COMPARISON OF SEVERAL MODELS FOR CARDIOVASCULAR DISEASES PREDICTION

Wang Yue,
Urumqi University, Urumqi, China, wynfnone@gmail.com

Lilia I. Voronova,
Moscow Technical University of Communications and Informatics, Moscow, Russia, voronova.lilia@ya.ru

Vyacheslav I. Voronov,
Moscow Technical University of Communications and Informatics, Moscow, Russia, vorvi@mail.ru

DOI: 10.36724/2664-066X-2020-6-6-24-28

ABSTRACT

Due to the rapid development of economy, science and technology, the pace of life of people has accelerated and their standard of living has increased. At the same time, the number of various chronic diseases, such as cardiovascular, cerebrovascular and chronic heart diseases, is increasing. These problems seriously affect people's quality of life. Therefore, the problem of predicting cardiovascular diseases has become extremely urgent. The article compares several models for predicting heart disease and evaluates quality of their prognosis.

KEYWORDS: *cardiovascular diseases, random forest, k-nearest neighbors, naive Bayesian classifier, decision tree, artificial neural network, disease prognosis*

INTRODUCTION

According to the World Health Organization, "Cardiovascular disease (CVD) is the leading cause of death worldwide – more people die from CVD every year than from any other disease. In 2030, about 23.6 million people will die from CVD, mainly from heart disease and stroke. These diseases are projected to remain the main single cause of death"[2].

This shows that predicting cardiovascular disease has become extremely important. In the field of disease prediction: S. M. K. Chaitanya et al. [3] proposed the use of artificial neural networks and gravity search algorithms to detect chronic kidney disease. Maryam Tayefi et al. [4] used a decision tree algorithm to create a predictive model of coronary heart disease to identify risk factors associated with coronary heart disease. Ms S. Kalaiarasi et al. [5] proposed using a mobile phone camera to collect facial data from the human body and create an application that can help diagnose and predict skin diseases using the device's image processing and machine learning functions. Pronevska and Claudia [6] use a random forest algorithm to classify biomedical signals, which increases the reliability of medical diagnostics. Patil et al. [7] presented a K-Means cluster algorithm for retrieving data suitable for heart attack from a data warehouse. Resul et al. [8] proposed forecasting using a set of neural networks that combine existing methods to create new models for predicting diseases.

At the Department of Intelligent Systems in Control and Automation of MTUCI, a large scientific and methodological work is being carried out on the use of modern methods of data mining in the educational process within different disciplines, which allows to form the competencies of bachelors and undergraduates in accordance with the challenges of Industria4.0. At the same time, classical approaches are combined and replenished with new methods. Big Data and databases [9], [10], research methods and models for predicting diseases [11], social adaptation of people with disabilities using computer vision [12].

In this paper, we compared several models for predicting cardiovascular disease based on decision trees, random forests, k-nearest neighbors, naive Bayesian classifiers, and artificial neural networks, and evaluated their predictive effects.

I. Data set

The biomedical dataset used in this study was created by Svetlana Ulyanova, Department of Data Science, Ryerson University, Toronto, Canada [13]:

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

The dataset contains 70,000 data records, including 11 characteristics such as age (age), gender (gender), height (height, cm), weight (weight, kg), systolic blood pressure (ap_hi), diastolic blood pressure (ap_io) , cholesterol (cholesterol, with values of 1: normal; 2: higher than normal; 3: much higher than normal), glucose (gluc, with values of 1: normal; 2: higher than normal; 3: much higher than normal), smoking (smoke), alcohol consumption (alco), and amount of exercise (active). two target outcome classes (cardio), 0: healthy and 1: suffering from cardiovascular disease.

Table 1 shows 21 records from the dataset [13].

Figure 2 shows the correlation between attributes. From the correlation map, it can be seen that cholesterol, blood pressure (ap_hi and ap_low both) and age are closely associated with cardiovascular disease. Glucogen and cholesterol are also closely related. Patients with cardiovascular disease have higher levels of cholesterol and blood sugar, and generally less exercise.

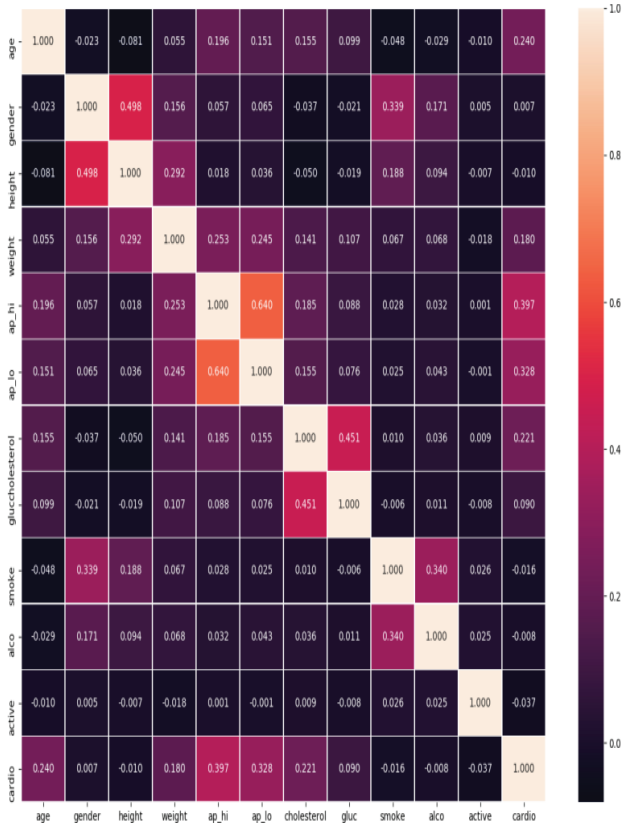

Figure 2. The relationship between every two properties

Table 1

Partial initial data from the set [13]

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62	110	80	1	1	0	0	1	0
1	20228	1	156	85	140	90	3	1	0	0	1	1
2	18857	1	165	64	130	70	3	1	0	0	0	1
3	17623	2	169	82	150	100	1	1	0	0	1	1
4	17474	1	156	56	100	60	1	1	0	0	0	0
8	21914	1	151	67	120	80	2	2	0	0	0	0
9	22113	1	157	93	130	80	3	1	0	0	1	0
12	22584	2	178	95	130	90	3	3	0	0	1	1
13	17668	1	158	71	110	70	1	1	0	0	1	0
14	19834	1	164	68	110	60	1	1	0	0	0	0
15	22530	1	169	80	120	80	1	1	0	0	1	0
16	18815	2	173	60	120	80	1	1	0	0	1	0
18	14791	2	165	60	120	80	1	1	0	0	0	0
21	19809	1	158	78	110	70	1	1	0	0	1	0
23	14532	2	181	95	130	90	1	1	1	1	1	0
24	16782	2	172	112	120	80	1	1	0	0	0	1
25	21296	1	170	75	130	70	1	1	0	0	0	0
27	16747	1	158	52	110	70	1	3	0	0	1	0
28	17482	1	154	68	100	70	1	1	0	0	0	0
29	21755	2	162	56	120	70	1	1	1	0	1	0
30	19778	2	163	83	120	80	1	1	0	0	1	0

II. Data preprocessing

Before performing data operations, you must cleanse and preprocess the data. Detecting and processing outliers can improve the estimation of forecast accuracy.

To compare the properties (age), (height), (weight), (ap_hi), (ap_lo) on the same scale, they must first be standardized. And let's use min-max normalization.

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X is the data, Xmin is minimum value of data sample, Xmax is the maximum value of data sample.

We randomly select 80% of data records in dataset as the training set for training model, and remaining 20% of data records as test case for testing model.

III. Designing and comparing multiple predictive models

In this work, we mainly use the Scikit-learn library [14] to complete model building based on decision trees, random forests, k-nearest neighbors, naive Bayesian classifiers, and artificial neural networks. Scikit-learn provides a variety of algorithms for Supervised Learning and Unsupervised Learning through an interface to the Python programming language.

• *Decision tre*

The decision tree adopts a greedy top-down algorithm that classifies the samples by choosing the best classification attribute at each node, and then continues the process until the tree can accurately classify the training samples or all properties have been used.

Listing 1. A snippet of program code

```
from sklearn.tree import DecisionTreeClassifier
dec = DecisionTreeClassifier()
dec.fit(x_train, y_train)       //training
scores["Decision tree"] = dec.score(x_test, y_test)
```

• *Random for st*

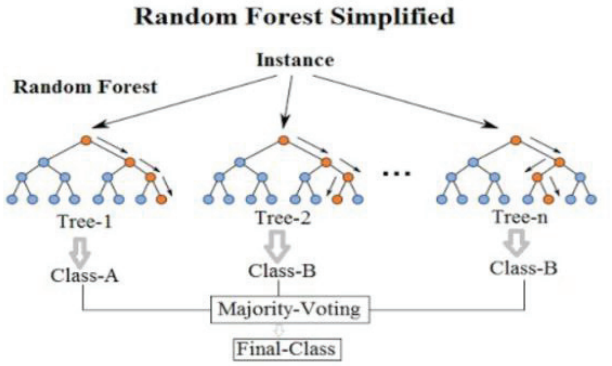

Figure 3. Simplified random forest

The random forest belongs to the bagging algorithm in Ensemble Learning. It is a technique for distinguishing and classifying data using multiple decision trees. Its main unit is a decision tree. He can assess the importance of each variable in categorical data, as well as assess the role of each variable in categorization.

Listing 2. A snippet of program code

```
from sklearn.ensemble import
RandomForestClassifier
ran = RandomForestClassifier(n_estimators=100)
ran .fit(x_train, y_train)     //training
scores["Random forest"] = ran .score(x_test, y_test)
```

• *K-nearest neighbors method*

The k-nearest neighbors method is a nonparametric approach in which the response of a data point is determined by the nature of its k neighbors from the training set. It can be used in both classification settings and regression.

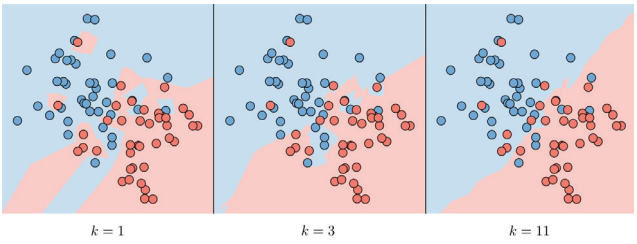

Figure 4. Method of k-nearest neighbors

Listing 3. A snippet of program code

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=100)
knn .fit(x_train, y_train)     // training
scores["KNN"] = knn .score(x_test, y_test)
```

• *Naive Bayesian Classifier*

Assumption – The Naive Bayesian model assumes that all characteristics of each data point are independent:

$$P(x \mid y) = P(x_1, x_2, ... \mid y) = P(x_1 \mid y)... = \prod_{i=1}^{n} P(x_i \mid y) \quad (2)$$

Solutions – The maximum log likelihood gives the following solutions with $k \in \{0, 1\}$, $l \in [1, L]$

$$P(x \mid y) = P(y = k) = \frac{1}{m} \times \#\{j \mid y^{(j)} = k\} \quad (3)$$

and

$$P(x \mid y) = P(x_i = l \mid y = k) = \frac{\#\{j \mid y^{(j)} = k \text{ and } x_i^{(j)} = l\}}{\#\{j \mid y^{(j)} = k\}} \quad (4)$$

Listing 4. A snippet of program code

```
from sklearn.naive_bayes import GaussianNB
naive = GaussianNB()
naive.fit(x_train, y_train)     //training
scores["Naive bayes"] = naive.score(x_test, y_test)
```

• Artificial eural network

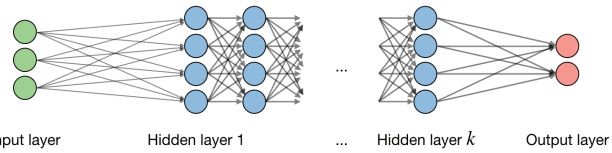

Figure 5. Simplified artificial neural network

Neural networks are a class of models built with layers. Marking the i-th layer of the network and the j-th hidden unit of the level, we get:

$$z_j^{|i|} = w_j^{|i|^T} x + b_j^{|i|} \quad (5)$$

where w-weight, b-bias, z-output respectively.

Examples of the use of neural network modeling in medical problems [16], [17].

Cross entropy loss. In the context of neural networks, cross-entropy loss $L(z, y)$ is usually used, which is defined as follows:

$$L(z, y) = -[y \log(z) + (1 - y) \log(1 - z)] \quad (6)$$

Learning rate – the rate of learning, often referred to α, and sometimes η, indicates how fast the weights are updated. This can be corrected or modified adaptively.

We first use the training set separately to train models based on Decision tree, Random forest, k-nearest neighbors (KNN) method, naive bayes classifiers (Naive bayes), and then use the test set to validate these models. The test results are shown in Fig. 6:

```
                Accuracy Score
Random forest         0.715953
KNN                   0.713054
Naive bayes           0.630862
Decision tree         0.630209
```

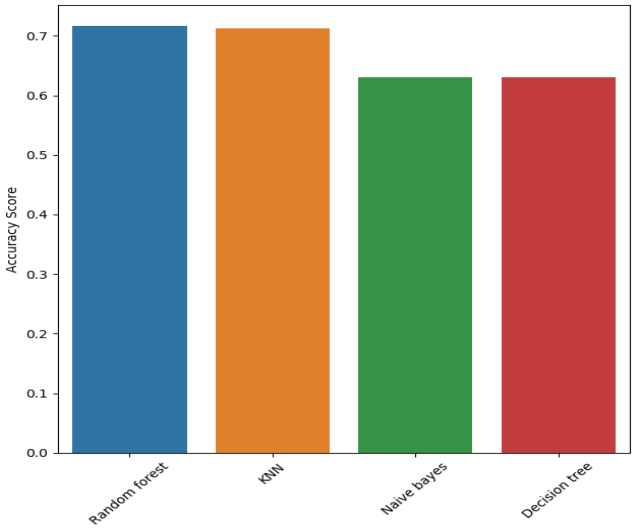

Figure 6. Accuracy test results for multiple models

From the above comparison, it can be concluded that among these models, the random forest model has the highest test accuracy.

Listing 5. A snippet of the program code

```
from sklearn.model_selection import GridSearchCV
grid = {"n_estimators" : np.arange(10,150,10)}
ran_cv = GridSearchCV(ran, grid, cv=3)
ran_cv.fit(x_train,y_train)
print("Tuned hyperparameter n_estimators:
{}".format(ran_cv.best_params_))
print("Best score: {}".format(ran_cv.best_score_))
```

Since we used the default parameters in the above model, next we need to determine the optimal value of random forest model parameter using a grid search algorithm to get the best prediction accuracy.

```
Tuned hyperparameter n_estimators: {'n_estimators': 130}
Best score: 0.7159967265417674
```

Next, we evaluate the effect of predicting an artificial neural network model on this dataset. First, we set the serialization model, set the number of neural network layers, select the activation functions sigmod (Logistic), ReLU (Linear Rectifier) and set the learning rate to 0.002.

Listing 6. A snippet of program code

```
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(6, input_dim=12,
activation='relu'))
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))
optimizer = RMSprop(learning_rate=0.002)
model.compile(loss='binary_crossentropy',
metrics=['accuracy'], optimizer=optimizer)
model.fit(x=x_train, y=y_train.values,
        batch_size=1024,epochs=1500,
        verbose=0,validation_data=(x_test,y_test.values),
        callbacks=[learning_rate_reduction,es],
        shuffle=True)
```

```
13797/13797 - 0s - loss: 0.5678 - acc: 0.7226
```

The accuracy of testing the trained model reaches 0.7226.Figure 7 shows that the accuracy increases with the number of training iterations. When the number of iterations reaches 1200, the accuracy does not increase significantly. Figure 8 shows that the loss function decreases with the number of training iterations. When the number of iterations reaches 1200, the loss function decreases slightly.

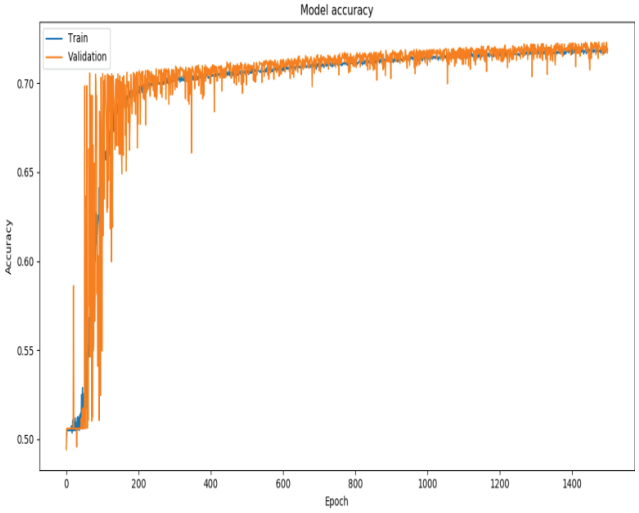

Figure 7. Accuracy changes with the number of training iterations

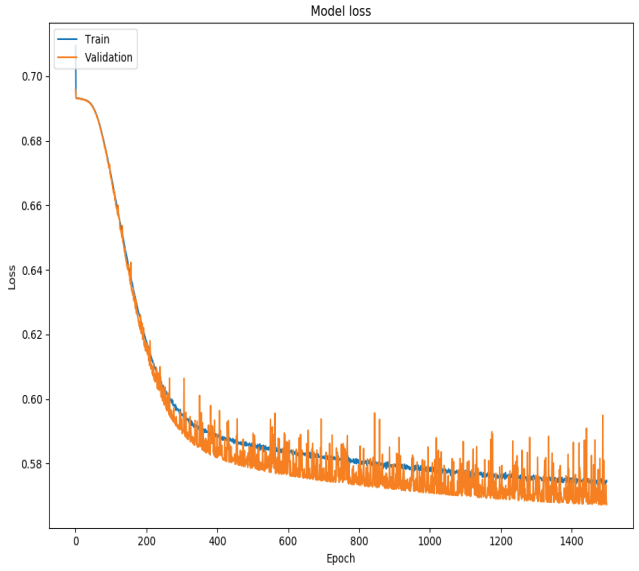

Figure 8. The loss function changes with the number of training iterations

IV. Test results

Table 2

Test result

Model types	Точность
Artificial neural network	0.7226
Random forest	0.7160
K-nearest neighbors method	0.7130
Naive Bayesian Classifier	0.6308
Decision tree	0.6302

Table 2 shows the test results of several models, of which models based on the artificial neural network, random forest, and k-nearest neighbors method have good prediction accuracy, of which artificial neural networks are slightly better than the other two.

Conclusion

This article describes the construction and comparison of several different prediction models applied to a dataset of cardiovascular disease. The forecasting accuracy of the model based on the artificial neural network reached 72%, which was slightly better than in other models. The authors believe that improving the quality of results is advisable to complicate the structure of neural network, improve the training set, and use methods to speed up computations.

References

1. Mortality statistics in Moscow, Moscow City Ritual Service, April 29, 2019 [Electronic resource] https://ritual.ru/poleznaya-informacia/articles/statistika-smertnosti-v-moskve/

2. About cardiovascular diseases, World Health Organization, [Electronic resource] https://www.who.int/cardiovascular_diseases/about_cvd/ru/

3. S. M. K. Chaitanya, P. Rajesh Kumar, Innovations in Electronics and Communication Engineering, January 2019, pp. 441-448.

4. Maryam Tayefi, Mohammad Tajfard, Sara Saffar, et al. Hs-CRP is strongly associated with coronary heart disease(CHD): A data mining approach using decision tree algorithm. *Computer Methods and Programs in Biomedicine*, 2017, 141(4), pp. 105-109.

5. Mrs. S. Kalaiarasi, Harsh Kumar, Sourav Patra,Dermatological Disease Detection using Image Processing and Neural Networks. *International Journal of Computer Science and Mobile Applications*. Vol. 6. Issue. 4, April 2018, pp. 109-118.

6. Proniewska, Klaudia. Data mining with Random Forests as a methodology for biomedical signal classification. Bio-Algorithms and Med-Systems, 2016, 12(2), pp. 89-92.

7. Patil S.B., Kumaraswamy Y.S. Intelligent and effective heart attack prediction system using data mining and artificial neural network[J]. *European Journal of Scientific Research*, 2009, 31(4), pp. 642-656.

8. Resul D., Turkoglu I., Sengur A. Effective diagnosis of heart disease through neural networks ensembles. *International Journal of Expert Systems with Applications*, 2009(36), pp. 7675-7680.

9. Voronova L.I. Teaching aid for the preparation and execution of course projects in the discipline of database technology. Moscow, 2016.

10. Voronov V.I., Voronova L.I., Usachev V.A. Development of a lab workshop for big data processing using hadoop. Moscow, 2018. 49 p.

11. Artemov M.D., Voronova L.I., Voronov V.I., Goncharenko A.A., Yezhov A.A. A software package for sign language recognition based on structural and parametric adaptation of a convolutional neural network. Certificate of registration of the computer program RU 2018666854, 21.12.2018. Application No. 2018664380 dated 13.12.2018.

12. Voronov V., Strelnikov V., Voronova L., Trunov A., Vovik A. Faces 2d-recognition and identification using the hog descriptors method. *Conference of Open Innovations Association, FRUCT*. 2019. No. 24. P. 783-789.

13. Cardiovascular Disease Dataset, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

14. Scikit-learn, Machine Learning in Python – [Electronic resource] https://scikit-learn.org/stable/

15. CS 229 - Machine Learning – [Electronic resource] https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning#other.

16. Wang Yue, Voronova L.I., Classification of the type of disease of the spine using a neural network. *Information Society Technologies. Materials of the XIII International Industrial Scientific and Technical Conference*. 2019. P. 407-410.

17. Kesyan G.R., Voronova L.I., Trunov A.S. Predicting the presence of diabetes mellitus using neural networks. *Information Society Technologies. Materials of the XIII International Industrial Scientific and Technical Conference*. 2019. P. 438-440.