Intelligent Discriminant Diagnosis of Heart Disease Cases

Qi Wang, Guici Chen

Wuhan University of Science and Technology, Wuhan, Hubei; 430065, China

Abstract

The application of artificial intelligence in the medical field has greatly alleviated the contradiction between people's growing demand for medical resources and the actual shortage of medical resources. In this paper, the combination of Fisher dimension reduction and Hidden Markov Model (HMM) is applied to the intelligent diagnosis of heart disease cases. The index sequence of heart disease cases was simplified by Fisher dimension reduction. The HMM of heart disease and non-heart disease is established by Baum-Welch algorithm. The matching score between the observation sequence and the two HMM is calculated by the Forward-Backward algorithm. The experimental results show that the diagnosis of heart disease cases by matching scores is reliable.

Keywords

Fisher dimension reduction, HMM, heart disease diagnosis, classification

1.Introduction

In recent years, the growing demand for medical resources and the shortage of medical resources have promoted the continuous infiltration of artificial intelligence into the medical field. Artificial intelligence has been successfully applied in intelligent diagnosis, intelligent drug research and development, intelligent image recognition, intelligent health management, medical robots, and other fields. The exact diagnosis and recognition of diseases is a significant task in medicine, as well as a major development field of artificial intelligence in medical applications. Domestic and foreign scholars do some research in the field of intelligent diagnosis. Yala et al. [1] developed a machine learning model to extract relevant tumor features from breast pathology reports. The accuracy of artificial intelligence in diagnosing pathological sections is 90%. Zhang et al. [2] studied a multimodal deep learning model to classify the abnormal behaviors of children in the collected videos, and combined with the information collected by other system modules. Fujioka et al. [3] used CNN to classify ultrasonic shear wave elastography of breast masses. They applied a series of CNN models, and the results showed that the best CNN model was DenseNet169, with sensitivity, specificity, and AUC of 85.7%, 78.9%, and 0.870.

This paper studies the differential diagnosis of heart cases. Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. How to diagnose heart disease quickly and effectively has always been one of the key issues in the field of life science. We apply the Fisher reduced dimension and HMM to the discrimination diagnosis of heart disease cases. The experimental results show that it is reliable.

2.Data Processing

The data comes from Kaggle's Heart Failure Prediction dataset, which contains 12 groups of data. First, we clean the data and delete the abnormal data whose RestingBP is 0 and Cholesterol is 0. There are 745 groups of data after cleaning. There were 335 groups of patients with heart disease,

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China

accounting for 47.65%, and 390 groups of patients without heart disease, accounting for 52.35%. Then, the data is deconstructed and mapped. The processed data is shown in Table 1.

Table 1
Data indicators

Age	1 (age>40), 0 (age<=40)
Sex	1 (Female), 0 (Male)
TA	1 (ChestPainType=Typical Angina), 0 (Otherwise)
ATA	1 (ChestPainType=Atypical Angina), 0 (Otherwise)
NAP	1 (ChestPainType=Non-Anginal Pain), 0 (Otherwise)
ASY	1 (ChestPainType=Asymptomatic), 0 (Otherwise)
RestingBP	1 (RestingBP>140), 0 (RestingBP<=140)
Cholesterol High	1 (Cholesterol>230), 0 (Otherwise)
Cholesterol Low	1 (Cholesterol<110), 0 (Otherwise)
FastingBS	1 (FastingBS > 120) , 0 (Otherwise)
Normal	1 (RestingECG = Normal), 0 (Otherwise)
ST	1 (RestingECG = ST), 0 (Otherwise)
LVH	1 (RestingECG = LVH), 0 (Otherwise)
MaxHR	1 (MaxHR is within the range of its average +/-2*variance), 0 (Otherwise)
ExerciseAngina	1 (ExerciseAngina = Yes), 0 (ExerciseAngina = No)
Oldpeak	1 (Oldpeak > 0.5), 0 (Otherwise)
ST_Slope-Up	1 (ST_Slope = Up), 0 (Otherwise)
ST_Slope-Flat	1 (ST_Slope = Flat), 0 (Otherwise)
ST_Slope-Down	1 (ST_Slope = Dowm), 0 (Otherwise)
HeartDisease	output class [1: heart disease, 0: Normal]

3. Combining HMM and Fisher's diagnostic model

3.1. Fisher dimension reduction

There are 19 groups of data indicators after deconstruction and mapping. To reduce the time complexity and space complexity, we use the Fisher dimension reduction method to extract significant features and simplify the prediction indicators. Fisher's idea of dimensionality reduction is to project high-dimensional pattern samples into the optimal discriminant vector space ω to extract classificatio

-n information and compress the dimension of feature space. After projection, the maximum inter-class distance $\omega^T S_b \omega$ and the minimum intra-class distance $\omega^T S_\omega \omega$ of pattern samples in the new subspace are guaranteed.

$$S_b = (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T,\ \mu_i = \frac{1}{N_i}\sum_{x \in X_i} x\ , \tag{1}$$

$$S_\omega = \sum_{x \in X_0}(x - \mu_0)(x - \mu_0)^T + \sum_{x \in X_1}(x - \mu_1)(x - \mu_1)^T, \tag{2}$$

The best projection direction ω is the direction that makes $\omega^T S_b \omega / \omega^T S_\omega \omega$ maximum. Here, the Lagrange multiplier method uses to solve the best projection direction and obtains the discriminant function $y = \omega^T x$.

3.2. Baum-Welch algorithm for solving HMM parameters

To establish HMM with heart disease and non-heart disease, we first need to obtain its parameter $\lambda = (A, B, \Pi)$. A is the state transition matrix composed of a_{ij}; B is the observation-generated probability matrix composed of $b_j(k)$, and Π is the initial state probability distribution. First, get the

numerical sequence of ST_Slope UP, ASY, ExerciseAngia, Sex, and RestingBP input into the HMM and the randomly given parameter $\pi_i, a_{ij}, b_j(k)$. Then calculate $\xi_t(i,j), \gamma_t(i)$ to update the model parameters. $\xi_t(i,j)$ describes the probability that t is in state q_i and t+1 is in state q_j at time t, and $\gamma_t(i)$ describes the probability that t is in state q at time t, which are recorded as:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ijt}b_j(o_{t+1})\beta_{t+1}(j)}{p(O|\lambda)}, \tag{3}$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(i)\beta_t(i)}, \tag{4}$$

Then update the model parameters,

$$\pi_i = \frac{\sum_{d=1}^{D}\gamma_1^{(d)}(i)}{D}, \tag{5}$$

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T-1}\xi_t^{(d)}(i,j)}{\sum_{d=1}^{D}\sum_{t=1}^{T-1}\gamma_t^{(d)}(i)}, \tag{6}$$

$$b_j(k) = \frac{\sum_{d=1}^{D}\sum_{t=1,o_t^{(d)}=v_k}^{T-1}\gamma_t^{(d)}(j)}{\sum_{d=1}^{D}\sum_{t=1}^{T-1}\gamma_t^{(d)}(j)}, \tag{7}$$

If the value has converged, the algorithm ends, otherwise, continues to iterate. The parameter $\lambda_1 = (A_1, B_1, \Pi_1)$ of the HMM of heart disease and the parameter $\lambda_0 = (A_0, B_0, \Pi_0)$of the HMM of non-heart disease are trained by the Baum-Welch algorithm.

3.3. Forward-Backward algorithm to distinguish the category

After obtaining the HMM parameters, the Forward-Backward algorithm uses to calculate the matching score $p(O|\lambda_i)$ (i=0,1) between the observation index sequence O and the two models. Compare the score size, and determine the category of the observation index sequence.The first step of the Forward algorithm is to calculate the forward probability $\alpha_1(i)$ of each state at time 1, the second step is to calculate the forward probability $\alpha_{t+1}(i)$ at times 2, 3,..., T, and finally calculate $p(O|\lambda)$.

$$\alpha_1(i) = \pi_i b_i(o_1), \tag{8}$$

$$\alpha_{t+1}(i) = (\sum_{j=1}^{N}\alpha_t(j)a_{ji})b_i(o_{t+1}), \tag{9}$$

$$p(O|\lambda) = \sum_{i=1}^{N}\alpha_T(i), \tag{10}$$

The Backward algorithm is the reverse process of the forward algorithm, so I won't repeat it here.

4.Model experiment test

Through the Fisher discriminant function, we extracted 5 significant indicators from the original 19 groups of indicator data. At the same time, the ranking of the importance of the five indicators is ST_ Slope UP>ASY > ExerciseAngina> Sex > RestingBP.

To unify the input length of the index series, we add five opposite indexes to the observation index series of the HMM. Therefore, the observation index series of the HMM includes RestingBP high, RestingBP normal, Male, Female, ChestPainType yes, ChestPainType as, ExerciseAngia yes, ExerciseAngia no, ST_ Slope-up, ST_ Ten indicators of Slope Flat Down. We select 80% of the observation index sequence data of all data sets as the training set to input the HMM. The model parameters are obtained by the Baum-Welch algorithm.

Four groups of observation index sequences are selected and the matching scores calculated by the Forward-Backward algorithm are shown in Figure 1. Through Figure 1, we can see that the matching scores of the same index sequence under different models have certain differences, which shows that the validity of observation sequence data can be judged by matching scores.

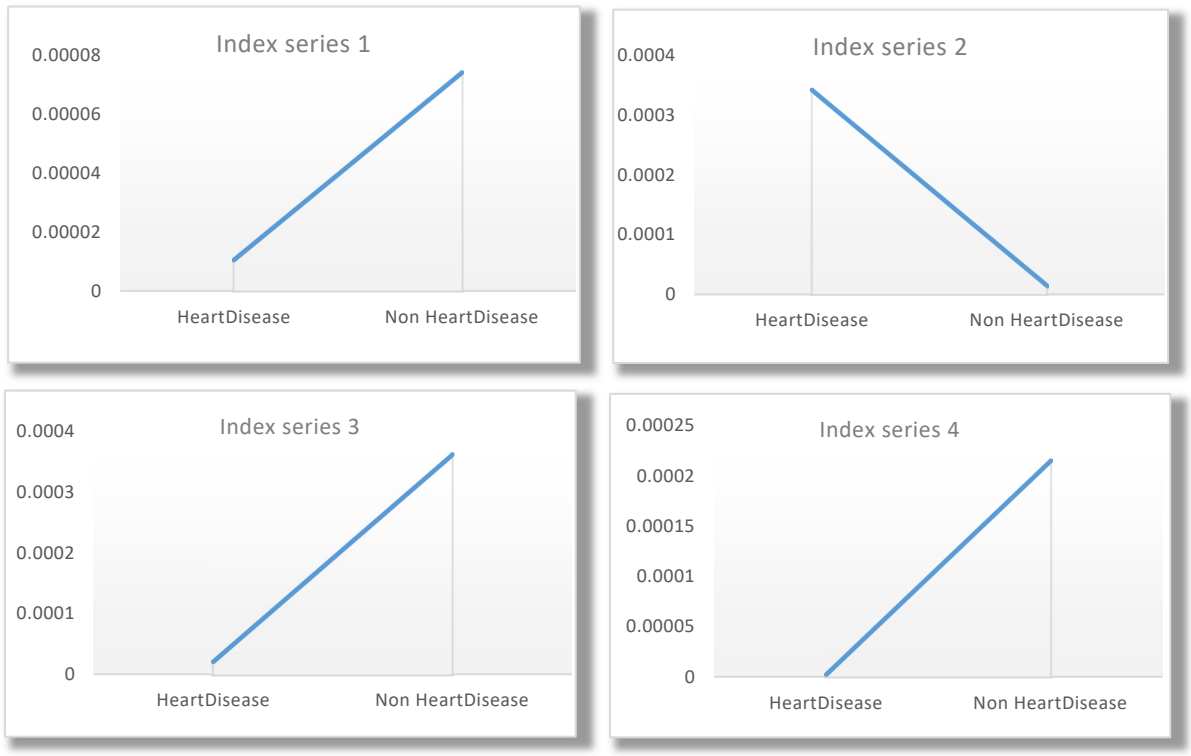

Figure 1: Matching score of indicator sequence and HMM

We take all the data as the validation set, and diagnosed the heart disease cases through the matching score between the observation index sequence and HMM, and the overall accuracy was 85.9%. To further verify the reliability of the model. We used ANN and Decision Tree to diagnose heart disease cases after data processing. The overall accuracy of ANN and Decision Tree was 84.3% and 82.5%. The detailed classification results of the three models are shown in Table 2 (the model proposed in this paper is abbreviated as F-H).

Table 2
Classification Results

F-H		HeartDisease	Non HeartDisease
	HeartDisease	92.39%	7.61%
	Non HeartDisease	20.00%	80.00%
ANN		HeartDisease	Non HeartDisease
	HeartDisease	85.47%	14.53%
	Non HeartDisease	16.54%	83.46%
Decision Tree		HeartDisease	Non HeartDisease
	HeartDisease	87.18%	12.82%
	Non HeartDisease	21.71%	78.29%

From Table 2, we can see that among the three models, 92.39% of the F-H model, 87.18% of the decision tree model, and 85.47% of the ANN model have the highest diagnostic accuracy. The highest diagnostic accuracy of non-heart disease was 83.46% in the ANN model, 80% in the F-H model, and 78.29% in the Decision Tree model.

5. Result analysis and summary

We deconstruct and map the original 11 groups of index data, expand the data to 19 groups, and then extract the important features that determine heart disease from the data indicators through Fisher

dimension reduction ST_ Slope UP, ChestPainType ASY, ExerciseAngia, Sex, RestingBP, and their importance ranking. Finally, the overall accuracy of the classification identified by HMM is 85.9%. Next, we will make some simple analysis of the results. We counted the number of patients with heart disease and non heart disease under several observation indicators, as shown in Table 3.

From Table 3, we can see that under the same observation index sequence, the number of patients with heart disease is equivalent to the number of non-heart patients. It leads to that even if the model parameters can fully fit the characteristics of the training data set, the accuracy of discrimination will not reach a very high level. It also shows that the original indicators cannot completely and accurately depict the characteristics of the heart disease population.

Table 3

Statistics of the number of people with heart disease and non heart disease under the same index

	HeartDisease	Non HeartDisease
ST_Slope-Flat-Down，ChestPainType yes, ExerciseAngina no, Male, RestingBP normal	25	18
ST_Slope-Flat-Down，ChestPainType yes, ExerciseAngina yes, Male, RestingBP normal	20	7
ST_Slope-Up，ChestPainType no, ExerciseAngina no, Male, RestingBP normal	10	36
ST_Slope-Flat-Down，ChestPainType yes, ExerciseAngina no, Male, RestingBP high	9	5
ST_Slope-Flat-Down，ChestPainType no, ExerciseAngina yes, Female, RestingBP normal	12	7

6.Acknowledgements

Thanks to the teachers, classmates, friends, and family who contributed to this article.

7.References

[1] Wang Tingting, Xing Dengxiang. Research on the progress of artificial intelligence in medical applications, J. Trauma and Critical Care Medicine, (2021).doi:10.16048/j.issn.2095-5561.

[2] Gong Gao, Huang Wenhua, Cao Shi, Chen Chaomin,. Research progress in the application of artificial intelligence in medicine J. Chinese Journal of Medical Physics, (2021): 1044-1047

[3] D.A.Qiu Shuang. Application of artificial intelligence diagnosis system based on DE Light framework in breast ultrasound, Master's thesis, University of Electronic Science and Technology of China, Chengdu, China,2022.

[4] Zong Changfu, Yang Xiao, Wang Chang, Zhang Guangcai. Driver's driving intention identification and behavior prediction during vehicle steering J. Journal of Jilin University (Engineering Edition), (2009) 27-32. doi: 10.13229/j.cnki.jdxbgxb2009.s1.023

[5] Sergios Theodoridis, Machine Learning (Second Edition),Chapter 7 - Classification: a Tour of the Classics, Editor(s): Sergios TheodOrid, Academic Press, 2020, Pages 301-350, doi:10.1016/B978-0-12-818803-3.00016-7.

[6] Shoba Ranganathan, Encyclopedia of Bioinformatics and Computat-ional Biology,Hidden Markov Models, Monica Franzese, Antonella Iuliano, Academic Press,2019,Pages 753-762, doi:10.1016/B978-0-12-809633-8.20488-3.