논문 2025-2-13 http://dx.doi.org/10.29056/jsav.2025.06.13

듀얼 가중치 KNN 알고리즘 기반 건강상태예측 시스템의 설계 및 분석

심정연*†

Design and Analysis of Dual Weighted KNN Algorithm_Based Health Status Prediction System

JeongYon Shim*†

요 약

현대 사회에서는 만성질환 및 심혈관 질환의 조기 예측 및 관리가 매우 중요해졌으며, 이를 위해 심박수, 혈압, 혈당과 같은 생체 신호를 기반으로 한 자동 건강 예측 시스템에 대한 연구가 활발히 진행되고 있다. 특히 KNN 알고리즘은 구현이 단순하면서도 특정 조건 하에서 높은 분류 성능을 보이는 비모수(non-parametric) 학습 기법으로, 의료 분야에서의 적용 사례가 다양하게 보고되고 있다. 그러나 기존 KNN은 각 속성의 중요도를 동일하게 간주하므로, 실제 임상적 중요도와는 괴리가 존재할 수 있다. 본 연구에서는 기존 KNN모델의 한계성을 극복하기 위해 이웃 샘플의 거리 기반 가중치와 각 특성(feature)의 중요도를 반영하는 특성 가중치를 동시에 적용하는 듀얼 가중치 기반 KNN 분류기를 제안하였다. 특히, 혈당 수치의 영향력을 강조하는 방식으로 가중치를 설계하였으며, 다양한 k값에 따른 성능을 비교하고, 사용자 입력 기반의 실시간 진단 기능을 포함하는 응용 시스템을 구현하였다. 실험 결과 가중치 설정에 따라 예측 정확도가 유의하게 향상됨을 보여주었다.

Abstract

Early prediction of chronic diseases have become increasingly critical in modern healthcare. To support this need, automated health prediction systems using biosignals such as heart rate, blood pressure, and blood glucose levels have been actively studied. Among various classification methods, KNN is widely used due to its simplicity and effectiveness in certain scenarios. However, conventional KNN treats all features as equally important, which may conflict with actual clinical priorities. In this study, we propose a dual-weighted KNN classifier that integrates both distance-based weights and feature-based weights. In particular, the model assigns higher importance to blood glucose levels. We also develop a system that enables real-time health status prediction for user input and evaluate the model across various values of k. Experimental results demonstrate that the proposed approach improves prediction accuracy compared to traditional KNN, validating its effectiveness for health monitoring applications

한글키워드 : 듀얼 가중치 KNN, 건강 예측, 학습, 분류, 헬스모니터링

keywords : Dual weighted KNN, Health prediction, learning, classification, health monitoring

* 강남대학교 참인재대학 컴퓨터 전공
† 교신저자: 심정연(email: mariashim@kangnam.ac.kr)

접수일자: 2025.06.06. 심사완료: 2025.06.12.
게재확정: 2025.06.20.

1. 서 론

최근 개인 건강 관리에 대한 관심이 높아지면서, 웨어러블 디바이스 및 IoT 기반 헬스케어 시스템을 활용한 실시간 건강 상태 분석에 대한 연구가 활발히 진행되고 있다. 특히 복잡한 의료 장비 없이도 기본적인 생체 지표만으로 건강 이상을 조기에 감지할 수 있는 기술에 대한 수요가 증가하고 있으며, 이는 인공지능 및 머신러닝 알고리즘을 활용한 자동화된 진단 기술로 이어지고 있다.

그중 KNN 알고리즘은 구현이 단순하면서도 특정 조건 하에서 높은 분류 성능을 보이는 비모수(non-parametric) 학습 기법으로, 의료 분야에서의 적용 사례가 다양하게 보고되고 있다. 그러나 기존 KNN은 각 속성의 중요도를 동일하게 간주하므로, 실제 임상적 중요도와는 괴리가 존재할 수 있다. 예를 들어, 고혈당 수치는 당뇨 및 심혈관 질환의 중요한 징후로 간주되지만, 기존 KNN은 심박수나 혈압과 동일한 가중치를 부여하는 문제가 있다.

이에 본 연구에서는 이웃 샘플의 거리 기반 가중치와 각 특성(feature)의 중요도를 반영하는 특성 가중치를 동시에 적용하는 듀얼 가중치 기반 KNN 분류기를 제안하고, 이를 바탕으로 실시간 건강 상태 예측 시스템을 구현하였다. 다양한 실험을 통해 해당 알고리즘의 유효성을 검증하고, 시각화 및 실시간 예측 기능까지 포함하는 통합적인 응용 가능성을 제시하고자 한다.

2. 관련 연구

2.1 K-최근접 이웃(KNN) 알고리즘

기계학습 알고리즘 중 K-최근접 이웃(K-Nearest Neighbors, KNN)은 간단한 원리와 높은 직관성으로 인해 다양한 분류 문제에 널리 사용되고 있다[1-3]. KNN은 훈련 과정을 요구하지 않는 비모수(non-parametric) 학습 방법으로, 데이터의 분포에 대한 특정한 가정을 두지 않고, 학습 데이터 공간 상의 거리(proximity) 를 기반으로 새로운 샘플에 대한 예측을 수행한다. 새로운 입력 샘플이 주어졌을 때, 기존 학습 데이터 중에서 가장 가까운 k개의 샘플을 선택하고 그들의 클래스 레이블 중 가장 빈도가 높은 값을 예측 결과로 반환한다.

주어진 입력 벡터 $X_{test \in} R^n$ 에 대해 KNN의 기본 작동원리는 다음 알고리즘에 따른다.

Step1. 거리계산 : 테스트 샘플과 학습 샘플 $X^{(i)} \in D$ 간의 거리를 계산한다. 일반적으로 유클리디안 거리(Euclidean distance)가 사용되며 계산식은 다음 식 (1)과 같다.

$$d(X_{test}, X^{(i)}) = \sqrt{\sum_{j=1}^{n} (x_j^{test} - x_j^{(i)})^2} \quad (1)$$

상황에 따라 Manhattan 거리, Minkowski 거리, Mahalanobis 거리 등이 활용될 수도 있다.

step2 최근접 이웃 선택 : 계산된 거리값을 기준으로 가장 가까운 k 개의 이웃 샘플들을 선택한다.

$$N_k(X_{test}) = \{(x^{(i)}, y^{(i)}) | i \in \text{상위 } k\text{개의 근접 샘플}\} \quad (2)$$

step3 예측 수행 :

- 분류 문제의 경우 k개의 이웃 중 다수결 투표를 통해 최종 클래스 $\hat{y}_{test}$를 결정한다.

$$\hat{y}_{test} = \arg\max_{c \in C} \sum_{(x^{(i)}, y^{(i)}) \in N_k} I(y^{(i)} = c) \quad (3)$$

여기서 $I(.)$는 지시함수이다.

-회귀 문제의 경우 k개 이웃 값의 평균을 예측값으로 사용한다.

$$\hat{y}_{test} = \frac{1}{k} \sum_{(x^{(i)}, y^{(i)}) \in N_k} y^{(i)} \quad (4)$$

KNN 알고리즘은 모델 구조가 단순하고 학습이 필요없다는 장점이 있으나 차원이 높아질수록 데이터가 희소해지며 거리기반 계산의 신뢰도가 저하되어 성능 저하로 이어질 수 있다. 또한 적절한 k 값을 선택하는 것이 중요한데 작은 k값은 과적합(overfitting)을, 큰 k값은 과소적합(underfitting)을 유발할 수 있다. 거리계산은 각 특성 척도에 민감하므로 표준화나 정규화가 필요하며 데이터셋이 클 경우, 모든 학습 샘플과의 거리 계산이 필요하므로 계산 복잡도가 높아진다. 모든 특성(feature)에 대해 동일한 영향을 가정하므로 실제 데이터 특성에 대한 고려가 부족하다. 특히 의료도메인과 같이 입력변수의 임상적 중요도가 다른 경우에는 예측 성능이 저하될 수 있다[4].

2.2 가중치 기반 KNN 알고리즘

기계학습 알고리즘 중 K-최근접 이웃(K-Nearest KNN)의 성능을 향상시키기 위해 가중치 기반 KNN(Weighted KNN) 방식이 제안되었다. 이는 다음 두 가지 방식으로 확장될 수 있다:

이웃 샘플 가중치 적용: 거리에 따라 가까운 이웃일수록 높은 가중치를 부여한다[5,6]:

$$w_i = \frac{1}{d(x_{test}, x^{(i)}) + \epsilon} \quad (5)$$

여기서 ϵ는 분모의 0 나눗셈을 방지하기 위한 작은 양수이다.

특성(feature)가중치 적용 : 각 특성의 중요도를 반영하여 거리 계산식을 수정한다[7].

$$d(X_{test}, X^{(i)}) = \sqrt{\sum_{j=1}^{n} w_j (x_j^{test} - x_j^{(i)})^2} \quad (6)$$

이때 w_j는 특성 j의 상대적 중요도를 의미하며 도메인 지식을 반영하거나 데이터 기반으로 설정될 수 있다. 이와 같은 확장은 특정 특성(예:혈당) 이 예측에 더 중요한 영향을 미치는 의료 분야와 같은 응용 분야에서 모델 성능을 개선시킬 수 있다.

3. 듀얼 가중치 KNN 알고리즘 기반 건강 상태 예측 시스템 설계

3.1 듀얼 가중치 기반 KNN

본 연구에서는 의료 예측의 정확도를 높이기 위하여 KNN알고리즘을 확장하여 거리에 따라 가까운 이웃일수록 높은 가중치를 부여하는 이웃 샘플 가중치와 특성의 중요도를 반영한 특성 가중치를 모두 고려한 듀얼 가중치 기반 KNN 알고리즘(Dual Weighted KNN)을 제안 하였다.

이 모델에서는 특성 중요도를 반영하여 건강 지표(혈당, 혈압 등)에 따라 거리 계산을 조정하고 거리 기반 이웃 가중치를 통해 가까운 샘플일수록 더 큰 영향을 미치도록 설계하여 두 가지 가중치를 곱합 형태로 통합하여 정밀도와 설명력을 동시에 강화하였다.

특성 가중치 기반 거리 (Feature-weighted Distance) : 특성 가중치 기반 거리 계산은 다음 식 (7)과 같다.

$$d(x, x_i') = \sqrt{\sum_{j=1}^{n} w_j (x_j - x_j')^2} \quad (7)$$

여기서 w_j는 특성 j의 중요도를 나타내는 가중치 (예: 혈당 > 심박수)이며 정규화된 w_j를 사용하여 전체거리 척도의 일관성을 유지한다. 의료 도메인에서는 전문가 지식이나 정보이득(Infromation gain)형태로 도출이 가능하다.

거리기반 이웃가중치 (Distance - based Neighbor Weight) : 선택된 이웃 $x^{(i)}$에 대하여 다음과 같은 가중치를 부여한다.

$$\alpha_i = \frac{1}{d(x_{test}, x^{(i)})^{\Upsilon} + \epsilon} \quad (8)$$

여기서 Υ은 거리민감도 조절 하이퍼파라메터로 $\Upsilon > 0$이며 예를 들어 1, 2와 같은 값을 갖는다. ϵ는 0으로 나누는 것을 방지하는 역할을 하는 값이고 아주 작은 값이 주어진다. 거리 d는 위에서 정의한 특성 가중치 거리를 의미한다.

듀얼 가중치 기반 예측함수(Dual Weighted Prediction): 특성 기반 가중치와 거리 기반 가중치가 결합된 듀얼 가중치 기반 예측함수는 다음과 같다.

- 분류 문제의 경우 :

$$\hat{y} = \arg\max_{c \in C} \sum_{i \in N_k} \alpha_i \cdot I(y^{(i)} = c) \quad (9)$$

-회귀 문제의 경우 :

$$\hat{y} = \frac{\sum_{i \in N_k} \alpha_i \cdot y^{(i)}}{\sum_{i \in N_k} \alpha_i} \quad (10)$$

이러한 기법들은 의료 분야 건강상태 예측에 많이 응용되고 있다[8,9,10]. 표 1은 데이터의 특성 가중치의 예를 보인 것으로. 비율로 특성별 거리 기여도를 조정하며, 거리 계산과 이웃 선택 모두에서 의료적 중요도를 반영하고 있다.

제안된 듀얼 가중치 기반 KNN은 어떤 특성이 예측에 얼마나 영향을 미쳤는지 분석 가능하여 설명 기능이 향상되고 정확도 개선으로 의미없는 특성의 영향을 감소시키고 중요한 의료 지표 중심의 예측이 가능하다. 또한 의료 전문가의 판단을 가중치 설계에 직접 반영 가능하여 도메인 지식 반영이 용이하다는 장점을 가지고 있다.

표 1. 의료 데이터: 특성 가중치의 예
Table 1. Medical data: example of Feature Weight

특성	중요도 w_j
혈당	0.50
수축기 혈압	0.20
이완기 혈압	0.15
심박수	0.10
체온	0.05

3.2 듀얼 가중치 KNN 알고리즘

제안된 듀얼 가중치 KNN (Dual Weighted KNN) 알고리즘은 다음과 같다.

Algorithm1 : Dual Weighted KNN algorithm

step1 : Receive the test sample x.

step2 : Compute the weighted Eucleadian distance between x and each training sample x_i:

$$d_w(x, x_i') = \sqrt{\sum_{j=1}^{n} \alpha_j (x_j - x_{i,j})^2}$$

step3 : Select the k nearest neighbors $N_k(x)$ based on the calculated distances.

step4 : For each neighbor, calculate the neighbor sample weight :

$$w_i = \frac{1}{d_w(x, x_i)^2 + \epsilon}$$

(where ϵ is a small positive value to avoid division by zero)

step5 : Calculate the weighted sum of neighbors for each class c:

$$S_c = \sum_{x_i \in N_k(x)} w_i \cdot 1_{y_i = c}$$

step6 : Determine the predicted class $\hat{y}$ as the class with maximum weighted sum:

$$\hat{y} = \arg\max_c S_c$$

4. 실 험

본 실험에서는 제안된 듀얼 가중치 기반 K-최근접 이웃(K-Nearest Neighbors, KNN) 알고리즘 기반 건강 상태를 예측하는 시스템을 개발하고, 그 성능을 예측 정확도, k 값에 따른 변화, 특성 가중치 적용 효과, 사용자 입력 시뮬레이션 등 다양한 관점에서 분석하였다. 합성된 생체 데이터(혈당, 혈압, 심박수 등)를 바탕으로, 이웃 샘플 가중치(weighted neighbor)와 특성 중요도 가중치(weighted feature distance)를 함께 적용하여 기존 KNN의 단점을 보완하였다. 그림 1은 k 값에 따른 예측정확도를 변화를 나타내고 있다. 다양한 k값(이웃 개수)을 변화시키며 예측 정확도를 비교한 결과, 일반적으로 너무 작은 k값은 과적합 (overfitting)의 경향을 보이며, 너무 큰 k 값은 다수 클래스에 편향된 결과를 낳는 경향이 있었다. 실험에서는 k=4 또는 k=6에서 가장 안정적이고 높은 정확도가 나타났으며, 이는 소수의 가장 유사한 샘플들로부터의 영향이 신뢰도 높은 예측을 도출함을 시사한다. k=5에서의 정확도 하락은 이웃 분포의 모호함과 다수결 투표 구조에서의 비정상적 클래스 분포에 의해 발생한 것으로 해석할 수 있다. 특히 이 데이터에서는 특성 및 거리 가중치의 조합에 따라 작은 변화에도 민감하게 반응했으며 k=4나 k=6에서는 상대적으로 안정적인 이웃 구성이 가능했던 것으로 보인다. 시각화된 정확도 그래프에서도 이러한 경향은 명확하게 확인되었다.

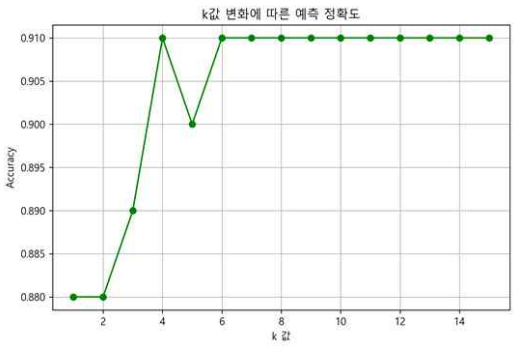

그림 1. k값 변화에 따른 예측 정확도
Fig. 1. Prediction accuracy under varying k values

또한, 특성별 중요도를 반영한 결과(그림 2)도 유의미한 영향을 미쳤다. 특히 혈당 수치에 가장 높은 가중치를 부여한 설정에서 예측 정확도가 뚜렷하게 향상되었으며, 이는 실제 의료 현장에서 혈당이 중요한 진단 지표로 작용함을 잘 반영한다. 반면, 상대적으로 중요도가 낮은 체온이나 심박수에 높은 가중치를 부여했을 경우, 오히려 성능 저하가 발생하는 경향이 관찰되었다. 이는 특성 가중치 설정이 임상적 또는 도메인 지식에 기반 해야 함을 강조한다. 혈당에 더 높은 가중치를 줄수록 정확도가 상승하였으며, 이는 혈당이 건강 이상 상태를 구분하는 주요 지표로 작용한다는 것을 시사한다. 반대로 심박수와 혈압의 중요도를 과도하게 높인 경우 정확도가 하락하는 현상이 관찰되었다.

가중치 설정(심박수,혈압,혈당)	예측정확도(%)
(1.0, 1.0, 1.0) 균등	85.25
(0.8, 0.8, 1.4)	88.75
(0.5, 0.5, 2.0)	**91.00**
(1.2, 1.2, 0.6)	83.75

그림 2. 특성별 중요도를 반영한 결과
Fig. 2. Results Reflecting Feature-Based Weights

표 2는 모델별 정확도를 비교한 수치를 보이고 있는데 듀얼 가중치 기반 모델이 이웃과 특성 양쪽의 중요도를 반영함으로써 가장 높은 정확도를 기록하고 있음을 알 수 있다. 이는 의료 데이터의 민감한 특성과 개별 변수의 중요성이 분류 성능에 영향을 미침을 뒷받침하는 결과이며, 단일 가중치 방식 대비 복합 가중치 방식의 효용성을 입증하고 있다. 또한 사용자 입력 기반의 실시간 예측 기능 실험(그림 3)에서는 임의로 입력된

생체 데이터를 통해 예측 결과를 실시간으로 확인할 수 있었으며, 예측 결과는 “정상” 또는 “주의 필요” 등으로 명확히 표현되어 사용자 친화적인 형태로 구현되었다. 이는 본 시스템이 개인 건강 모니터링이나 간단한 자기 진단 도구로도 활용될 수 있음을 보여준다.

표 2. 모델별 정확도 비교
Table 2. Comparative analysis of model prediction accuracies

모델	정확도(%)
기본 KNN	85.0
이웃 가중치 KNN	86.7
특성 가중치 KNN	86.7
듀얼 가중치 KNN	88.3

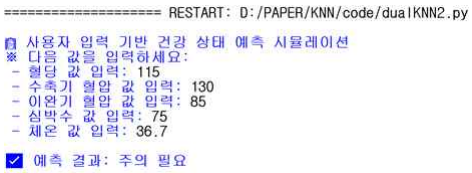

그림 3. 사용자 입력 기반의 실시간 예측 기능 실험
Fig. 3. Real-Time Health Status Prediction Using User-Provided Input

종합적으로, 본 시스템은 가중치 기반의 KNN 알고리즘을 통하여 기존 단순 KNN의 단점인 “모든 특성 동등 처리”와 “이웃의 거리 차이 무시” 문제를 효과적으로 보완하였으며, 특히 의료 데이터를 다룰 때 중요한 특성 중심의 진단과 이웃 가중치를 동시에 반영한 접근 방식이 높은 실용성과 정확도를 보여주었다. 추후 실제 환자 데이터를 기반으로 한 임상 적용이나 웨어러블 헬스 모니터링 시스템과의 연동 등을 통해 확장 가능성이 높다고 판단된다.

5. 결 론

본 연구에서는 심박수, 혈압, 혈당 등 주요 생체 정보를 기반으로 개인의 건강 상태를 실시간으로 예측할 수 있는 시스템을 제안하였다. 특히, 기존의 K-최근접 이웃(KNN) 알고리즘에 이웃 샘플의 거리 기반 가중치와 각 특성의 중요도를 반영한 특성 가중치를 동시에 적용한 듀얼 가중치 기반 KNN 알고리즘을 제안함으로써 예측 정확도를 향상시켰다.

실험 결과 단순 KNN에 비해 제안된 방법은 높은 정확도를 보였으며, 특히 혈당 수치와 같은 주요 변수의 영향력을 정량적으로 반영함으로써 보다 정밀한 진단이 가능함을 입증하였다. 또한, 사용자 입력 기반의 실시간 진단 기능을 통해 직관적인 사용자 인터페이스 및 상호 작용성을 제공함으로써, 개인 건강 모니터링 시스템으로의 실용성을 높였다.

향후에는 보다 다양한 생체 데이터를 활용한 모델 고도화, 실제 IoT 센서와의 연동, 딥러닝 기반 예측 모델과의 비교 및 통합 등을 통해 본 시스템의 성능과 활용 가능성을 더욱 확장할 수 있을 것이다.

참 고 문 헌

[1] Cover, T. M., & Hart, P. E. “Nearest neighbor pattern classification”, IEEE Transactions on Information Theory, 13(1), 21 - 27, 1967. https://doi.org/10.1109/TIT.1967.1053964
[2] Zhang, Z. “Introduction to machine learning: k-nearest neighbors.”, Annals of Translational Medicine, 4(11), 218, 2016. https://doi.org/10.21037/atm.2016.03.37
[3] Han, J., Kamber, M., & Pei, J. “Data Mining: Concepts and Techniques”, 3rd

ed., Morgan Kaufmann, 2011. ISBN: 9780123814791
[4] Duda. Hart & Stork, "Pattern Classification(2nd Edition)", Wiley, 2000. ISBN: 9780471056690
[5] Dudani, S. A. "The distance-weighted k-nearest-neighbor rule", IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(4), 325 - 327, 1976. https://doi.org/10.1109/TSMC.1976.5408784
[6] Keller,Gray & Givens, "A fuzzy k-nearest neighbor algorithm", IEEE Transactions on Systems, Man and Cybernetics", 580-585, 1985. DOI: 10.1109/TSMC.1985.6313426
[7] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In C. C. Aggarwal & C. Zhai (Eds.), Data classification: Algorithms and applications (pp. 37 - 64). CRC Press, 2014. ISBN : 9781466558212
[8] Chowdhury, M. E. H., et al., "Wearable real-time heart attack detection and warning system to reduce road accidents", Sensors, 19(12), 2780, 2019. https://doi.org/10.3390/s19122780
[9] T. Shaik et al., "Remote patient monitoring using artificial intelligence: Current state, applications, and challenges", arXiv preprint arXiv:2301.10009, Jan. 2023. https://arxiv.org/abs/2301.10009
[10] Dey et al.,"Machine Learning techniques for diabetic risk prediction", International Journal of Advanced Computer Science and Applications, 9(9), 359-365. DOI: 10.14569/IJACSA.2018.090950

저 자 소 개

심정연(JeongYon Shim)

1989.2 고려대학교 컴퓨터학과 졸업
1991.2 고려대학교 컴퓨터학과 석사
1998.8 고려대학교 컴퓨터학과 박사
2000 CUHK Post Doc.
2003.3-현재 : 강남대학교 교수
<주관심분야> 인공지능, 지식공학 시스템, Machine Learning, ICA, Information System