

가중치 조정 파라미터를 활용한 확장된 가중 상관계수 연구

윤종찬*, 윤소영**

A Study on the Extended Weighted Correlation Coefficient using a Weighted Adjustment Parameter

Jongchan Yun*, Soyoung Yun**

요 약

본 논문에서는 피어슨 상관계수, 가중 상관계수 그리고 가중치 조정 파라미터 상관계수를 비교 분석하였다. 기존 피어슨 상관계수는 모든 데이터 포인트를 동일한 중요도로 고려하여 특정 데이터의 가중치를 반영하지 못하는 한계가 있다. 이를 보완하기 위해 가중 상관계수가 제안되었으나, 극단적인 가중치의 영향을 조정할 수 없는 문제점이 존재한다. 본 논문에서는 가중치 조정 파라미터(α)를 도입한 확장된 가중 상관계수를 제안하여, 기존 방법보다 가중치가 높은 데이터의 영향을 정밀하게 조절하고 특정 그룹의 데이터를 강조하면서 극단적인 값의 왜곡을 최소화하였다. 이를 검증하기 위해 병원환자 실제 설문 데이터를 사용하여 기존 방법과 비교 분석한 결과, 확장된 가중 상관계수가 보다 정밀한 관계 측정에 기여할 수 있음을 확인하였다.

Abstract

In this paper, we compare and analyze the pearson correlation coefficient, the weighted correlation coefficient and the Weighted Adjustment Parameter Correlation Coefficient (WAPCC). The traditional Pearson correlation coefficient considers all data points with equal importance, limiting its ability to reflect the influence of specific data points. To address this limitation, the weighted correlation coefficient has been proposed; however, it lacks the ability to adjust for extreme weight distortions. To overcome this issue, we propose an extended weighted correlation coefficient incorporating a weight adjustment parameter(α), which allows for more precise control of highly weighted data while minimizing distortions caused by extreme values. Additionally, this approach enables the emphasis of specific groups in the analysis. To validate the effectiveness of WAPCC, we conducted a comparative analysis using real-world hospital patient survey data. The results confirm that the proposed method contributes to a more precise measurement of relationships compared to existing approaches.

Keywords

correlation coefficient, weighted correlation coefficient, outlier, optimization

* 국립부경대학교 정보전산원

- ORCID: <https://orcid.org/0009-0007-3003-2870>

** 국립부경대학교 정보전산원(교신저자)

- ORCID: <https://orcid.org/0009-0001-7477-2113>

• Received: Feb. 20, 2025, Revised: Apr. 06, 2025, Accepted: Apr. 09, 2025

• Corresponding Author: Soyoung Yun

Information & Compter Center, Pukyong National University, Pusan 48513, Korea

Tel : +82-51-629-6831, Email : ysmallzero@pknu.ac.kr

I. 서 론

웹의 데이터 분석에서 변수 간의 관계를 측정하는 것은 매우 중요한 과정이며, 이는 경제학, 교육학, 사회과학, 데이터 과학 등 다양한 분야에서 활용되고 있다. 특히 상관계수(Correlation coefficient)는 두 변수 간의 선형 관계를 정량적으로 측정하는 대표적인 방법이다. 그중 피어슨 상관계수(Pearson correlation coefficient)는 가장 널리 사용되는 방식으로, 두 변수의 공분산을 표준편차로 정규화하여 상관성을 측정한다[1].

그러나 피어슨 상관계수는 모든 데이터 포인트를 동일한 중요도(weight)로 고려한다는 한계를 가진다. 즉, 특정 데이터가 더 중요한 경우(예: 특정 그룹의 응답 또는 신뢰도가 높은 데이터)가 있더라도 이를 반영하지 못하는 문제가 발생한다. 이를 해결하기 위해 가중 상관계수(Weighted correlation coefficient)가 제안되었으며, 데이터 포인트별 중요도를 다르게 설정하여 보다 현실적인 관계를 측정할 수 있도록 하였다[2].

하지만 기존의 가중 상관계수 또한 한계를 가지며, 극단적인 가중치(예: 데이터의 편중)에 대한 조정이 어렵다는 문제가 있다. 따라서 본 논문에서는 이러한 한계를 보완하기 위해 가중치 조정 파라미터 상관계수(WAPCC, Weighted Adjustment Parameter Correlation Coefficient)를 제안한다. 본 논문의 목표는 다음과 같다. 첫째, 기존 상관계수(피어슨 상관계수, 가중 상관계수)의 장단점을 분석하고, 어떤 상황에서 기존 방법이 적절한지를 평가한다. 둘째, 새로운 방법론인 WAPCC를 제안하고, 데이터 분석에서의 활용 가능성을 탐색한다. 셋째, 기존 방법들과 WAPCC를 실제 데이터를 이용하여 비교 분석하고, 확장된 가중 상관계수가 가지는 실질적 이점을 논의한다. 넷째, 실제 설문 데이터를 사용하여 피어슨 상관계수, 가중 상관계수, WAPCC를 비교 분석한다. 다섯째, WAPCC가 데이터 분석에서 가지는 실질적인 장점을 평가하고 논의한다. 여섯째, 극단적인 가중치를 조정했을 때 기존 방법과의 차이를 실험적으로 분석한다.

본 논문을 통해 WAPCC가 데이터 분석에서 보

다 정밀한 관계 측정에 기여할 수 있음을 검증하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서 관련 연구를 분석하고, 제3장에서는 본 연구에서 제안하는 WAPCC를 상세히 기술한다. 제4장에서는 실험 및 분석을 제시하며, 제5장에서 결론 및 향후 과제를 논의한다.

II. 관련 연구

2.1 상관계수

피어슨 상관계수는 두 연속형 변수 간의 선형 관계를 측정하는 대표적인 지표로, 통계학 및 데이터 분석 분야에서 가장 널리 사용되는 상관관계 측정 방법 중 하나이다. 이 계수는 -1에서 1사이의 값을 가지며, 값의 크기와 부호를 통해 변수 간의 관계를 직관적으로 해석할 수 있다[3].

피어슨 상관계수는 다음 식 (1)으로 정의된다[4].

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \quad (1)$$

X_i, Y_i 는 각각 두 변수의 개별 데이터 값

$\bar{X}, \bar{Y}$ 는 각 변수의 평균값

피어슨 상관계수는 여러 연구에서 다음과 같은 이유로 널리 활용된다.

첫째, 계산이 간단하고 직관적이며 평균과 표준편차를 이용하여 쉽게 계산할 수 있고, 데이터의 선형 관계를 명확하게 평가할 수 있다.

둘째, 선형 관계를 명확히 측정할 수 있으며 두 변수 간의 선형적 관계를 분석하는 데 적합하고, 변수 간의 방향성과 강도를 동시에 제공한다. 단순한 공분산과 달리 피어슨 상관계수는 변수 간의 관계 강도와 방향을 함께 보여준다[4].

셋째, 정규화된 값으로 제공하기 때문에 두 변수의 단위에 영향을 받지 않으며, 서로 다른 척도의 변수도 비교할 수 있다. 그러나 피어슨 상관계수는 다음과 같은 한계를 가진다.

첫째, 선형 관계만 측정 가능하고 비선형 관계는 반영하지 못한다. 두 변수 간에 비선형 관계가 존재할 경우 피어슨 상관계수는 낮거나 0에 가까운 값을 가질 수 있다.

둘째, 이상치에 민감하다. 이상치가 존재하면 상관계수 값이 크게 왜곡될 수 있다. 예를 들어 대부분의 데이터가 약한 상관을 가지더라도 몇 개의 극단적인 데이터 포인트가 상관계수를 인위적으로 높이거나 낮출 수 있다.

셋째, 변수의 척도에 영향을 받을 수 있다. 특정 분포에서는 데이터 변환(예: 로그 변환)을 하지 않으면 상관계수가 왜곡될 수 있다.

넷째, 모든 데이터에 동일한 가중치를 적용한다. 피어슨 상관계수는 모든 데이터 포인트를 동일한 중요도로 고려하므로, 신뢰도가 높은 특정 데이터의 영향을 강조할 수 없다.

다섯째, 인과관계를 나타내지 않는다. 높은 상관계수 값이 두 변수 간 인과관계를 의미하는 것은 아니다. 예를 들어 아이스크림 판매량과 익사 사고 간의 상관관계가 높을 수 있지만, 이는 두 변수 모두 '기온'이라는 공변량의 영향을 받은 결과일 수 있다.

피어슨 상관계수는 이와 같은 장단점을 가지고 있으므로, 연구의 목적과 데이터 특성을 고려하여 적절한 상관 측정 방법을 선택하는 것이 중요하다. 필요한 경우 스피어만 상관계수, 켄달의 타우, 가중 상관계수 등의 대체 방법을 활용할 수 있다.

2.2 가중 상관계수

가중 상관계수는 피어슨 상관계수의 확장된 개념으로, 각 데이터 포인트의 중요도를 반영하여 상관관계를 측정하는 방법이다. 즉, 특정 데이터가 더 중요한 경우 이를 반영하여 보다 현실적인 상관관계를 계산할 수 있도록 설계되었다[5]. 가중 상관계수는 다음 식 (2)로 계산된다.

$$r_w = \frac{\sum w_i (X_i - \bar{X}_w)(Y_i - \bar{Y}_w)}{\sqrt{\sum w_i (X_i - \bar{X}_w)^2} \sqrt{\sum w_i (Y_i - \bar{Y}_w)^2}} \quad (2)$$

w_i 는 개별 데이터 i 의 가중치(weight)

$\bar{X}_w, \bar{Y}_w$ 는 가중 평균(weighted mean)

가중 상관계수는 각 데이터 포인트에 가중치를 부여한 후, 가중 평균과 가중 공분산, 가중 표준편차를 기반으로 상관관계를 측정하는 방식이다.

이 방법은 다음과 같은 경우에 유용하게 활용될 수 있다.

첫째, 데이터의 중요도가 다를 때이다. 예를 들어 경제 분석에서 GDP 규모가 큰 국가의 데이터가 더 중요한 경우, 혹은 실험 데이터에서 신뢰성이 높은 데이터의 가중치를 더 높여야 하는 경우가 해당한다.

둘째, 표본 크기가 다를 때이다. 특정 관측값이 다른 값들보다 더 자주 나타나는 경우, 동일한 가중치를 부여하면 왜곡된 결과가 나올 수 있다. 예를 들어 설문조사에서 특정 응답자의 의견이 전체 모집단을 대표하지 않을 경우, 가중치를 조정하여 분석 정확도를 높일 수 있다.

셋째, 데이터 신뢰성이 다를 때이다. 측정 오차가 큰 데이터의 경우, 해당 데이터를 덜 신뢰하고 가중치를 낮추는 방식이 필요하다. 예를 들어 센서 데이터 분석에서 오차율이 높은 센서의 가중치를 조정하면 더 신뢰성 있는 결과를 도출할 수 있다.

넷째, 특정 구간을 강조할 필요가 있을 때이다. 데이터의 특정 구간이 더 중요한 경우, 가중치를 이용해 해당 구간을 강조할 수 있다. 예를 들어 시간 시계열 데이터에서 최근 데이터에 가중치를 높이면 보다 중요한 정보 반영이 가능하다.

가중 상관계수의 장점으로서는 데이터의 중요도를 반영 가능하고 가중치를 활용하여 보다 신뢰성 있는 데이터 분석이 가능하다. 표본 크기의 차이를 보정도 가능하고, 조사나 연구에서 특정 그룹의 표본 크기가 차이 나는 경우 이를 조정할 수 있다. 또한 이상치의 영향을 완화할 수 있다. 이상치에 낮은 가중치를 부여하여 피어슨 상관계수보다 더 안정적인 결과를 얻을 수 있다. 특정 구간에 대한 분석 강화가 가능하고 특정 시간대와 특정 집단 등에 대해 보다 정밀한 분석이 가능하다[6][7].

가중 상관계수의 한계점을 보면, 가중치 선정이 연구자의 주관에 의존할 가능성이 있고 적절한 가중치를 결정하는 방법이 연구의 신뢰성을 크게 좌우된다. 가중치가 잘못 설정될 경우, 결과가 왜곡될 가능성도 있다. 또한 비합리적인 가중치 적용은 분석의 객관성을 해칠 수 있고, 계산 복잡성이 증가한다.

그리고 단순 피어슨 상관계수보다 추가적인 연산이 필요하며 대규모 데이터에서는 계산 부담이 커질 수 있다.

가중 상관계수는 피어슨 상관계수의 확장된 개념으로, 데이터의 중요도를 반영한 보다 정밀한 상관관계 분석이 가능한 강력한 도구이다.

특히 경제학, 금융, 의료, 심리학, 환경 과학 등 다양한 분야에서 신뢰성 있는 데이터 분석을 수행하는 데 유용하게 활용될 수 있다. 그러나 가중치 선정 과정에서 연구자의 주관이 개입될 가능성이 있으므로 신중하게 설계하는 것이 중요하며, 연구 목적에 따라 적절한 방법론을 선택해야 한다[8].

III. 가중치 조정 파라미터 상관계수

기존의 피어슨 상관계수와 가중 상관계수는 변수 간의 관계를 측정하는 데 활용되어 왔다. 그러나 기존 방법들은 모든 데이터 포인트에 동일한 가중치를 부여하거나, 미리 정의된 가중치를 적용하기 때문에 몇 가지 한계를 갖는다.

본 논문에서는 기존 가중 상관계수의 한계를 극복하기 위해 가중치 조정 파라미터(α)를 추가한 확장된 가중 상관계수(WAPCC)를 제안한다.

기존의 가중 상관계수는 가중치를 부여하여 각 데이터의 중요도를 반영하지만, 이상치에 높은 가중치가 부여될 경우 전체 상관관계 결과가 왜곡될 수 있다. 이를 해결하기 위해 본 논문에서는 가중치 조정 파라미터를 도입하여 이상치의 영향을 적절히 조절하고자 한다.

즉, 가중치가 높은 데이터일수록 상관관계에 미치는 영향이 크지만, WAPCC는 가중치 조정 파라미터를 통해 이상치가 존재하더라도 전체 상관계수 계산에서 편향이 발생하지 않도록 가중치를 부드럽게 조정하는 메커니즘을 제공한다. 이러한 조정 과정을 통해 이상치가 포함된 데이터에서도 신뢰할 수 있는 상관계수를 산출할 수 있으며, 기존 방법 대비 이상치에 덜 민감한 보다 견고한 상관계수 계산이 가능해진다.

WAPCC를 적용하기 전에 변수 정규화 과정이 필요하다. 변수들을 0~1 범위로 정규화하는 이유는 다음과 같다. 첫째, 변수마다 크기나 단위가 다를

경우 직접 비교가 어려우므로, 정규화를 통해 분석의 신뢰성을 높일 수 있다. 둘째, 데이터의 스케일을 조정함으로써 가중치가 보다 효과적으로 반영될 수 있다.

변수 X 를 0~1 범위로 정규화하는 공식은 식 (3)과 같다.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

X : 원본 데이터 값

X_{max} : 해당 변수 X 의 최대값

X_{min} : 해당 변수 X 의 최소값

X_{norm} : 정규화된 변수 값 (0~1 범위)

최적의 α 값을 찾기 위한 알고리즘은 다음과 같이 구성된다.

1단계는 실험적 $\alpha \in \{1.0, 1.1, 1.2, \dots, 2.0\}$ 범위를 설정한다.

이 범위를 설정한 이유는 $\alpha \in \{1.0, 1.1, 1.2, \dots, 2.0\}$ 으로 설정한 이유는 기존 가중 상관계수($\alpha = 1.0$)와의 차이를 비교하고, 조정 강도 탐색 ($1.0 \leq \alpha \leq 2.0$) 탐색을 통해 이상치 영향 최소화와 데이터 보존의 균형점을 찾기 위함이다. 0.1 간격으로 세밀한 변화 분석을 통해 효과를 정밀하게 평가하며, α 값이 너무 크면 데이터 왜곡이 발생하고 너무 작으면 효과가 부족하기 때문이다.

2단계는 각 α 값에 대한 WAPCC 행렬 계산이다.

각 α 에 대한 WAPCC 행렬을 계산하고, 전체 상관계수의 평균을 측정한다. 그 계산식은 식 (4)와 같다.

$$M_{\alpha} = \frac{1}{n} \sum_{i=1}^n r_{ew}^{(\alpha)}(X, Y) \quad (4)$$

3단계는 가장 높은 평균 상관계수 M_{α} 를 제공하는 α 값을 선택한다. 계산식은 식 (5)와 같다.

$$\alpha^{\dagger} = \arg \max_{\alpha} M_{\alpha} \quad (5)$$

4단계는 최적의 α 를 선택하여 최종 WAPCC를 계산한다. 1~3단계를 거쳐 최적의 α 를 찾는 계산식은 식 (6)과 같다.

$$\alpha^{\ddagger} = \arg \max_{\alpha} \left(\frac{1}{n} \sum_{i=1}^n r_{ew}^{(\alpha)}(X, Y) \right) \quad (6)$$

각 α 값에 대해 WAPCC (r_{ew}) 계산은 모든 변수의 상관 행렬을 평균하여, 가장 높은 값을 제공하는 α 선택한다.

$$r_{ew} = \frac{\sum (w_i^{\alpha})(X_i - \bar{X}_{ew})(Y_i - \bar{Y}_{ew})}{\sqrt{\sum (w_i^{\alpha})(X_i - \bar{X}_{ew})^2} \cdot \sqrt{\sum (w_i^{\alpha})(Y_i - \bar{Y}_{ew})^2}} \quad (7)$$

X_i, Y_i : 변수 X 와 Y 의 개별 값

w_i : 가중치(weight), 특정 데이터의 중요도를 나타냄

α : 가중치 조정 파라미터(최적화 대상)

$\bar{X}_{ew}, \bar{Y}_{ew}$: 조정된 가중 평균(weighted adjustment mean)

가중치 조정 파라미터 α 의 역할은 다음과 같다.

첫째, $\alpha > 1$ (강한 가중치 차이 강조)일 때 가중치가 높은 데이터를 더욱 중요하게 반영한다. 최근 변동성이 크거나 심각도가 높은 데이터를 우선적으로 고려하는 경우 유용하다.

둘째, $0 < \alpha < 1$ (가중치 차이 완화)일 때, 데이터 포인트 간 가중치 차이를 줄여 분석의 균형을 유지하고 이상치의 영향을 완화하여 보다 안정적인 분석이 가능하다.

표 1은 전체 알고리즘의 의사코드이다. 1단계는 데이터 로드 및 전처리하는 단계로써 데이터를 불러오고 이상치(0 이하 값)를 제거하여 분석에 적합한 데이터를 구축하는 단계이다. 2단계는 가중치 계산 및 정규화하는 단계로써 두 변수를 Min-Max Scaling하여 0~1 범위로 정규화 하는 단계이다. 두 변수를 반영한 가중치를 계산하는 방법은 다음 수식 (8)과 같다.

$$w_i = 1 + 0.5 \times var1_{norm} + 0.5 \times var2_{norm} \quad (8)$$

표 1. 제안한 방식의 의사코드

Table 1. Pseudocode of the proposed method

```

Step1. Data Loading and Preprocessing
Function load_data(file_path):
    Load the data and keep only samples with "var1"
    and
    "var2" greater than 0
    Return Cleaned DataFrame
Step2. Weighted Calculation
Function normalize_weights(dataframe):
    Normalize "var1" and "var2" to a range [0,1]
    weight = 1 + 0.5 * var1norm + 0.5 * var2norm
    Return DataFrame with Added Weights
Step3. Calculation of Weighted Correlation Coefficient
Function weighted_correlation(X, Y, weight, alpha=2.0):
    Compute weighted mean of X and Y
    Compute weighted covariance of X and Y using weight^alpha
    weighted_corr = Covariance / (Standard Deviation of
    X *
    Standard Deviation of Y)
    Return weighted_corr
Step 4. Weighted Correlation Matrix Calculation
Function
compute_weighted_correlation_matrix(dataframe, alpha=2.0):
    variable_List = ["var1", "var3", "var4", "var2",
    "var5"]
    For each X, Y in variable_List:
        If X == Y:
            Weighted_Correlation_Matrix[X, Y] = 1.0
        Else:
            Weighted_Correlation_Matrix[X, Y] =
            wighted_corr(X, Y, dataframe["weight"], alpha)
    Return Weighted_Correlation_Matrix
Step 5. Execution and Result Saving
df = load_data("your_data.xlsx")
df = normalize_weights(df)
weighted_corr_matrix =
compute_weighted_correlation_matrix(df, alpha=2.0)
Save weighted_corr_matrix to
"weighted_correlation_result.xlsx"
Print completion message

```

3단계는 가중 상관계수를 계산하는 단계로, 기존 피어슨 상관계수를 확장하여 가중치 조정 파라미터 (α)를 반영한 상관 행렬 계산을 하고, 특정 그룹의 데이터를 강조하여 더욱 정밀한 관계 분석을 수행한다. 4단계는 가중 상관 행렬 계산을 하는 단계로, 모든 변수 간의 가중 상관 행렬을 생성한다. 마지막 5단계는 실행결과를 파일로 저장하는 단계이다.

본 논문에서 제안한 가중치 조정 파라미터 상관계수는 기존 가중 상관계수의 한계를 극복할 수 있도록 설계되었다.

특히, 가중치 조정 파라미터를 적용함으로써 데이터의 중요도를 동적으로 조정할 수 있으며, 이상치 영향을 최소화하면서 더욱 정교한 상관 분석이 가능하다. 이 방법론은 금융, 의료, 환경 과학, 머신러닝 등 다양한 응용 분야에서 데이터의 중요도를 반영한 분석을 수행하는 데 강력한 도구가 될 수 있다.

IV. 실험 및 분석

4.1 실험 데이터

본 논문에서는 지방간 발생 여부와 관련된 주요 생체지표를 포함한 실제 의료 데이터를 익명화된 상태로 분석을 수행하였다.

연구 대상은 총 497명이며, 모든 데이터는 결측치 처리를 거친 후 분석에 사용되었다. 각 변수의

기술 통계값은 표 2에 제시하였다.

4.2 실험 분석 및 결과

4.2.1 상관 분석 및 교차 분석 결과

상관 분석을 통해 수치형 변수 간의 관계를 살펴본 결과, 다음과 같은 경향이 확인되었다. 표 3은 변수 간의 피어슨 상관계수의 결과를 나타낸다.

간기능_GOT과 간기능_GPT 사이의 상관계수는 0.78로 강한 양의 상관관계를 보였고, 지방간 유무와 중성지방 사이의 상관계수는 0.31로 중등도 양의 상관관계를 나타냈다. 또한 BMI와 중성지방 사이의 상관계수는 -0.26로 약한 음의 상관관계를 나타냈다. 일반적으로 상관계수의 값이 -0.1~0.1의 경우는 두 변수 간의 상관관계가 매우 약하거나 거의 없다고 판단하여 분석 대상에서 제외하는 경우가 많다. 따라서 본 논문에서는 모든 변수간의 관계를 일일이 설명하지는 않는다.

표 2. 각 변수의 기술 통계값
Table 2. Descriptive statistics of each variable

Variable name	Explication	Unit	Average	Standard deviation	Minimum	Maximum
liver function_GOT	liver function test values(GOT)	U/L	xx.xx	xx.xx	X	X
liver function_GPT	liver function test values(GPT)	U/L	xx.xx	xx.xx	X	X
Diabetes	blood glucose level	mg/dL	xx.xx	xx.xx	X	X
triglycerides	blood triglyceride level	mg/dL	xx.xx	xx.xx	X	X
fatty liver presence	fatty liver presence (absent:0,present:1)	-	xx.xx	xx.xx	0	1
BMI	body mass index	kg/m ²	xx.xx	xx.xx	X	X
hemoglobin	hemoglobin level	g/dL	xx.xx	xx.xx	X	X

표 3. 변수 간 피어슨 상관계수
Table 3. Pearson correlation coefficients among variables

Variable name	Liver function_GOT	Liver function_GPT	Diabetes	Triglycerides	Fatty liver presence	BMI	Hemoglobin
liver function_GOT	1	0.78	-0.12	0.19	0.19	0.16	0.08
liver function_GPT	0.78	1	0.21	0.25	0.36	0.32	0.23
Diabetes	0.12	0.21	1	0.16	0.26	0.22	0.16
triglycerides	0.19	0.25	0.16	1	0.31	-0.26	0.27
fatty liver presence	0.19	0.36	0.26	0.31	1	0.24	0.29
BMI	0.16	0.32	0.22	-0.26	0.24	1	0.21
hemoglobin	0.08	0.23	0.16	0.27	0.29	0.21	1

표 4는 가중 상관계수의 결과를 나타낸다.

간기능_GPT와 지방간 유무 사이의 상관계수는 0.34로, 기존 피어슨 상관계수보다 조금 낮아졌지만 여전히 중등도 이상의 관계를 유지하였다.

$\alpha=2.0$ 을 적용함으로써 특정 변수(예: 간기능, BMI)의 가중치가 조정되어 보다 정밀한 상관 분석 가능했다.

따라서 연구나 분석 상황에 따라 적절한 상관 계수를 선택하는 것이 중요하며, 특정 그룹의 영향을 반영하고자 할 경우에는 가중 상관계수를 적용하는 것이 더 나은 선택이 될 수 있다.

표 5는 WAPCC의 결과를 보여준다. 최적의 가중치 조정 파라미터 α 값은 2.0으로 선정되었다. 이는 간기능_GPT와 체질량지수(BMI)를 가중치로 설정했을 때, $\alpha=2.0$ 이 가장 최적화된 결과를 제공하는 것으로 나타났다. $1.0 \leq \alpha \leq 2.0$ 범위에서 다양한 값을 실험한 결과, $\alpha=2.0$ 이 가장 최적의 분석 결과를 제공하는 것으로 나타났다. α 값이 1.0에 가까우면 기존 상관계수와 차이가 크지 않으며 이상치의 조정 효과가 미미하다. 반대로 α 값이 2.0을 초과하면 이상치 조정이 과도하게 이루어져 분석의 신뢰성이

떨어질 수 있다. 따라서 2.0은 이상치 조정과 데이터 반영의 균형을 가장 적절히 유지하는 값으로 판단되었다.

이 결과는 간기능과 체질량지수(BMI)가 높을수록 상관관계를 더 강하게 반영하는 것이 데이터 분석에 적합함을 의미한다.

따라서 간 기능과 체질량지수(BMI) 기반의 가중치를 적용한 최적 상관계수를 활용하면, 보다 의미 있는 해석이 가능할 것이다.

표 6의 분석 결과에서 간기능_GOT 값은 기존 방식보다 약간 증가하였다. 이는 기존 분석에서 충분히 고려되지 못했던 대사질환 환자의 영향을 보다 정확하게 반영한 결과이다. 반면 간기능_GPT, 중성지방, BMI, 혈색소 값들은 기존 방식보다 약간 감소하였다. 이는 기존 방식에서 과대평가되었을 가능성이 있는 관계를 보정함으로써, 보다 신뢰성 있고 균형 잡힌 분석 결과를 제공했기 때문이다. 결과적으로 WAPCC는 기존 방식보다 더 정밀한 상관관계 분석을 수행할 수 있는 방법으로, 연구 및 데이터 분석에서 보다 신뢰할 수 있는 결과를 도출하는 데 기여할 수 있다.

표 4. 가중 상관계수의 결과

Table 4. Results of weighted correlation coefficient

Variable name	Liver function _GOT	Liver function _GPT	Diabetes	Triglycerides	Fatty liver presence	BMI	Hemoglobin
liver function_GOT	1	0.77	0.15	0.17	0.18	0.14	0.05
liver function_GPT	0.77	1	0.21	0.24	0.34	0.29	0.19
Diabetes	0.15	0.21	1	0.16	0.25	0.20	0.14
triglycerides	0.17	0.24	0.16	1	0.31	-0.26	0.27
fatty liver presence	0.18	0.34	0.25	0.31	1	0.24	0.29
BMI	0.14	0.29	0.20	-0.26	0.24	1	0.21
hemoglobin	0.05	0.19	0.14	0.27	0.29	0.21	1

표 5. 가중치 조정 파라미터 상관계수 결과

Table 5. Weight adjustment parameter correlation coefficient results

Variable name	Liver function _GOT	Liver function _GPT	Diabetes	Triglycerides	Fatty liver presence	BMI	Hemoglobin
liver function_GOT	1	0.79	0.16	0.18	0.19	0.15	0.06
liver function_GPT	0.79	1	0.23	0.25	0.36	0.31	0.21
Diabetes	0.16	0.23	1	0.17	0.26	0.21	0.15
triglycerides	0.18	0.25	0.17	1	0.31	-0.26	0.27
fatty liver presence	0.19	0.36	0.26	0.31	1	0.24	0.29
BMI	0.15	0.31	0.21	-0.26	0.24	1	0.21
hemoglobin	0.06	0.21	0.15	0.27	0.29	0.21	1

그림 1의 분석에서는 세 가지 상관계수의 분석 결과를 시각적으로 비교하고 있다.

표 7은 세 가지의 상관계수의 전반적인 통계 결과를 나타낸 표이다.

표 6. 3개의 상관계수 분석 결과
Table 6. Results of three correlation coefficient analysis

Variable name	Pearson correlation coefficients	Weighted correlation coefficients	Weighted adjustment parameter correlation coefficient
liver function_GOT	0.1919	0.2053	0.2075
liver function_GPT	0.3667	0.3663	0.3533
triglycerides	0.2675	0.2435	0.2233
BMI	0.2435	0.2233	0.2150
hemoglobin	-0.7423	-0.7001	-0.6825

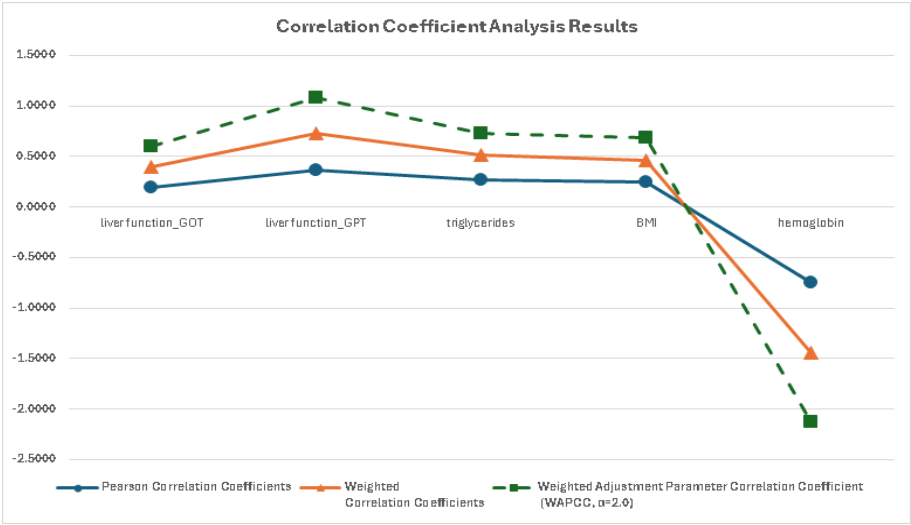


그림 1. 3개의 상관계수 분석 결과
Fig. 1. Three correlation coefficient analysis results

표 7. 3개의 상관계수의 전반적인 통계 결과
Table 7. Overall statistical results of the three correlation coefficients

Analysis method	Applied weight	α value	Outlier effect	Group highlighting feature availability	Resolution of analysis results
Pearson correlation coefficient	liver function test values(GOT)	N/A	highly affected	impossible	Measures the basic linear relationship between variables but does not reflect logic and the influence of specific group
Weighted correlation coefficients	liver function test values(GPT)	N/A	partial mitigation	possible (Fixed weight)	It is possible to analyze by emphasizing data with fatty liver, but the weighting is fixed
Weighted Adjustment Parameter Correlation Coefficient (WAPCC)	blood glucose level	optimized $\alpha=2.0$	maximum mitigation	possible(dynamic weighting)	A more precise analysis is possible by dynamically adjusting the weights based on liver function and BMI values

V. 결론 및 향후 과제

본 논문에서는 변수 간의 관계를 측정하는데 널리 사용되는 피어슨 상관계수, 가중 상관계수의 데이터 포인트 처리상의 한계를 보완하기 위해 WAPCC를 제안하였다.

WAPCC는 가중치가 높은 데이터의 영향을 조절하고, 이상치로 인한 왜곡을 최소화하기 위해 가중치 조정 파라미터(α)를 사용하였다. 실제 의료 데이터를 활용한 실증 분석을 통해 제안한 방식이 변수 간 관계를 보다 신뢰성 있게 분석할 수 있음을 검증하였다. 실험 결과, WAPCC는 α 값을 최적화함으로써 이상치의 영향을 줄이면서도 특정 데이터 그룹을 강조할 수 있는 최적의 값을 제공하였으며, 기존 상관계수보다 더욱 정교한 관계 분석이 가능함을 확인하였다. 또한 기존 피어슨 상관계수나 가중 상관계수와 비교했을 때, 데이터의 특성을 보다 정밀하게 반영할 수 있음을 보였다. 이를 통해 WAPCC는 변수 간 관계 분석에서 보다 신뢰성 있는 결과를 제공할 수 있는 효과적인 방법임을 입증하였다.

다만 본 연구는 단일 병원의 데이터를 사용하여 분석을 수행했기 때문에, 데이터 종류, 크기, 변수범위가 제한적이라는 한계를 가진다. 또한 상관관계 분석은 변수 간의 연관성을 측정하는 방법일 뿐, 직접적인 인과관계를 밝히는 것은 아니다. 따라서 향후 WAPCC의 활용 가능성을 더욱 확장하기 위해 다양한 데이터셋을 적용하여 그 일반화 가능성을 검증할 필요가 있다. 아울러 머신러닝 및 딥러닝 모델과 결합하여 보다 정교한 예측 모델을 개발하고, 실용적인 분석 기법으로 발전시킬 수 있도록 추가적인 연구가 요구된다.

References

- [1] G.p Noh, "Basic Statistics", Hanbit Academy, pp. 292-309, Dec. 2023.
- [2] Y. Park, "Correlation and Regression analysis", J Korean Acad Farm Med, Vol. 22, No. 1, pp. 43-51, Jan. 2001.
- [3] A. G. Asuero, A. Sayago, and A. G. González, "The Correlation Coefficient: An Overview", Critical Reviews in Analytical Chemistry, Vol. 36, No. 1, pp. 41-59, 2006. <https://doi.org/10.1080/10408340500526766>.
- [4] Veher Costa, "Correlation and Regression", Fundamentals of Statistical Hydrology, Springer, pp. 391-440, 2017. <https://doi.org/10.1007/978-3-319-43561-9>.
- [5] K.-Y. Jung, S.-Y. Choi, K.-W. Rim, and J.-H. Lee, "Preference Prediction System using Similarity Weight granted Bayesian estimated value and Associative User Clustering", Korean Institute of Information Scientists and Engineers, Vol. 30, No. 4, pp. 316-325, Apr. 2003.
- [6] B. Lee, "Improvement of the Semantic Information Retrieval using Ontology and Spearman Correlation Coefficients", The Society of Digital Policy and Management, Vol. 11, No. 11, pp. 351-357, Nov. 2013. <https://doi.org/10.14400/JDPM.2013.11.11.351>.
- [7] S. Kim, H.-S. Ahn, C. Yoon, J.-P. Chung, and S. S. Nam, "Development of Outlier Detection Method Using Statistical Analysis and Correlation Coefficient for Environmental Sensor Data", Journal of Advanced Navigation Technology, Vol. 28, No. 6, pp. 986-994, Dec. 2024. <http://dx.doi.org/10.12673/jant.2024.28.6.986>.
- [8] J. Cha and J. Kim, "Analysis of fine dust correlation between air quality and meteorological factors using SPSS", Journal of the Korea Institute of Information and Communication Engineering, Vol. 22, No. 5, pp. 722-727, May 2018. <https://doi.org/10.6109/jkiice.2018.22.4.722>.

저자소개

윤 중 찬(Jongchan Yun)



2006년 2월 : 부경대학교
산업대학원
전산정보학과(공학석사)
2008년 2월 : 부경대학교
전자상거래 협동과정(공학박사)
2005년 9월 ~ 현재 : 부경대학교
정보전산원 강사

관심분야 : 전산교육, 데이터마이닝, 데이터분석

윤 소 영(Soyoung Yun)



2007년 2월 : 부경대학교
경영대학원
국제물류학과(경영학석사)
2011년 2월 : 부경대학교
전자상거래 협동과정(공학박사)
2011년 9월 ~ 현재 : 부경대학교
정보전산원 강사

관심분야 : 추천시스템, m-commerce, 빅데이터