HYBRID METHOD FOR EVALUATING FEATURE IMPORTANCE FOR PREDICTING CHRONIC HEART

Diseases Rashid Nasimov
Artificial Intelligence,Tashkent State University of Economics
Tashkent, Uzbekistan
rashid.nasimov@tsue.uz

Abstract

Predicting the impact of different factors on the patient's health is as important as diagnosing diseases, especially when monitoring patients with chronic diseases. To perform this by Artificial Intelligence (AI) methods, it is recommended to determine the features importance (FI) of data. There are a number of methods to evaluate FI. However, we can see a big variation in their results which is difficult to interpret. To solve this issue, we proposed new method which aim is minimizing the differences. Furthermore, to demonstrate the effectiveness of the proposed method we used the extracted FIs as weights of the weighted KNN and compared performances.

Keywords: Artificial Intelligence, chronic heart diseases, feature importance, disease prediction, decision tree, KNN.

Introduction

Using AI algorithms in medicine opens up huge opportunities for the different domain of healthcare. As a consequence, todays, various algorithms of AI are widely used in the direction of disease (e.i. cancer, cardiovascular and skin diseases) detection, interpretation and segmentation of medical images as well as classification of diseases [1-3]. However, in healthcare, there are cases when it is more important to determine the factors (drug, food, physical activity etc.) affecting the patient's condition and to determine the degree of influence of these factors than to diagnose the disease. In particular, the monitoring of chronic diseases is one of this crucial case. Because, in the process of daily monitoring of the patient, it is necessary to determine what factors and to what extent cause its condition to improve or to worsen. To solve such kind of problem, machine-learning methods of evaluating the importance of features in the dataset has been used.

To be more precise, let's develop a dataset of factors affecting the patient's condition, these factors are called features. These features are used to classify this dataset. Depending on the type of features, they can have different effects on classification accuracy. The influence of some features can be very strong (that is they have great feature importance) while some may have almost no importance for classification task.

With the development of machine learning methods, several feature importance determination methods have been developed, which will be discussed in detail in Section II. However, the problem is that these methods produce different importance values for the same feature. For their intended use in medicine, they must have identical feature importance with little variation. Otherwise, these calculated values will not have any value, or may lead to an incorrect medical conclusion [4]. This is considered very dangerous for human health. Therefore, it is a very important task to choose a feature importance estimation method that is suitable for medical use.

Therefore, in this paper, a new method was proposed to choose a feature importance estimation method for medical aims. For this reason, the FI evaluated by Logistic Regression (LR) and Decision Tree (DT) algorithms was converted into other values using special expressions. In order to determine the effectiveness of the proposed method, the resulting values was entered as weights for scaling the data in the dataset. The KNN network was trained with this scaled data. To show the advantage of this method over the LR and DT methods, the KNN was trained by scaling the data with the values determined by these methods and compared the results of three cases.

This paper is structured as follows. The first section is an introduction, and II section provides information on methods for determining feature importance, and provides brief information on LR and DT methods. III section considers research works on increasing the accuracy of KNN through feature importance. IV section provides information about the dataset used. In V section, proposed method is discussed. And final section, VI section, covers information about the verification the effectiveness of the proposed methods

Feature Importance Estimation Methods

Since feature selection methods are based on machine learning methods, they can be divided into three large groups: supervised, unsupervised and semi-supervised. In turn, each of these three groups is divided into four classes according to the evaluation criteria [4, 5].

1) Filters methods. These methods measure the relevance of features by their correlation with dependent variable. Missing value, information gain, square test and Fisher’s score are counted as most popular filter methods.

2) Wrappers methods. These methods estimate the usefulness of a subset of feature by actually training a model on it. It includes methods like: forward feature selection,

backward feature selection, exhaustive feature selection and recursive feature elimination.

3) Embedding methods. These methods integrate the quality of two above mentioned methods. In this method feature selection process is embedded in the learning or the model building phase. Most common examples are Regularization L1, L2, Random forest importance.

4) Hybrid methods. Sometimes using the advantages of several different methods gives better result than using a single method. The method of determining FI using several different methods is called hybrid method. F**uzzy random forest**-based **feature selection, hybrid genetic algorithms, hybrid and colony optimization, or mixed gravitational search algorithm are good examples of this method.** Some of these methods are designed to solve classification problems, while others are used to solve regression problems.

Since the assessment of feature importance in medicine is often accompanied by such tasks as predicting the patient's condition, classifying the severity of the disease. In thispaper has been considered only two of the most suitable methods for classification: logistic regression and random forest decision tree methods.

Logistic regression

Unlike linear regression, linear function isn't used in Logistic regression, but, instead it, sigmoid function is used as a regression function. And, since there are only 2 classes in the dataset, only the issue of classifying data into 2 classes are seen. Let $P_{(y=1)}$ be the probability of data belonging to the first class and $P_{(y=0)}$ be the probability of belonging to the second class, then the ratio of these probabilities is determined as follows:

$$\frac{P_{y=1}}{P_{y=0}} = \frac{P_{y=1}}{(1-P_{y=1})} \qquad (1)$$

Here, if it takes into account that $P_{(y=1)} = [\, 1 / (1 + e^{-z})]$ and put this value into the (1) formula and simplify it, the following expression is obtained:

$$\frac{P_{y=1}}{(1-P_{y=1})} = e^{Z} \qquad (2)$$

Here $z = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$. If this value put in the (2) formula, the following expression is obtained:

$$\frac{P_{y=1}}{P_{y=0}} = e^{(w_0+w_1x_1+w_2x_2+w_3x_3+w_4x_4)} \qquad (3)$$

If the Euler numbers on the right side of this equation are extracted, these coefficients can be considered as a feature importance of appropriate feature.

Decision tree

Although decision tree can be used to solve classification and regression problems, here we decided to use decision tree only for classification. Usually, when solving classification problems, the Gini impurity criterion is used to extract the desired feature or divide it into classes. Gini impurity is determined by the following formula:

$$\sum_{i=1}^{c} f_i(1 - f_i) \qquad (4)$$

Here, f_i is the frequency of label; i is at a particular node; while C is the number of unique labels. In this method, determining the feature importance in the dataset is based

on measuring how much the impurity criterion decreases. More precisely, if the tree is a binary tree (just like in our case), the Gini importance is calculated as follows, taking into account only two child nodes:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \qquad (5)$$

Here, ni_j – the importance of j node; w_j – weighted number of samples reaching j node; C_j – the impurity value of j node; left(j) – child node from left split on j node; right(j) – child node from right split on j node. Using these n values, the FI for each feature can be determined by the following expression:

$$fi_i = \frac{\sum_{j:\text{node j splits on feature i}} ni_j}{\sum_{k\in\text{all nodes}} ni_k} \qquad (6)$$

Here, fi_i – the importance of feature i; ni_j – the importance of j node; this method is widely used because it is simple and easy to use. However, this way of determining FI also has a number of disadvantages. For example, it tends to overestimate the importance of a continuous numeric category and shows poor performance to work with huge dataset.

Related Works

We aimed to use weighted KNN algorithm (i.e., using FI coefficients as a weights) to evaluate proposed method's effectiveness. Because, the accuracy of KNN is strongly dependent on the distance between features, several scientific studies have been conducted on feature scaling [6-11]. These studies divided into two groups, the first group of studies was conducted on the development of feature-sensitive dynamic custom metric [6, 7], and the second group proposed scaling of features [8-11]. Custom metric allows to change the distance as you like, but it is a much slower algorithm than the built-in methods. Even if the Sython library is used to increase the speed, it takes hundred times more time than the built-in methods. If we take into account that KNN is lazy learning, that is, it is a method that spends little time on training and a lot of time on testing, the time of classification with custom distance increases even more. For time consuming reason, this method is considered ineffective.

In terms of speed and feature scaling are considered much faster. To date, several methods have been proposed to determine the necessary coefficients for feature scaling. In particular, [8, 9] was determined based on the Random Forest Model. However [8] is suggested to get weights. In this case, the difference between the errors that occurred when Random forest was trained with and without a certain feature was determined. For each feature, the sum of the error differences detected in all trees was divided by the number of trees, and the average error was determined. This average error is taken as feature importance and multiplied by the corresponding feature. It can be seen a similar approach in [9]. In general, the research done in recent years, feature importance is the most widely used method as a multiplying coefficient. However, what caught our attention is that, despite the fact that there are many methods for determining FI, most of the studies used the random forest DT method. This may be due to several advantages of DT. However, the conclusion made in [5] is very important. The authors compared the random forest DT and LR methods and found that although the random forest DT method is better than the LR method in extracting some of the most important features. These features can still change their importance depending on the algorithm used.

Dataset

In the United States, telephone surveys are conducted annually by the Centers for Disease Control, and these data are systematized to create the BRFSS dataset. To perform our research, the database was used provided by [12] and prepared based on the 2015 database of BRFSS. This dataset is a modified representation of the BRFSS dataset, the original dataset contains data from 441,455 patients, each data set consists of 330 features. The database contains numerical and textual data. But in the modified database, only 253,680 of these data were selected and only the 21 features considered most important were extracted. Also, these features are presented in numerical form convenient for classification. There are no null values and the data belong to 2 classes, people with and without heart disease. But these two classes have a huge difference, that is, about 9.4% of participants had heart disease to 90.6% without heart disease. In particular, KNN networks are very sensitive to imbalanced data, and the accuracy of multi-class data is high, that is, they often identify data as belonging to a large number of classes. Taking this into account, we tried to equalize the data. 27,392 healthy patients and 23,893 patients with heart disease from the database was selected. Then, 51285 pieces of data-based dataset was developed. After that, we divided the data into training and test set in the ratio of 8:2. Since the values of the data differ sharply from each other, normalized has done using L2 norm.

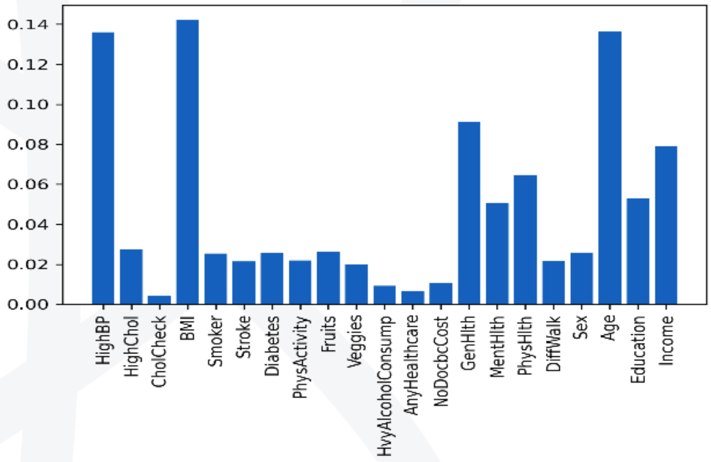
Proposed Methods

Unlike previously presented methods, feature importance determined was used in several different ways as feature scaling weights and compared the performance of KNN in each of them. The values determined by each method are illustrated in Fig.1.

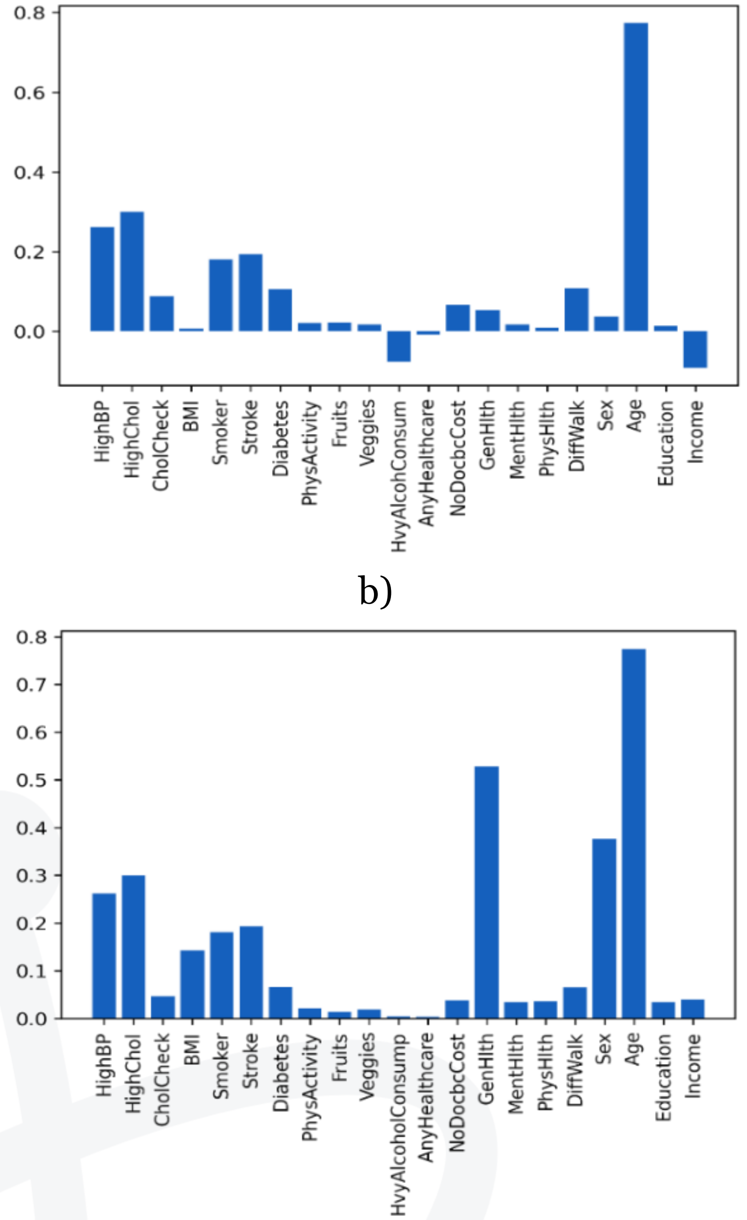
As can be seen from the graphs, the results of each algorithm are different. For example, BMI was considered as an important feature by DT while it was found to be not so important by LR. Moreover, the most important aspect is that the values considered to be the most important by these algorithms are defined as less important features in real medical case. For example, in [13], 49.5-year-old and obese people were observed for 10.9 years. It has been studied that in obese people, only the increase of Body Mass-Index (BMI) over the years can lead to cardiovascular diseases, while stable BMI does not cause cardiovascular heart diseases (CHD). A more interesting conclusion was reached in [14]. A very large database was collected in this research work (9 278 433 people of different ages were observed for 8,2 years). In this case, it was found that the effect of BMI on CHD was different for different ages. It can be concluded that high BMI is not always the cause of CHD, on the contrary, it can have the opposite effect in a certain age range (40-64). However, in the LR method, this feature is estimated to be more important than even the most important factor - age. In the DT method, on the contrary, its importance is underestimated.

As it can be seen above, each algorithm uses different aspects to determine FI. Therefore, it is needed to check whether it is possible to show a better result by selecting the advantages from the results of both algorithms. To do this, the following operation was performed on FI values with a large difference between them:

$$F_i = \begin{cases} F_i^{I} \text{ if } F_i^{I} > F_i^{II} \\ F_i^{II} \text{ if } F_i^{I} < F_i^{II} \end{cases} \qquad (7)$$



a)



b)

c)

Fig.1. Values of feature importance determined by
a) decision tree, b) logistic regression, c) proposed method

For some feature whose values detected by two methods with very little difference, the following operation is performed:

$$F_i = (F_i^I + F_i^{II})/2 \qquad (8)$$

To use these two formulas, it was necessary to choose a threshold value that determines whether the difference between the values is large or small. In this case, the threshold value was chosen as 0.1 value.

Verification The Effectiveness of the Proposed Methods

The KNN network was used to check the effectiveness of the proposed method. To be more precise, the obtained values was entered as weights into the KNN network and took the classification accuracy of the KNN network as an evaluation criterion.

For this, first of all, we determined the most important parameter of this network, which is the number k at which the network works most efficiently. It is known that the KNN algorithm depends on the number of k neighbors, so the performance of the KNN model was checked by changing the number of k from 1 to 8. The result is shown in Fig.2.

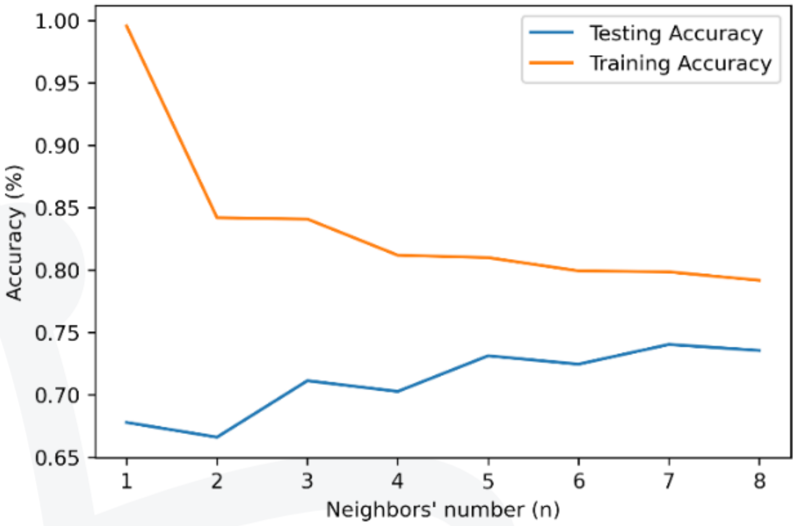


Fig.2. Training and test accuracy values of KNN algorithm for different k numbers. The red and the blue lines are line training and test accuracy, respectively

The Manhattan method was used as a distance metrics to train the KNN network to perform our experiment. As can be seen from the graph, the highest accuracy is with k=7. That's why we basically tried to get this value. After that, using logistic regression and decision tree, the extracted coefficients into features were multiplied.

$$X_i = Y_i \cdot Z_i \qquad (9)$$

Here, X_i -is the new value of the generated feature, Y_i -is the value of the first feature in the data, Z_i -is the determined impotance coefficient of the first feature. After that, the KNN network was trained with the new values generated, and the obtained results are presented in Tab.1.

TABLE I. ACCURACY OF KNN NETWORK FOR CASES WHERE FI VALUES EXTRACTED USING DIFFERENT METHODS ARE USED AS WEIGHTS

Values of k	**The Name of the Coefficient**			
	Logistic tree	**Decision tree**	**Original value**	**Proposed method**
1	67,2	65,6	66,6	**67.6**
2	72	70,4	71,1	**72.7**
3	70,9	69,9	70,3	**71.4**
4	73,1	72,2	73,1	**73.8**
5	72,7	71,7	72,5	**73.3**
6	73,6	73,1	74	**74.9**
7	73,4	72,8	73,6	**74.3**

As can be seen from the table, the coefficients allocated by logistic regression increased the accuracy of KNN merely less than 1%, and Decision tree, on the contrary, caused a decrease in the accuracy of the network. In the case of the proposed method, a greater result was obtained than both methods. But this value was very close to the coefficients obtained by the logistic tree method. The main reason for this is that the main part of the FIs identified with a large difference in the two methods was identified as having a large value in the Logistic regression method, while in the Decision tree, on the contrary, these features were identified as insignificant values.

When the literature to check whether the values determined was turned by the proposed method are medically compatible with practice. It was found that age is indeed important, while importance of BMI is not very high, nor very low.

Conclusion

In this paper, the ability of LR and DT algorithms to distinguish FI as well as the proposed hybrid method that is more effective than both methods for medical use and tested its effectiveness using the KNN algorithm. Indeed, the values obtained by the proposed method were found to be suitable for real medical situations.

Acknowledgment

This work was supported by the ИЗ-2020082913 project funded by the Ministry of Innovation Development of the Republic of Uzbekistan (MIDRU). We gratefully acknowledge the financial support of the MIDRU.

REFERENCES

[1] P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, “AI in health and medicine,” Nat Med, vol. 28, pp. 31–38,January 2022.

[2] N. Nasimova, B. Muminov, R. Nasimov, et al “Comparative Analysis of the Results of Algorithms for Dilated Cardiomyopathy and Hypertrophic Cardiomyopathy using Deep Learning,” ICISCT, 2021, pp. 1-5

[3] A. Turgunov, K. Zohirov, R. Nasimov and S. Mirzakhalilov, “Comparative Analysis of the Results of EMG Signal Classification Based on Machine Learning Algorithms,” ICISCT, 2021, pp. 1-4

[4] D. Jain, V. Singh, “Feature selection and classification systems for chronic disease prediction: A review,” Egyptian Informatics Journal, Vol. 19, pp. 179–189, April 2018.

[5] M. Saarela, S. Jauhiainen, “Comparison of feature importance measures as explanations for classification models,” SN Appl. Sci., Vol 3, Art. Num. 272, February 2021.

[6] Wu, Z. Cai, "Dynamic K-Nearest-Neighbor with Distance and attribute weighted for classification," ICEIE, 2010 International Conference, Vol. 1, September 2010.

[7] Ch. Zhang, P. Zhong, M. Liu, Q. Song, Zh. Liang, X. Wang, “Hybrid metrics K-nearest neighbor algorithm and applications,” Mathematical Problems in Engineering, Vol 2022, Art. ID 8212546, January 2022.

[8] C. Zhu, R. Hou and X. Ding, "An Improved Hybrid RFVIM-KNN Method for High Dimensional Data," 2016 8th International Conference on IHMSC, 2016, pp. 164-167

[9] C. A. Bhardwaj, M. Mishra, K. Desikan, “Dynamic Feature Scaling for K-Nearest Neighbor Algorithm,” ArXiv, abs/1811.05062, 2018

[10] S. Basak and M. Huber, "Evolutionary Feature Scaling in K-Nearest Neighbors Based on Label Dispersion Minimization," 2020 IEEE International Conference on SMC, 2020, pp. 928-935

[11] D. Li, B. Zhang, C. Li, “A Feature-Scaling-Based k-Nearest Neighbor Algorithm for Indoor Positioning Systems,” IEEE Internet of Things Journal, Vol 3, pp. 590-597, August 2016

[12] A. Teboul, M. Limam, F. S. Rizqi, “Heart disease health indicators dataset”, Version 3, November 2021

[13] B. Iyen, S. Weng, Y. Vinogradova, “Long-term body mass index changes in overweight and obese adults and the risk of heart failure, CVD and mortality: a

cohort study of over 260,000 adults in the UK" BMC Public Health, Vol.21, Art. num.576, April 2021

[14] H. J. Lee, H. K. Kim and et al, "Age-dependent associations of body mass index with myocardial infarction, heart failure, and mortality in over 9 million Koreans," European Journal of Preventive Cardiology, May 2022;, zwac09.