고혈압 예측을 위한 이상치 탐지 알고리즘 및 데이터 통합 기법

홍고르출 [1], 김미혜 [2], 송미화 [3†]
[1]가천대학교 컴퓨터공학과 교수
[2]충북대학교 컴퓨터공학과 교수
[3]세명대학교 스마트 IT 학부 교수
khongorzul63@gmail.com, mhkim@cbnu.ac.kr, mhsong@semyung.ac.kr

An Outlier Detection Algorithm and Data Integration Technique for Prediction of Hypertension

Khongorzul Dashdondov[1], Mi-Hye Kim[2], and Mi-Hwa Song[3†],
[1]Department of Computer Engineering, College of IT convergence, Gachon University
[2]Department of Computer Engineering, Chungbuk National University
[3]School of Smart IT, Semyung University
Corresponding author: Mi-Hwa Song (e-mail: mhsong@semyung.ac.kr)

Abstract

Hypertension is one of the leading causes of mortality worldwide. In recent years, the incidence of hypertension has increased dramatically, not only among the elderly but also among young people. In this regard, the use of machine-learning methods to diagnose the causes of hypertension has increased in recent years. In this study, we improved the prediction of hypertension detection using Mahalanobis distance-based multivariate outlier removal using the KNHANES database from the Korean national health data and the COVID-19 dataset from Kaggle. This study was divided into two modules. Initially, the data preprocessing step used merged datasets and decision-tree classifier-based feature selection. The next module applies a predictive analysis step to remove multivariate outliers using the Mahalanobis distance from the experimental dataset and makes a prediction of hypertension. In this study, we compared the accuracy of each classification model. The best results showed that the proposed MAH_RF algorithm had an accuracy of 82.66%. The proposed method can be used not only for hypertension but also for the detection of various diseases such as stroke and cardiovascular disease.

1. Introduction

Hypertension is a chronic disease that leads to heart, brain, kidney, diabetes, and other serious conditions [1], [2]. Hypertension, which occurs because of high blood pressure, is a condition in which blood vessels have constantly increased pressure. The human heart is difficult to pump because of the higher pressure. It is a major cause of early death worldwide, with up to one in four men and one in five women in over a billion people with hypertension [3]. South Korea is one of the countries in which hypertension is common [4]. In 2019, COVID-19 confronted the world. There have been many research activities related to the coronavirus disease of 2019 (COVID-19), and the use of machine learning methods to diagnose the causes of hypertension diseases has increased in recent years [5]-[6].

In this study, we aimed to infer the association between COVID-19 and hypertension by using ML methods to identify hypertension based on the characteristics of COVID-19. Machine learning is the process of learning that starts with observations or data, such as cases, real-world experience, or instructions, to look for patterns in the data and make better decisions in the future based on the examples that we supply. Machine learning helps make decisions automatically without a person using models learned from the data. In addition, it can be used to diagnose various diseases.

2. Methodology

In this section, we describe the components of our proposed prediction method. Figure 1 shows the proposed framework based on feature selection-based hypertension prediction method. The proposed framework consists of two main modules: data preprocessing and predictive analysis.

The data preprocessing and predictive analysis modules were implemented in Python using the Sklearn library [28]. The data preprocessing module was implemented using SPSS 23.0.

1) Data-preprocessing

Initially, we merged by region, age, and gender value for KHNANES and Kaggle data. Next, we removed a row of missing values and features unrelated to hypertension. After removing unrelated features with hypertension and missing

values from the dataset, 7965 records and 76 features were removed. We then removed attributes based on DT classifier, which included a total of 4926 records, 37 features for experimental dataset. Figure 2 shows the procedure for creating the target dataset.

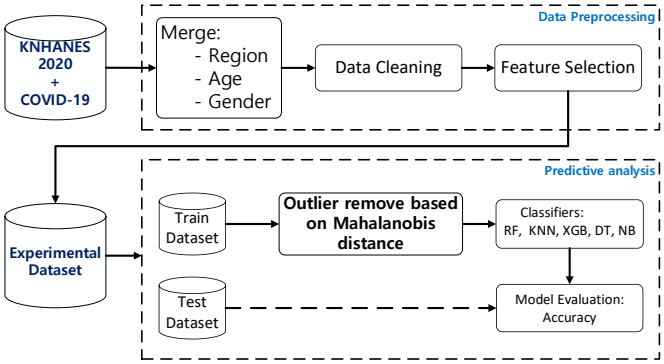

Figure 1. System architecture.

2) Outlier Detection Based on Mahalanobis Distance

Multivariate outliers can be identified with the use of Mahalanobis distance, which is the distance of a data point from the calculated centroid of the other cases where the centroid is calculated as the intersection of the mean of the variables being assessed. Each point is recognized as an X, Y combination and multivariate outliers lie a given distance from the other cases. The distances are interpreted using a $p < 0.001$ and the corresponding χ^2 value with the degrees of freedom equal to the number of variables [4].

Multivariate outliers can also be recognized using leverage, discrepancy, and influence. Leverage is related to Mahalanobis distance but is measured on a different scale so that the χ^2 distribution does not apply. Large scores indicate the case if further out however may still lie on the same line. Discrepancy assesses the extent that the case is in line with the other cases. Influence is determined by leverage and discrepancy and assesses changes in coefficients when cases are removed. Cases > 1.0 are likely to be considered outliers. It was introduced by Prof. P. C. Mahalanobis in 1936 and has been used in various statistical applications ever since. The formula to compute Mahalanobis distance is as follows:

$$D^2 = (x - m)^T C^{-1}(x - m) \quad (1)$$

where is the square of the Mahalanobis distance, x is the vector of the observation (row in a dataset), m is the vector of mean values of independent variables (mean of each column), is the inverse covariance matrix of independent variables, and (x – m) is essentially the distance of the vector from the mean, then divide this by the covariance matrix (or multiply by the inverse of the covariance matrix). P-value probability is shown as the following equation:

$$P = 1 - \chi^2(MAH, df) \quad (2)$$

3. Experimental results

1) Integrated dataset

In this study, the Korean National Health and Nutrition Examination Survey datasets were used to build a model for hypertension prediction. KNHANES data were collected by the Disease Control and Prevention (KCDC) [2]. It consists of a health examination of various diseases, health interviews, and nutrition surveys of the Korean population. The dataset duration was 2020 with COVID-19 [6]. We generated a target value for the upper 19-year-old patients with hypertension. Additionally, this target hypertension group included subjects who had a history of diabetes, prediabetes, heart disease, heart attack, and stroke.

First, we measured the performance of the baseline models for comparison with our proposed method. We trained the baseline models directly on the raw dataset using the machine learning algorithms shown in Figure 1. To investigate the correlation between COVID-19 and hypertension, we trained the ML-based classifiers on two different datasets: a dataset without COVID-19 features and a dataset with COVID-19 features. Predicting hypertension using COVID-19 features increased all the evaluation measures.

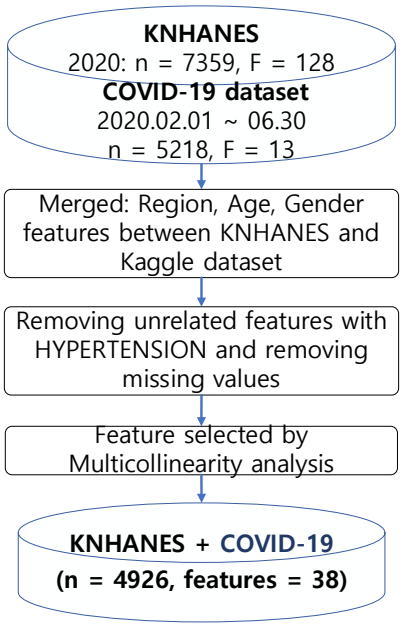

Figure 2. Experimental dataset.

2) Classifier results:

In other words, because the COVID-19 feature affects the prediction of hypertension, these two features can be related to each other. We also analyzed this relationship with the chi-square test, and the results are shown in Table I. The accuracy measurement of the performance results is shown in Table 1, and the highest values of evaluation scores are marked in bold. The RF with the MAH model achieved the highest accuracy of 82.66%. As can be seen, KNN and NB based both predictive models performed lower results compared with other predictive models in terms of the evaluation metrics. The receiver operating characteristic

(ROC) curve is one of the important evaluation metrics for evaluating detection performance.

Table 1. Results of accuracy analysis of two methods on the experimental dataset.

	Accuracy	
	Without outlier	With outlier
KNN	58.70	58.31
DT	73.08	72.15
RF	82.66	82.47
NB	60.10	58.33

Finally, Figure 3 and Figure 4 show the AUC score of the proposed method was improved by MAH -based feature, the accuracy of the ML-based KNN, DT, RF, and NB approaches improved by 0.15%, 0.14%, 0.15%, and 0.02%, respectively.

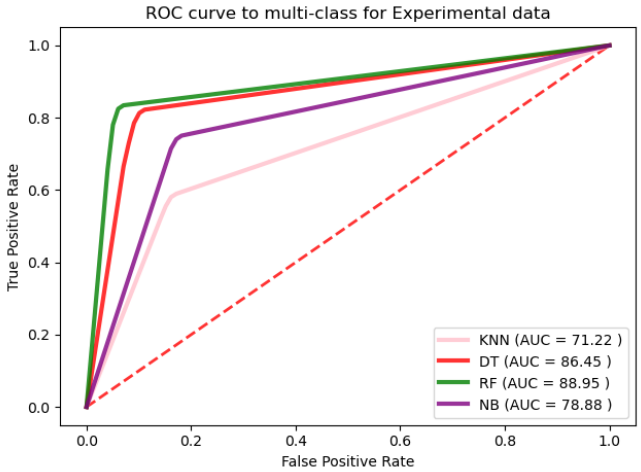

Figure 3. The ROC curves of compared algorithms on experimental dataset.

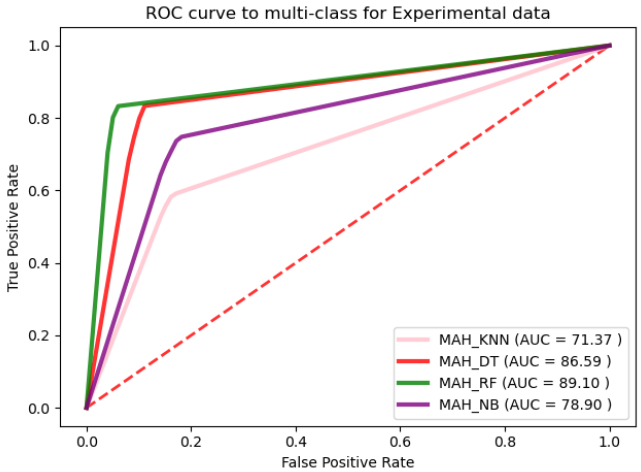

Figure 4. The ROC curves of compared algorithms on outlier removed dataset.

4. Conclusion

This study proposed a method consisting of two modules for predicting hypertension. First, external features of KHNANES and Kaggle data for COVID-19 patients were aggregated by region, age, and gender, followed by data preprocessing by performing DT-based feature selection. The next, module predictive analysis uses to predict hypertension based on multivariate outlier removal using a Mahalanobis distance from integrated experimental data. In this study, we compared the accuracy and AUC of each classification model. Four classifiers were compared in this experiment. The evaluation results showed that the proposed method improved the accuracy, and AUC of KNN, DT, RF, and NB by (0.39, 0.93, 0.19, 1.77), and (0.15, 0.14, 0.15, 0.02), respectively. Moreover, by using MAH_RF our proposed method, we successfully increased the predictive performance of the classifiers used in all experiments by preparing a high-quality training dataset. We experimentally demonstrate how the steps of our proposed method improve performance.

References

[1] Korea Centers for Disease Control & Prevention. Online, Available: http://knhanes.cdc.go.kr.

[2] C. Wang et al., A novel coronavirus outbreak of global health concern, Lancet. 2020: 395: 470-473

[3] World Health Organization. Online, Available: https://www.who.int/health-topics/hypertension/#tab=tab_1

[4] D. Khongorzul, and M.H. Kim, Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction, Neural Processing Letters, 2023: 55: 265-277

[5] K. Song, et al., Change in Prevalence of Hypertension among Korean Children and Adolescents during the Coronavirus Disease 2019 (COVID-19) Outbreak: A Population-Based Study, Children, 2023: 10: 159.

[6] J. Kim, et al., DS4C patient policy province dataset: a comprehensive COVID-19 dataset for causal and epidemiological analysis, in Proc. NeurIPS 2020

[7] NeurIPS 2020: data science for COVID-19, DS4C: data science for COVID-19 in South Korea, San Francisco: Kaggle, 2020. Accessed 2021 Mar 11. Online available: https://www.kaggle.com/kimjihoo/coronavirusdataset.

Roger S. Pressman "Software Engineering A Practltiners' Approach" 3rd Ed. McGraw Hill