniwersytet
Ekonomiczny
w Katowicach

Zastosowanie uczenia maszynowego w predykcji rezultatów meczów piłki nożnej

Szymon Głowania
Jan Kozak

Katedra Uczenia Maszynowego

Zastosowanie uczenia maszynowego w predykcji rezultatów meczów piłki nożnej

Inspiracje

International Journal of Forecasting 35 (2019) 741–755

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Predictive analysis and modelling football results using machine learning approach for English Premier League

Rahul Baboota [a], Harleen Kaur [b,*]

[a] Guru Gobind Singh Indraprastha University, New Delhi, India
[b] Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India

ARTICLE INFO

Keywords:
Machine learning
Feature engineering
Data mining
Predictive analysis
Random forest
Support vector machines (SVM)
Ranked probability score (RPS)
Gradient boosting

ABSTRACT

The introduction of artificial intelligence has given us the ability to build predictive systems with unprecedented accuracy. Machine learning is being used in virtually all areas in one way or another, due to its extreme effectiveness. One such area where predictive systems have gained a lot of popularity is the prediction of football match results. This paper demonstrates our work on the building of a generalized predictive model for predicting the results of the English Premier League. Using feature engineering and exploratory data analysis, we create a feature set for determining the most important factors for predicting the results of a football match, and consequently create a highly accurate predictive system using machine learning. We demonstrate the strong dependence of our models' performances on important features. Our best model using gradient boosting achieved a performance of 0.2156 on the ranked probability score (RPS) metric for game weeks 6 to 38 for the English Premier League aggregated over two seasons (2014–2015 and 2015–2016), whereas the betting organizations that we consider (*Bet365* and *Pinnacle Sports*) obtained an RPS value of 0.2012 for the same period. Since a lower RPS value represents a higher predictive accuracy, our model was not able to outperform the bookmaker's predictions, despite obtaining promising results.

Cel i motywacja

Celem jest poprawa predykcji wyników meczów piłkarskich na podstawie polskiej PKO Ekstraklasy, w stosunku do przewidywań bukmacherów.
Aby osiągnąć cel konieczne jest przygotowanie zbioru uczącego oraz opracowanie modelu uczącego. Dobrze, aby przygotowane podejście mogło być skalowalne na inne ligi.

Przegląd literatury

- R. Baboota and H. Kaur. **Predictive analysis and modelling football results using machine learning approach for english premier league**. International Journal of Forecasting, 35(2):741–755, 2019.
- R. P. Bunker and F. Thabtah. **A machine learning framework for sport result prediction. Applied Computing and Informatics**, 15(1):27–33, 2019.
- W. Cai, D. Yu, Z. Wu, X. Du, and T. Zhou. **A hybrid ensemble learning framework for basketball outcomes prediction. Physica A: Statistical Mechanics and its Applications**, 528:121461, 2019.
- D. Delen, D. Cogdell, and N. Kasap. **A comparative analysis of data mining methods in predicting ncaa bowl outcomes**. International Journal of Forecasting, 28(2):543–552, 2012.
- E. Eryarsoy and D. Delen. **Predicting the outcome of a football game: Acomparative analysis of single and ensemble analytics methods**. In HICSS, 2019.
- A. Joseph, N. E. Fenton, and M. Neil. **Predicting football results using bayesian nets and other machine learning techniques**. Knowledge-Based Systems, 19(7):544–553, 2006. Creative Systems.

Przegląd literatury

- J. Kahn. **Neural network prediction of nfl football games**. World Wide Web electronic publication, pages 9–15, 2003.
- C. K. Leung and K. W. Joseph. **Sports data mining: Predicting results for the college football games**. Procedia Computer Science, 35:710–719,2014. Knowledge-Based and Intelligent Information Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings.
- A. Maszczyk, A. Gołaś, P. Pietraszewski, R. Roczniok, A. Zając, and A. Stanula. **Application of neural and regression models in sports results prediction**. Procediasoci Behavio Sci, 117:482–487, 2014.
- A. McCabe and J. Trevathan. **Artificial intelligence in sports prediction**.In Fifth International Conference on Information Technology: New Gene-rations (itng 2008), pages 1194–1197, 2008.
- N. Razali, A. Mustapha, F. A. Yatim, and R. Ab Aziz. **Predicting football matches results using bayesian networks for english premier league (EPL)**. IOP Conference Series: Materials Science and Engineering, 226:012099, aug 2017.
- A. P. Rotshtein, M. Posner, and A. B. Rakityanskaya. **Football predictions based on a fuzzy model with genetic and neural tuning.Cybernetics and Systems Analysis**, 41(4):619–630, 2005.

Przegląd literatury

- H. Rue and O. Salvesen. **Prediction and retrospective analysis of soccer matches in a league**. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3):399–418, 2000.
- G. Schauberger, A. Groll, and G. Tutz. **Modeling football results in the german bundesliga using match-specific covariates**, 2016.
- R. P. Schumaker, T. A. Jarmoszko, and C. S. Labedz. **Predicting wins and spread in the premier league using a sentiment analysis of twitter**. Decision Support Systems, 88:76–84, 2016.
- K Sujatha, T. Godhavari, and N. P. G. Bhavani. **Football match statistics prediction using artificial neural networks**. International Journal of Mathematical and Computational Methods, 3, 2018.

Przegląd literatury

- L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. **Classification and Regression Trees**. Chapman & Hall, New York, 1984.
- L. Breiman. **Bagging predictors**. Machine learning, 24(2):123–140, 1996.
- L. Breiman. **Random forests**. Machine learning, 45(1):5–32, 2001. C. Cortes and V. Vapnik. **Support-vector networks**. Machine Learning, 20(3):273–297, September 1995.
- Y. Freund and R. E. Schapire. **A decision-theoretic generalization of on–line learning and an application to boosting**. Journal of computer and system sciences, 55(1):119–139, 1997.
- Y. Freund, R. E. Schapire, et al. **Experiments with a new boosting algorithm**. Inicml, volume 96, pages 148–156. Citeseer, 1996.
- T. Morzy and A. Leceniewska. **Eksploracja danych**.
- T. Morzy. **Eksploracja danych: problemy i rozwiązania**. In5th PLOUG Conference Zakopane, 1999.

Metodologia badań – odkrywanie wiedzy z danych

Metodologia badań

Metodologia badań - Określenie problemu

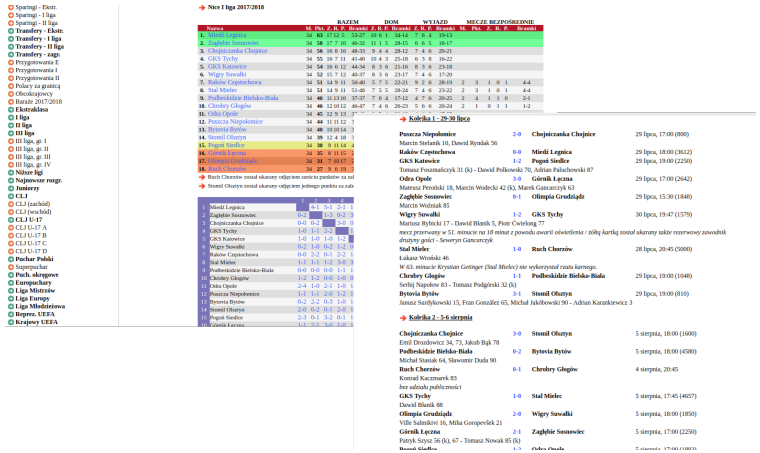
- Złożoność problemu: 6 561 kombinacji w każdej kolejce (3 możliwe wyniki dla 8 spotkań, 3^8 możliwości, 30-37 kolejek w ramach sezonu).
- Konfrontacja z podejściem bukmacherów.
- Możliwość optymalizacji.

Metodologia badań

Zbiór danych

Zbiór danych przygotowany został na podstawie archiwalnych danych dostępnych w serwisie 90minut.pl – wszystkie dane musiały zostać sparsowane.

Do zbioru uczącego włączone zostały wszystkie sezony od 2013/2014 do 2018/2019.



Dane w liczbach:

6 sezonów;

5 pierwszych kolejek nie jest branych pod uwagę;

1536 przypadków w tabeli decyzyjnej (dodanych spotkań);

6 atrybutów warunkowych;

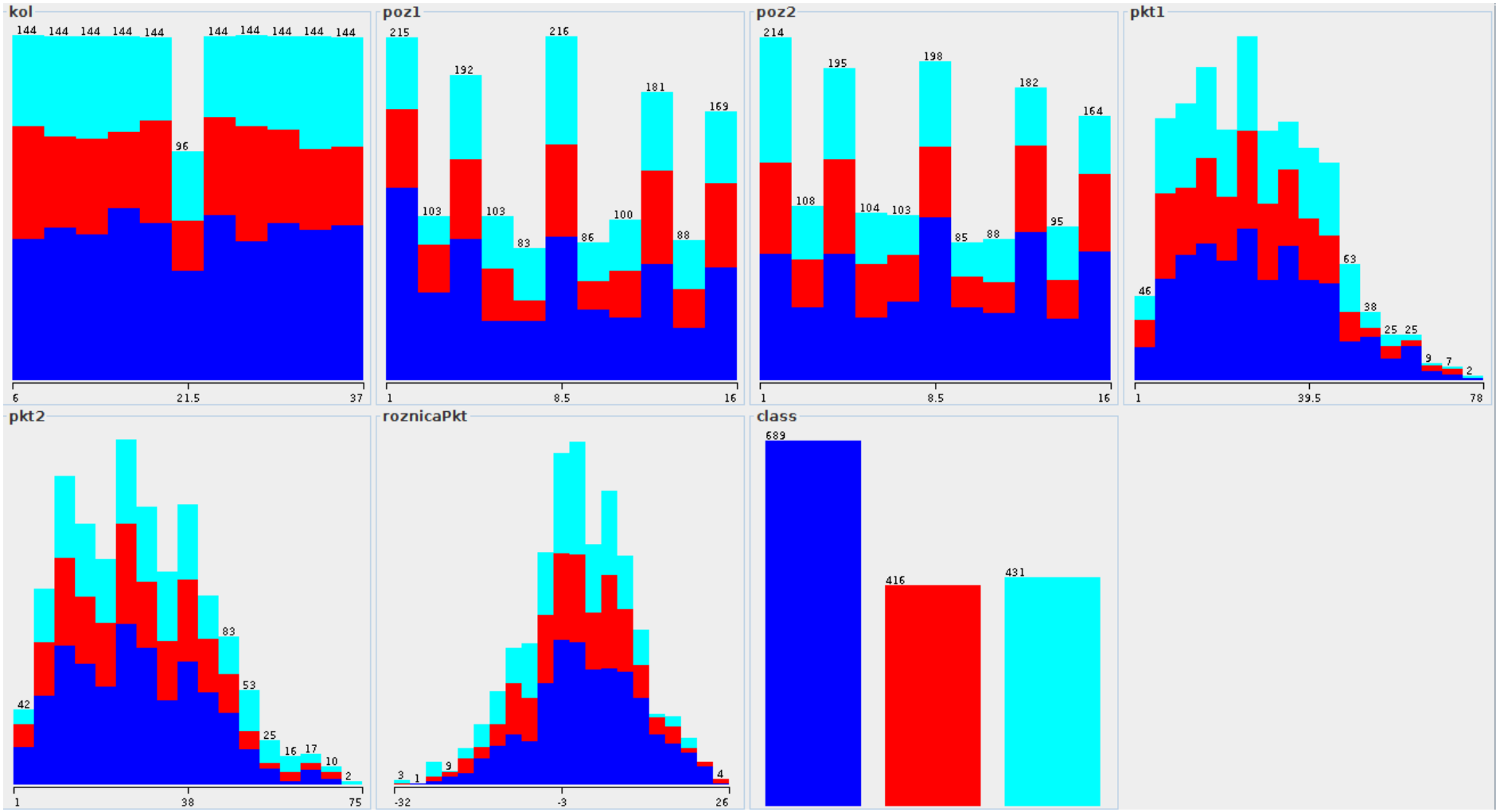
3 klasy decyzyjne.

Zbiór danych

Atrybuty warunkowe	**Wartości**
Kolejka	6 – 37
Pozycja 1	1 – 16
Pozycja 2	1 – 16
Punkty 1	0 – 90
Punkty 2	0 – 90
Różnica punktów	-90 – 90

Klasa decyzyjna	**Wystąpienia**	**Udział**
1	689	44,86%
0	416	27,08%
2	431	28,06%

Zbiór danych



Metodologia badań

Klasyfikatory

- Drzewa decyzyjne - struktura przypominająca swym kształtem drzewo, przedstawiająca różne możliwe ścieżki decyzyjne i ich skutki.
- Maszyna wektorów nośnych - szuka hiperpłaszczyzny, która „najlepiej” rozdziela klasy treningowego zbioru danych poprzez maksymalizację odległość do najbliższego punktu w każdej z klas.
- Lasy losowe - polegającej na tworzeniu wielu drzew decyzyjnych i pozwoleniu im na wybór sposobu klasyfikacji danych wejściowych w drodze głosowania.
- Boosting - ogólna metoda służącą zwiększeniu skuteczności dowolnego algorytmu uczenia. Idea budowania „mocnego i złożonego klasyfikatora” ze „słabych i prostych klasyfikatorów”.

Zastosowane klasyfikatory

Do przygotowania prototypu zastosowane zostały algorytmy dostępne w bibliotece scikit-learn języka Python.

Podejście przetestowane zostało z algorytmami:

- Drzewa decyzyjne CART (bez przycinania), DT0;
- Drzewa decyzyjne CART (przycięte do 3 poziomu), DT;
- Maszyna wektorów nośnych, SVM;
- Lasy losowe (100 drzew przyciętych na 3 poziomie), RF;
- AdaBoost, AB.

Metodologia badań

Badania

W celu weryfikacji proponowanego rozwiązania zostały wytrenowane i przetestowane odpowiednie modele, które następnie zostały użyte do predykcji wyników rozgrywek.

Badane były wyniki rozgrywek polskiej PKO Ekstraklasy w kolejkach od 16 do 30 sezonu 2019-2020. Mecze były rozgrywane pomiędzy 22 listopada 2019 roku a 14 czerwca 2020 roku i obejmowały 120 zdarzeń.

Rzeczywiste zastosowanie podejścia – miary oceny jakości klasyfikacji

$$\textbf{dokładność klasyfikacji} = \frac{TP+TN}{TP+TN+FP+FN}$$

	Predykcja pozytywna	Predykcja negatywna
Pozytywne przypadki (P)	Prawdziwie pozytywne (TP)	Fałszywie negatywne (FN)
Negatywne przypadki (N)	Fałszywie pozytywne (FP)	Prawdziwie negatywne (TN)

Dokładność klasyfikacji

Podejście	**Dokładność klasyfikacji**
Losowe	33,33%
DT0	37,50%
DT	51,67%
SVM	50,83%
RF	54,17%
AB	52,50%

Typowanie u bukmachera

Dla ułatwienia obliczeń, każde zdarzenie zostało zagrane za kwotę 10 zł u legalnego bukmachera. W związku z tym kurs (w wypadku trafnego wytypowania) mnożony jest przez 8,8. Zastosowano rzeczywiste kursy z dnia wykonania klasyfikacji.

Podejście	**Suma wkładu**	**Suma wygranych** (PLN)	**Bilans** (PLN)	**Procentowy zysk/strata**
DT0	1200,00	1148,17	-51,83	-4,32%
DT	1200,00	1246,17	+46,17	3,85%
SVM	1200,00	1214,94	+14,94	1,25%
RF	1200,00	1268,53	+68,53	5,71%
AB	1200,00	1250,41	+50,41	4,20%

Podsumowanie

- Zaproponowane podejście pozwala na predykcję wyników meczów sportowych z większym prawdopodobieństwem niż podejście losowe.
- Predykcja jest wystarczająca by przy zastosowaniu podstawowych mechanizmów uczenia się, możliwe było osiągnięcie zysków w wirtualnych zakładach u bukmacherów.
- Dodatkowe analizy wskazują, że podejście jest skalowane (podobnie sprawdza się w innych ligach).
- Dużym problemem jest brak predykcji remisu.

Dalsze prace

- Jest to prototyp rozwiązania, które należy udoskonalić, zarówno pod względem przygotowania zbioru danych (nowe atrybuty), jak i dodatkowej wiedzy zewnętrznej poprzez inne mechanizmy.
- W dalszych pracach oprócz rozszerzenia listy atrybutów przetestowana będzie możliwość zastosowania sztucznych sieci neuronowych, oraz heterogenicznych zespołów klasyfikatorów co może poprawić wyniki klasyfikacji.