HANSHIN UNIVERSITY

Coursework Research Paper

Dept, of IT Transmedia Contents, Hanshin University

K-NN 알고리즘을 활용한 농구선수 포지션 예측

Predicting Basketball Player Position using K-NN Algorithm

Seungjae Moon

Dept, of IT Transmedia Contents, Hanshin University

[Abstract]

This paper analyzes basketball players' record data and classifies and predicts each player's position through K-NN machine learning algorithm to determine whether it can help assign appropriate positions to players. After finding optimal K for K-NN algorithm by 10-fold cross validation and predicting, we were able to confirm that the classification results showed very high classification accuracy.

Key word : Appropriate position, K-NN algorithm, Classification, Prediction, Machine learning.

Ⅰ. 서 론

기계학습을 이용한 분류 및 회귀 방법이 발달하면서, 다양한 분야에서 기계학습을 활용하여 분석하고 예측하기 시작하였다[1]. 특히, 스포츠 과학 분야에서는 기계학습을 활용한 분석이 활발해지면서 큰 성장을 이루었다[2].

축구, 농구, 야구 등의 많은 스포츠에서 각 포지션은 주어진 역할이 다르고, 해당 포지션마다 선수의 필요 능력이 다르다. 예를 들어 골 결정력이 높고, 위치선정 능력이 높은 축구선수는 스트라이커 포지션에 적합하다[3]. 또, 키가 크고, 리바운드를 잘 따내며, 블록을 잘 하는 농구선수는 센터 포지션에 적합하다[4]. 이렇게, 스포츠 분야에서는 선수들의 기록과 능력 데이터를 분석하여 선수들에게 적합한 다양한 전술들을 구사할 수 있고, 이는 승리에 엄청난 영향력을 준다. 이때, 기계학습은 데이터를 분석하고 예측하는 데에 많은 도움을 줄 수 있다.

본 논문은 농구선수들의 기록 데이터를 분석하고 기계학습 알고리즘 중 분류에 적합한 K-NN 알고리즘을 이용하여 각 선수의 포지션을 분류 예측하여, 선수들에게 적절한 포지션을 배정하는 데에 도움을 줄 수 있는지 가능성을 확인한다.

Ⅱ. 관련 연구

2-1 K-NN (K-Nearest Neighbor)알고리즘

K-NN 알고리즘은 비슷한 특성을 가진 데이터는 비슷한 범주에 속한다는 경향을 가정하고 주변에 있는 데이터를 참고하여 어떤 그룹에 속할지 결정하는 알고리즘이다[5]. 이때, 주변에 있는 참고할 가까운 이웃의 개수를 K로 정하고, K값에 따라 분류 정확도가 달라지므로 최적의 K값을 정하는 것이 중요하다.

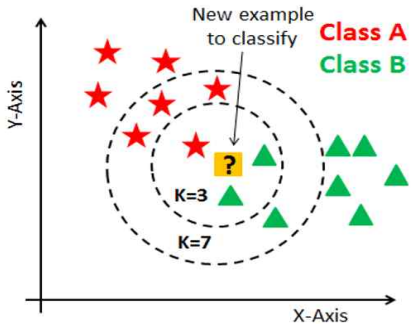

그림 2 [6]. K-NN (K 최근접 이웃)
Fig. 2 [6]. K-NN (K-Nearest Neighbor)

본 연구에서는 적절한 변수를 선택하고, K-NN 알고리즘을 이용하여 농구선수의 포지션을 분류하여 예측한다.

2-2 K-fold 교차검증

K-fold 교차검증이란, 모델의 성능을 평가하는 검증방식으로, 데이터를 K개의 블록으로 나누어 한 개의 검증 데이터와 K-1개의 학습데이터로 사용하여 총 K 번 반복하여 검증하는 방식이다[7].

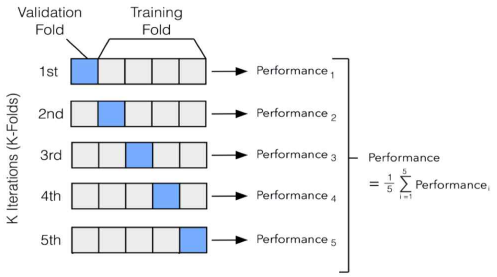

그림 3 [8]. K-fold 교차검증
Fig. 3 [8]. K-fold Cross Validation

본 연구에서는 K-NN 알고리즘의 최적의 K값을 찾기 위해, 10-fold 교차검증을 이용하여 총 10번 반복하여, K값에 따른 분류 정확도를 도출하여 비교한다.

Ⅲ. 데이터

3-1 데이터 구성

데이터 세트는 오픈소스 데이터 세트를 활용하였으며, 총 100명의 선수별 Pos(포지션), 3P(3점), 2P(2점), TRB(리바운드), AST(어시스트), STL(스틸), BLK(블록) 기록으로 구성되어 있다. 포지션은 SG (Shooting Guard) 50명과 C(Center) 50명 두 종류로 구성되어 있다.

3-2 데이터 시각화 및 전처리

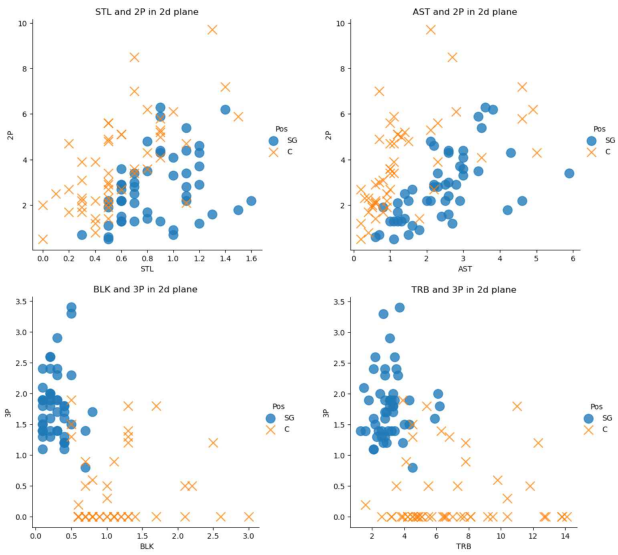

그림 1. 변수별 데이터 분포
Fig. 1. Data distribution by variable

그림 1은 데이터의 분포를 시각화한 결과이다. 데이터를 시각화하였을 때, 분포상, 2P, STL, AST 특징에 대한 포지션별 경계가 명확하지 않고, 그 경계가 매우 근접한 것을 확인할 수 있다. 3P, BLK, TRB 변수의 경우 특징에 대한 포지션별 경계가 뚜렷하고, 구분이 확실한 것을 확인할 수 있다. 따라서, 데이터의 분별력이 없다고 판단되는 특징인 2P, STL, AST는 데이터 세트에서 제거하고, 3P, BLK, TRB 세 개의 변수만 분류에 사용하도록 한다. 또한, 분류를 진행하기 전, 데이터를 8:2의 비율로 학습데이터와 테스트 데이터로 분리한다.

Ⅳ. 실험 결과 및 분석

그림4는 변수 선정과 각 K값에 따른 검증 정확도를 나타낸 그래프이다. 3P, BLK, TRB 총 세 개의 변수로 최적의 K값을 찾은 결과, K는 9일 때 0.9375의 가장 높은 검증 정확도를 보였으며, 3P, BLK 두 개의 변수로 최적의 K값을 찾은 결과, K값은 5일 때 0.9625의 가장 높은 검증 정확도를 보였다. 3P, TRB 두 개의 변수로 최적의 K값을 찾은 결과, K값은 17일 때 0.9375의 가장 높은 검증 정확도를 보였으며. BLK, TRB 두 개의 변수로 최적의 K값을 찾은 결과 K값은 7일 때 0.8875의 가장 높은 검증 정확도를 보였다.

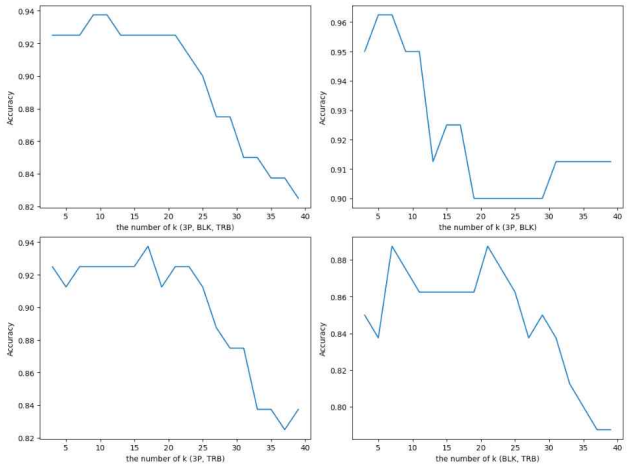

그림 4. 변수 선정과 K값에 따른 검증 정확도
Fig. 4. Validation accuracy according to variable selection and K value

Variables (K=optimal K)	Validation Accuracy	Test Accuracy
3P, BLK, TRB (K=9)	0.9375	0.85
3P, BLK (K=5)	0.9625	0.95
3P, TRB (K=17)	0.9375	0.85
BLK, TRB (K=7)	0.8875	0.90

표 1. 변수별 최적의 K 값에 따른 분류 정확도
Fig. 1. Classification accuracy according to optimal k value for each variable

표 1은 변수별 최적의 K값으로 K를 정하고, 테스트 데이터를 분류한 결과이다. 네 가지 케이스로 나누어 분류한 결과, 3P, BLK 두 변수를 사용하고 K값을 5로 설정하였을 때, 95%의 매우 높은 정확도를 확인할 수 있었다. 또한, 검증 정확도와 테스트 정확도의 차이가 약 1%로 네 가지 케이스 중 가장 작은 차이를 보인다.

Ⅴ. 결 론

본 논문에서는 농구선수의 기록 데이터를 분석하여 K-NN 알고리즘으로 분류하였다. 분류 결과, 변수 선택이 분류 정확도에 큰 영향을 미쳤으며, 3점 슛과 블로킹 점유율이 센터 포지션과 슈팅가드 포지션을 분류하는 데에 큰 영향을 미쳤다. 또한, 95% 이상의 매우 높은 정확도를 보였으며, 농구선수의 기록 데이터를 통한 K-NN 알고리즘으로 분류가 적절한 포지션을 배정하는 데에 도움을 줄 수 있다고 판단하였으며, 스포츠 과학 분야에 기계학습이 유용하게 사용될 것으로 기대된다.

References

[1] Iqbal H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions”, SN Computer Science. SCI, 2, 160, March 2021.

[2] Chris Richter, Martin O’Reilly, Eamonn Delahunt, “Machine learning in sports science: challenges and opportunities”, *Sports Biomechanics Journal*, Apr 2021.

[3] Wonjae Lee, SoJung Lee, JungJae Lee, “A Study on the Analysis of Stamina, Anaerobic Power and Performance of Varying Positions among High School Soccer Players”, *International Journal of Coaching Science,* vol.15, no.2, 2013.

[4] “Needs Analysis For Basketball Positional Role.” *Edubirdie*, 21 Feb. 2022.

[5] Martin E. Hellman, “The Nearest Neighbor Classification Rule with a Reject Option”, *IEEE Transactions on Systems Science and Cybernetics,* Vol. 6, pp. 179-185, July 1970.

[6] Rajvi Shah, “Introduction to k-Nearest Neighbors (kNN) Algorithm”[Online image], *Artificial Intelligence in Plain English,* Mar 2021, https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8

[7] Shanthababu Pandian, “K-Fold Cross Validation Technique and its Essentials.”, *Analytics Vidhya*, July 2023.

[8] Shivaadith Anbarasu, “What Is Cross-Validation In Machine Learning? Why We Need To Do It?”[Online image], *Pianalytix*, https://pianalytix.com/what-is-cross-validation-in-machine-learning/