한국 남성의 고혈압에 대한 특징 선택 기반 위험 예측

홍고르출 [1], 김미혜 [1]*
[1] 충북대학교 컴퓨터공학과
khongorzul63@gmail.com, mhkim@cbnu.ac.kr*

Feature selection-based Risk Prediction for Hypertension in Korean men

Khongorzul Dashdondov[1], and Mi-Hye Kim[1*]
[1] Department of Computer Engineering, Chungbuk National University,
Chungbuk 28644, Korea

Abstract

In this article, we have improved the prediction of hypertension detection using the feature selection method for the Korean national health data named by the KNHANES database. The study identified a variety of risk factors associated with chronic hypertension. The paper is divided into two modules. The first of these is a data pre-processing step that uses a factor analysis (FA) based feature selection method from the dataset. The next module applies a predictive analysis step to detect and predict hypertension risk prediction. In this study, we compare the mean standard error (MSE), F1-score, and area under the ROC curve (AUC) for each classification model. The test results show that the proposed FIFA-OE-NB algorithm has an MSE, F1-score, and AUC outcomes 0.259, 0.460, and 64.70%, respectively. These results demonstrate that the proposed FIFA-OE method outperforms other models for hypertension risk predictions.

Keywords: KNHANES, Hypertension, risk prediction, feature importance, feature selection

1. Introduction

Hypertension is a serious medical condition and can increase the risk of heart, brain, kidney, and other diseases [1-2]. It is a major cause of premature death worldwide, with upwards of 1 in 4 men and 1 in 5 women over a billion people having the condition [1].

In recent years, the incidence of hypertension diseases 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 diseases has increased in recent years. Machine learning is the process of learning that begins with observations or data, such as examples, direct experience, or instruction, to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly. [2] have done a hybrid feature selection analysis to accurately classify hypertensions. They applied a Bayesian network algorithm on selected features of the KNHANES 2007–2014 dataset. Our previous related work [4] has implemented a Mahalanobis-based multivariate outlier removing approach to improving the performance of the predictive analysis. By this research work, we have proposed the feature selection based on a FA was conducted with social and demographic characteristics in relation to hypertension then detected with greater accuracy with the commonly used ML algorithm. Another advantage is that the data has normalized by the OE method resulting in theoretically a distribution of the data identical to that in earlier work [3].

2. Materials and Methods

In this section, we will describe the components of the proposed prediction method. Fig. 1 shows the proposed framework based on FA based FS method. The proposed framework consists of two main modules including data preprocessing and predictive analysis.

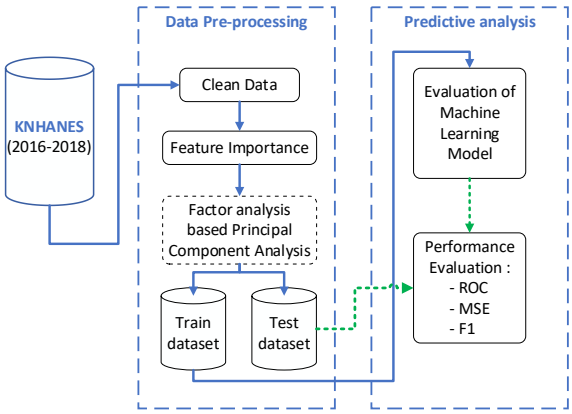

Fig. 1. The experimental architecture of the proposed method.

2.1. Data pre-processing.

This model includes two main parts feature importance and FA-based PCA feature selection.

Feature importance: We applied a Decision tree (DT) classifier in the machine-learning package to selected important features in hypertension. In this step, the features will be select if their importance score is greater than zero. Our proposed DT-based important features interpretability prediction model across the KNHANES dataset has described in Fig. 2. For the KNHANES dataset, “Vitamin A”, “Niacin”, “Water”, “Monounsaturated fatty acid”, “Retinol”, “n3 fatty acid”, “income”, “iron”, “blood sugar” and “potassium” were maintained as most useful features with importance scores of 0.3472, 0.18334, 0.11692, 0.11592, 0.05683, 0.01777, 0.01666, 0.01561, 0.01318, and 0.01306 to predict DT among the Korean men as shown Fig. 2.

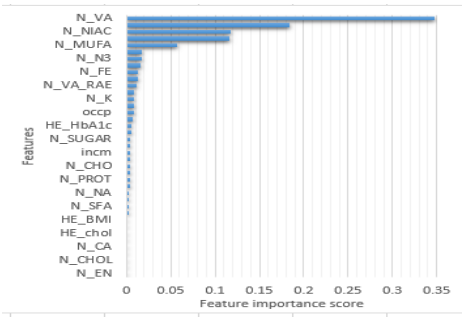

Fig. 2. Feature importance of the DT based feature selection in the KNHANES dataset.

Feature Selection: Factor analysis is a technique used to reduce a huge number of variables to a few factors. This method calculates the most common variance of all variables and scores them. In this paper, we used the most common method for principal component analysis for the extracted factor from the data set. PCA extracts the maximum variance and sets them into the first factor [6]. The variance explained by the first factor is then calculated, and then the maximum variance of the second factor is calculated. This process moves to the last factor. This model has determinant = 0.022, no multicollinearity, KMO factor = 0,843 sample sizes are compatible, and significant has equal to 0.001 (alpha < 0.005), rejected the null hypothesis, therefore shows the correlation between the variables. Here, we can perform a factor analysis. Table 1 shown the Rotated component matrix for the suggested factor model. There also we extracted the PCA model and rotated varimax with the Kaiser normalization method converged in 7 iterations and five components were extracted.

2.2. Predictive analysis.

To improve the performance of predictive analysis, we focused on the training dataset. In other words, instead of directly train classifiers, we reduced features dimension from the training dataset using by PCA based FA. After that, the prepared training dataset applied KNN, DT, RF, and NB algorithms [3-5].

Table 1. Rotated Component Matrix

Features	Component				
	1	2	3	4	5
N_FAT	.882				
N_PUFA	.874				
N_N6	.860				
N_MUFA	.826				
N_PROT	.653	.594			
N_N3	.615				
N_CHOL	.613				
N_B2	.603	.547			
N_K		.844			
N_CHO		.798			
N_PHOS	.580	.716			
N_FE		.656			
N_SUGAR		.645			
N_VITC		.616			
N_NIAC	.530	.578			
N_NA		.571			
incm			.925		
ho_incm			.910		
age				.774	
HE_HPdg				.729	
HE_glu				.556	
HE_TG					.778

3. Experimental Results

The datasets were used to propose a model for risk predictions of hypertension in experimental data from earlier work [3]. This is referred to here as “KNHANES” the for open data [1]. Initially, we removed a row of missing values and features unrelated to hypertension, after which there were a total of 3840 records from 7870 records originally, and 40 features from 106. Next, we selected important features DT based model which there a total 26 features from 40. Afterward, we selected features according to the FIFA-OE model of hypertension risk detection, in this case five features.

Table 2. Evaluation comparison of the proposed methods for experimental target dataset.

Algorithms	*MSE*	*F1-score*	*AUC*
FIFA-OE- KNN	0.296	0.527	**64.61**
FIFA-OE-DT	0.339	0.498	61.75
FIFA-OE-RF	0.280	0.486	63.59
FIFA-OE-NB	0.259	0.460	**64.70**
KNN	0.291	0.428	61.01
DT	0.424	0.195	49.98
RF	0.423	0.194	50.00
NB	0.370	0.340	55.95

The MSE, F1-score, and AUC measurements of the performance results are shown in Table 2, and the highest values of evaluation scores are marked in bold. If we did not use FIFA-OE method to predict hypertension, the KNN algorithm showed the highest performance. The NB with the FIFA-OE model achieved the highest MSE of 0.259, and AUC of 64.7%. Following it, the second-best MSE of 0.296, F1-score of 52.7%, and AUC of 64.61% were achieved by KNN with the FIFA-OE. As can be seen, DT and RF based both predictive models performed lower results compared with other predictive models in terms of the evaluation metrics.

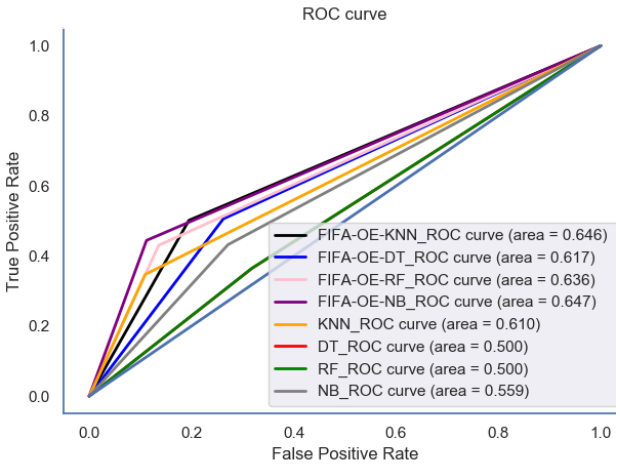

Fig. 3. Comparison of ROC area of prediction algorithms

It can be clearly seen from Table 2 that 26 dimensions were reduced to 5 dimensions by factor analysis, the AUC of hypertension risk detection increased in all algorithms. This shows that the proposed factor analysis-based feature reduction method is suitable for the risk prediction of hypertension.

The ROC curve and AUC are one of the important evaluation metrics for evaluating detection performance. We provided ROC curves for each class on the experimental dataset Fig. 3.

4. Conclusion

In this article, we have improved hypertension prediction using feature selection based on a factor analysis of actual open data of Korean national health data. The analysis revealed various risk factors associated with hypertension. These important features were selected DT classifier, after that, extracted from a PCA-based factorial analysis model. The analysis found that hypertension was related to social and demographic characteristics, such as age, sex, household income, education, occupation, as well as BMI, glucose, hemoglobin, cholesterol, HDL, triglycerides, and other health factors. In addition, hypertension was associated with nutritional information such as energy, water, protein, fat, and others. According to the test results, the proposed FIFA-OE-NB algorithm has an MSE, F1-score, and AUC outcomes 0.259, 0.460, and 64.70%, respectively. The system has implemented SPSS and Python, including its performance, is tested on real open data.

Acknowledgment

This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE) of Korea under the “Regional Specialized Industry Development Program” (R&D, P0002072) supervised by the Korea Institute for Advancement of Technology (KIAT). Other support came from the MSIT (Ministry of Science and ICT) of Korea under the Grand Information Technology Research Center support program (IITP-2020-1711120023) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

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

[2] Park, H.W., Li, D., Piao, Y. and Ryu, K.H., “A hybrid feature selection method to classification and its application in hypertension diagnosis”, Inter. Conf. on Information Tech. in Bio-and Medical Informatics (pp. 11-19). 2017, August. Springer.

[3] Khongorzul Dashdondov, Mi-Hye Kim. “Multivariate Outlier Removing for the Risk Prediction of Gas Leakage based Methane Gas”, Journal of the Korea Convergence Society. 11(12), pp. 23-30, 2020.

[4] Khongorzul Dashdondov, Mi-Hye Kim. “Prediction of Hypertension in Korean Men using the Outlier Detection Method”, Int. Conf. on the Multimedia and Ubiquitous Engineering (MUE2021), April 22-24, 2021, Jeju, Korea.

[5] Chang, W.; Liu, Y.; Xiao, Y.; Yuan, X.; Xu, X.; Zhang, S.; Zhou, S. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. Diagnostics 2019, 9, 178.

[6] Vapnik, V. N.: The nature of statistical learning theory., Springer, New York (1995).