BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Integrating functional data analysis with case-based reasoning for hypertension prognosis and diagnosis based on real-world electronic health records

Ping Qi[1*], Fucheng Wang[1], Yong Huang[2] and Xiaoling Yang[3]

Abstract

Background: Hypertension is the fifth chronic disease causing death worldwide. The early prognosis and diagnosis are critical in the hypertension care process. Inspired by human philosophy, CBR is an empirical knowledge reasoning method for early detection and intervention of hypertension by only reusing electronic health records. However, the traditional similarity calculation method often ignores the internal characteristics and potential information of medical examination data.

Methods: In this paper, we first calculate the weights of input attributes by a random forest algorithm. Then, the risk value of hypertension from each medical examination can be evaluated according to the input data and the attribute weights. By fitting the risk values into a risk curve of hypertension, we calculate the similarity between different community residents, and obtain the most similar case according to the similarity. Finally, the diagnosis and treatment protocol of the new case can be given.

Results: The experiment data comes from the medical examination of Tianqiao Community (Tongling City, Anhui Province, China) from 2012 to 2021. It contains 4143 community residents and 43,676 medical examination records. We first discuss the effect of the influence factor and the decay factor on similarity calculation. Then we evaluate the performance of the proposed FDA-CBR algorithm against the GRA-CBR algorithm and the CS-CBR algorithm. The experimental results demonstrate that the proposed algorithm is highly efficient and accurate.

Conclusions: The experiment results show that the proposed FDA-CBR algorithm can effectively describe the variation tendency of the risk value and always find the most similar case. The accuracy of FDA-CBR algorithm is higher than GRA-CBR algorithm and CS-CBR algorithm, increasing by 9.94 and 16.41%, respectively.

Keywords: Case-based reasoning, Functional data analysis, Time series, Hypertension

Introduction

According to the China Cardiovascular Disease Report, there are currently 270 million adult hypertensive patients and 290 million cardiovascular disease patients in China [1]. A systematic analysis of data reveals that China is one of the top nine countries with the most severe rises in both male and female morbidity rates of hypertension [2]. Hypertension has posed a severe threat

*Correspondence: qiping929@tlu.edu.cn

[1] Department of Mathematics and Computer Science, Tongling University, Tongling 244061, China
Full list of author information is available at the end of the article

to public health, and it creates a lot of chain problems. On the one side, hypertensive patients are more likely to develop diabetes, heart failure, angina pectoris, myocardial infarction, and other adverse health outcomes. On the other side, hypertension costs the country about 366 billion in 2020, and the annual cost is rising substantially [3]. However, due to the developing symptoms of hypertension being mostly hidden, most people do not know they have pre-hypertension or already have hypertension. Therefore, the early prognosis and diagnosis is a critical step in the hypertension care process.

Using of data from electronic health record (EHR) has shown great promise for the early detection of chronic disease [4–6]. Various methods have been proposed to provide disease prediction and clinical decision-making aid based on retrospective electronic health records data, such as regression model [7], decision-making tree [8], recurrent neural network [9], and case-based reasoning (CBR) [10]. Inspired by human philosophy, CBR is an empirical knowledge reasoning method to find the recorded case most relevant to the target case, which avoids the training process of machine learning algorithms. The recorded case that has occurred in the past is referred to as the source case, and the source cases are used to guide the solution of the target case. Because of its better learning capability and interpretability than the rule-based and model-based reasoning algorithm, CBR has been widely used in medical diagnosis [11].

As mentioned above, the core idea of CBR is to handle similar problems. In the medical field, the diagnosis and treatment protocol of the new case can be given according to the most similar source case in the case base. However, the traditional CBR algorithms tend to focus on cross-sectional study in the medical field and ignore the impact of time. As shown in Fig. 1, the weights of Zhang and Wang are the same in September. But we do not consider that the importance of A and B remain the same in October. The main reason is that the various tendencies of two people are significantly different. Obviously, the similarity between source case and target case is also affected by time. The more recent the history information is, the more impact the time factor has.

In addition, although the community health service stations organize free medical examinations every half year, some community residents do not attend continuously. Therefore, for real-world electronic health records, the time and frequency of medical examination for different people may be different as well. In this case, how to evaluate the similarity between two people is of great importance.

Because of the aforementioned limitations mentioned above, we first analyze and evaluate the risk value of hypertension based on the medical examination data of community resident. As each community resident has multiple medical examination records, these data are fitted into a risk curve of hypertension for each resident based on functional data analysis (FDA) technique. Then, we can calculate the similarity between two different curves, which can be defined as the similarity between the recorded case and the target case. Finally, a novel FDA-based CBR model named FDA-CBR is proposed in this paper.

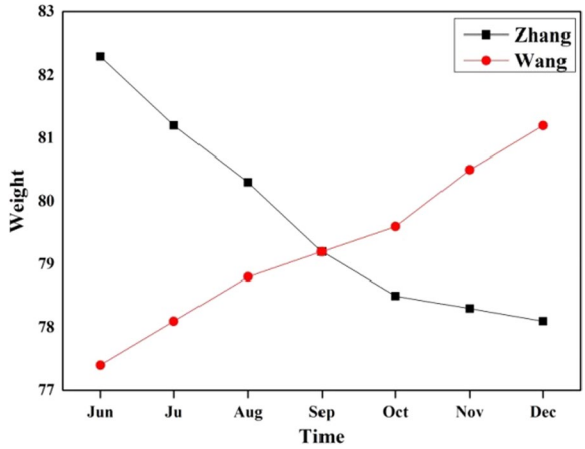

Fig. 1 Weight changes of Zhang and Wang during half a year

The remainder of this paper is listed as follows: "Related work" section presents the related work. "Material and method" section proposes the material and method. "Discussion" section demonstrates the experimental results. Finally, the conclusions and improvement directions are presented in section "Conclusions".

Related work

Since Roger Schank first came up with CBR in 1982, CBR has become one of the prominent reasoning mechanisms, and has been widely used in medical diagnosis, industrial production and engineering planning. Using the previous experience or knowledge, CBR can realize the reuse of essential knowledge and effectively extract the complex rules. The processing of the CBR system contains case reuse, case retrieve, case revise, and case retain.

As is well-known, the similarity calculation between the target case and the source cases is the critical step for case retrieval. Therefore, similarity calculation has become a research hotspot recently. Euclidean distance function and Mahalanobis distance function are the most commonly used distance functions of similarity in many studies [12, 13].

Some studies integrate Cosine similarity or Jaccard similarity into the CBR system, which makes the applicability of similarity measure of the traditional CBR widely

extended. Zhang et al. [14] calculate the angle between two eigenvectors by using the cosine theorem and Euclidean distance, which is defined as the Cosine similarity. In the meantime, the traditional single-category attribute is extended to relative entropy model-based multi-attribute. Baharav et al. [15] use Jaccard distance to measure similarity between sample sets. Min-hashes are employed to efficiently estimate these similarities. Chen et al. [16] present an emergency decision model based on grey relational analysis, which can effectively quantify the attribute weights and the similarity for the heterogeneous multi-attributes decision-making problem.

In addition, there are so many risk factors for hypertension, i.e. obesity, smoking, alcohol consumption and waistline. A reasonable weight assignment of attributes has a significance influence on the decision result. Multivariate logistic regression model [17] and cox regression model [18] are the most representative risk prediction model. However, with the growth of data in volume and dimensionality, the ability of data mining algorithms to deal with mass-data becomes more important. Some classification based data mining technique, such as random forest [19] and SVM [20], has performed well for multilabel classification using knowledge-driven features. It also can reduce the complexity of the model by reducing the number of features required to train a machine learning model.

In recent years, with the rapid development of machine learning techniques, some machine learning algorithms have been used to learn the similarity between two record cases. Zhang et al. [21] adopt the earth mover's distance as the similarity between two dense images, which is used for classification. Vij et al. [22] present a machine learning-based approach to find out the similarity between two texts. Unfortunately, although machine learning technique-based algorithms are very useful, these algorithms are not widely used due to lack of samples.

In summary, distance measure function based similarity calculation method can only reflect the relationship in spatial location, but ignores the time series and variation tendency of the record cases. Fortunately, functional data analysis is a statistical analysis technique especially suited for the analysis of curves, which can be used for s table estimates and accurate predictions [23]. To overcome the above shortcomings, functional data analysis is a suitable method to capture the time series similarity of two data series in the system.

Material and method

This study received ethical approval from the Ethics Committee of Tongling Municipal Hospital and Anhui Medical University. The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and research regulations of the country. Considering retrospective nature of the study, Informed consent was waived by the Ethics Committee of Tongling Municipal Hospital.

Material

The data comes from the medical examination of the Tianqiao Community (Tongling City, Anhui Province, China) from 2012 to 2021. It contains 4143 community residents and 43,676 medical examination records. Each record includes more than 100 attributes, such as demographic information, physical examination, physiology, biochemistry, and so on. As shown in Table 1, the quantization assignment method is One-Hot encoding. Because of the community health service station organizes free medical examination in March and September each year, the time of medical examination can be marked with its ordinal number. For example, the medical examination in March 2012 is marked as "1". Similarly, September 2012 and September 2021 can be marked as "2" and "20" separately.

Method

In this subsection, we describe the novel FDA-based CBR model in detail. Firstly, we calculate the weights of input attributes by random forest algorithm. Secondly, for every community resident, the risk value of hypertension from each medical examination can be evaluated according to the input data and the attribute weights. Then, these continuous or not completely continuous risk values are fitted into a curve of risk value by using the medical examination time as the variable. Based on this, we can calculate the similarity between two curves, and the most similar case is extracted according to the similarity. Finally, the diagnosis and treatment protocol of the new case can be given.

The case extract strategy is as follows: when the similarity between the new case and the source case is over 90%, we can directly reuse the diagnosis and treatment protocol of the source case. When the similarity is between 70 and 90%, the diagnosis and treatment protocol of the source case can be regarded as an alternative treatment plan. When the similarity is between 60 and 70%, the extracted case can be used as an auxiliary reference plan. The whole workflow of FDA-CBR is shown in Fig. 2.

Weight assignment based on random forest algorithm

In order to make the weight calculation of attributes more reasonable, a random forest algorithm is employed in this paper. Random forest algorithm combines different decision trees (decision tree, DT). Each

Table 1 Medical examination information

Variable	Quantitative assignment
Hypertension	NO→1; YES→1
Gender	Female→0;Male→1
Age	>65→1; 35~65→2; <35→3
Exercise frequency	Never→1; Everyday→2; Once a week or more→3; Occasionally→4
Dietary habit	Meat diet→1; Vegetarian diet→2; Equilibrium→3
Smoking	Yes→1; Never→2; Quitting→3
Drinking	Everyday→1; Frequently→2; Never→3; Occasionally→4
Heart rhythm	Normal→0; Arrhythmia→1
Central obesity	<90 cm(Male) or <80 cm(Female)→0; >90 cm(Male) or >80 cm(Female)→1
BMI	18.5~24→1; 24~28→2; >28→3; <18.5→4
Diabetes	No→0; Yes→1
Heart rate	60~100→1; >100→2; <60→3
Blood urea	3.2~7.1→1; >7.2→2; <3.2→3
…	…
Total cholesterol	>5.2→1; 3.0~5.2→2; <3.0→3
Triglyceride	<1.7→1; 1.7~5.65→2; ≥5.65→3
Low-density lipoprotein	<4.14→0; ≥4.14→1
High-density lipoprotein	≥1.2→0; <1.2→1

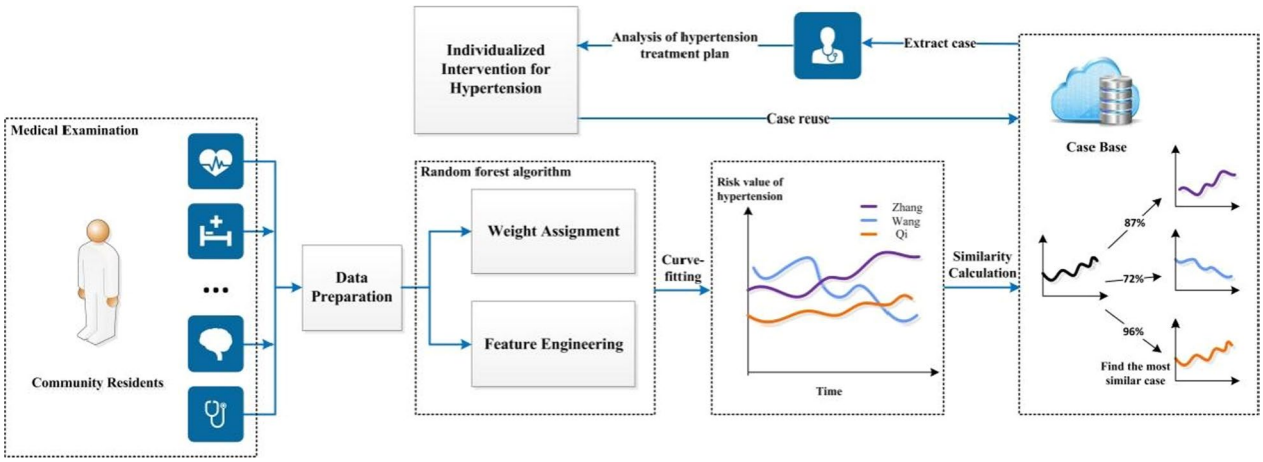

Fig. 2 Whole workflow of FDA-CBR

decision tree depends on the values of independently sampled random vectors. As shown in Fig. 3, the weight assignment is obtained by casting a vote for the most effective class. Assuming that each case is represented by an n-dimensional feature vector $X=\{x_1,x_2,\ldots,x_n\}$, the weight vector of the attributes can be described as follows: $W=\{w_1, w_2,\ldots, w_n\}$. The algorithm flow is listed as follows:

(1) $S=\{S_1,S_2,\ldots,S_k\}$ are sampled randomly from the medical examination data, and the bootstrap sampling method is employed in this process;
(2) Different decision trees are constructed based on S. During the construction phase of each decision tree, when the value of $Gini(t)$ increases, less available information can be gained. Therefore, the total

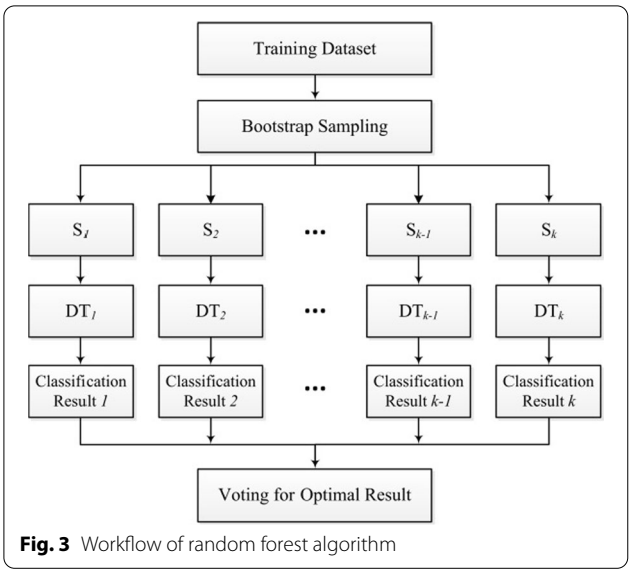

Fig. 3 Workflow of random forest algorithm

Gini value of all derived nodes should be less than that of the parent node. The minimum *Gini* value is used as the best splitting standard of the nodes, calculated by Formula (1).

$$Gini(t) = 1 - \sum_{j=1}^{k} \left[p(j|t)\right]^2 \tag{1}$$

where $p(j|t)$ denotes the probability of risk class j at node t.

(3) Each decision tree votes for the most effective classification, and the vote results determine the optimal weight assignment. Assuming that D_i is the mean *Gini* decrease for i-th variable. w_i is an i-th variable weight, which can be calculated by Formula (2):

$$w_i = \frac{D_i}{\sum_{i=1}^{n} D_i} \tag{2}$$

Curve-fitting and similarity calculation

Based on weight assignment, the risk value of hypertension from the i-th medical examination can be evaluated according to the input data and the attribute weights by Formula (3).

$$risk_i = X \times W = (x_1, x_2, \ldots x_n) \times (w_1, w_2, \ldots w_n) \tag{3}$$

For community residents, each of them may have multiple medical examination results. Therefore, these dynamically changing risk values are fitted into a curve of risk value by using the medical examination time as the variable. Then, we can calculate the similarity between the two curves. The specific calculation steps are as follows: basis function selection, smoothing function, calibration function, and similarity calculation.

(1) Basis function

Basis function fitting is the most common method of FDA. Basis function is a series of independent function, which is defined as $R(t) = \sum_{k=1}^{K} c_k \phi_k(t)$, where $\Phi_k(t)$ ($k=1, 2,\ldots, K$) are k selected basis functions, c_k is the coefficient matrix. In general, the B-spline basis function is more appropriate for aperiodic functional data. Assuming that the time interval [1, 20] (medical examination from March 2012 to September 2021) is divided into several subintervals $[t_{i-1},t_i]$, where t_i is the time of i-th medical examination, $risk_i$ is the risk value of hypertension from i-th medical examination, $1 \le t_0 < t_1 < \ldots < t_N \le 20$. $B_{i,k}(t)$ is defined recursively as the B-spline basis function of order k by Formula (4) and Formula (5).

$$B_{i,0}(t) = \begin{cases} 1, & t_i \le t \le t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k} - t_i} B_{i,k+1}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(t) \tag{5}$$

(2) Smoothing function

According to the basis function, the coefficient vector should be calculated by the least square method. That is to minimize the following formula.

$$\mathrm{SMSSE}(Risk|C) = \sum_{j=1}^{n} \left[risk_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2 \tag{6}$$

where $Risk$ and C are the matrix form of $\{risk_j\}$, $\{c_k\}$.

(3) Calibration function

Unlike point data, the properties of functional data include amplitude and phase. Therefore, the purpose of the calibration function is to move the misaligned variable to the same standard by adjusting the translation parameters. Let the translation parameter be δ_i, $R_i^*(t) = R_i(t+\delta_i)$, δ_i can be calculated by minimizing Formula (7).

$$REGSSE = \sum_{i=1}^{n} \int_{t_1}^{t_2} \left[R_i(t+\delta_i) - \hat{\mu}(t)\right]^2 dt = \sum_{i=1}^{n} \int_{t_1}^{t_2} \left[R_i^*(t+\delta_i) - \hat{\mu}(t)\right]^2 dt \tag{7}$$

where $\hat{\mu}(t)$ is the mean value of all the functional data in $[t_1,t_2]$. This mean value function is updated iteratively until it stabilizes, which makes the translation parameter more rational.

(4) Similarity calculation

The resampling technique is used in this paper to collect data. Firstly, we transform the risk scores into functional data using FDA, and then define the continuity of data exactly by using function properties. In the calculation process, the fitting function is divided into 19 intervals: {(1,2),(2,3),...(19,20)}. We can obtain the continuous function in any one of these intervals. Then the interval similarities are calculated separately and integrated into the global similarity with decay factor, which makes the calculation more accurate.

The interval similarity between two curves is calculated in two parts: actual distance and derived function distance. The actual distance describes the data discrepancy, and the derived function distance describes the discrepancy of inherent characteristics. Let $R_{org}(t)$ and $R_{tgt}(t)$ be the functional descriptions of the original case and target case, respectively. Then the actual distance d_{act} between $R_{org}(t)$ and $R_{tgt}(t)$ on $[t_1,t_2]$ can be calculated by Formula(8).

$$d_{act} = \sqrt[2]{\int_{t_1}^{t_2} \left(R_{org}(t) - R_{tgt}(t)\right)^2 dt} \tag{8}$$

Let $R'_{org}(t)$ and $R'_{tgt}(t)$ be the derived functions of $R_{org}(t)$ and $R_{tgt}(t)$ respectively. Then the derived function distance $d_{der}(R'_{org}(t), R'_{tgt}(t))$ on $[t_1,t_2]$ can be calculated by Formula(9)

$$d_{der} = \sqrt[2]{\int_{t_1}^{t_2} \left(R'_{org}(t) - R'_{tgt}(t)\right)^2 dt} \tag{9}$$

Thus, two kinds of distance between the original case and target case can be aggregated into an integrated similarity $Sim(org, tgt)$ as follows:

$$Sim(org, tgt) = \theta \cdot d_{act} + (1-\theta) \cdot d_{der}, \quad \theta \in (0,1) \tag{10}$$

where θ is the influence factor of actual distance and derived function distance. Briefly, $\theta \in (0, 0.5)$ indicates that we are more concerned with the variation tendency of the risk value of medical examination data.

(5) Decay factor

As discussed above, recent information has more influence on similarity calculation. The decay factor μ is employed to reflect the importance of the historical information, which decreases as time pass on. The similarity with decay factor can be calculated as follows:

$$Sim(n) = \sum_{i=1}^{n} Sim_i \times \mu^{(n-i)} \tag{11}$$

Discussion

Weight assignment

With the aid of the grid searching technique (GridSearchCV), the depth and number of decision trees are set to 5 and 500 separately. The experimental result is shown in Table 2.

As shown in Table 2, the top 10 weighted attributes are age, diabetes, exercise frequency, BMI, total cholesterol, smoking, drinking, central obesity, triglyceride, and blood urea. According to Formula (3), the risk value of hypertension from each medical examination can be evaluated according to the input data and the attribute weights. Table 3 shows a community resident's risk value of hypertension in 10 consecutive medical examinations.

In the real world, this community resident is 52 years old. He does not drink or smoke. However, he seldom does exercise, and his BMI has been increasing since 2017. He was diagnosed with hypertension in 2019. The experimental result shows that the risk value can effectively reflect the hypertension risk of community residents.

Influence factor

In this experiment, we discuss the effect of the influence factor θ in Formula (8). In order to evaluate the performance of the proposed algorithm, we define "correct

Table 2 Weights of input attributes

Variable	Weights of input attributes
Age	0.301
Diabetes	0.152
Exercise frequency	0.112
BMI	0.103
Total cholesterol	0.073
Smoking	0.051
Drinking	0.049
Central obesity	0.041
Triglyceride	0.032
Blood urea	0.031
Serum high lipoprotein cholesterol	0.028
Heart rhythm	0.027
Gender	0.024
Heart rate	0.023
Serum low lipoprotein cholesterol	0.019
Dietary habit	0.017
...	...

Table 3 Risk value of hypertension in 10 continuous medical examinations

Time	2017.3	2017.9	2018.3	2018.9	2019.3	2019.9	2020.3	2020.9	2021.3	2021.9
Risk value	13.6	13.2	13.8	14.5	15.6	15.9	15.4	15.8	15.7	16.2

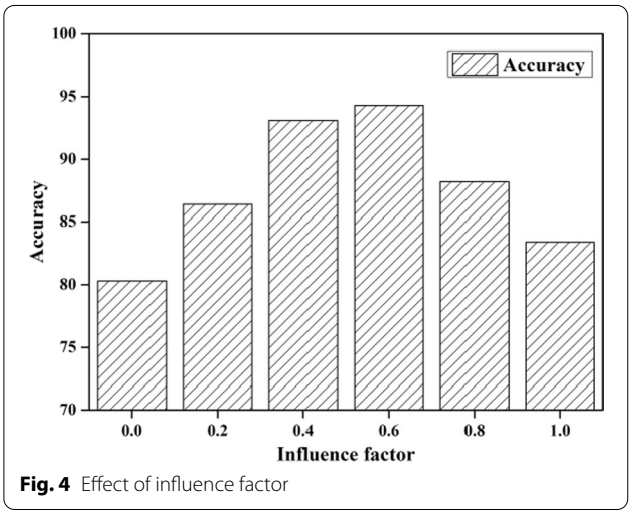

Fig. 4 Effect of influence factor

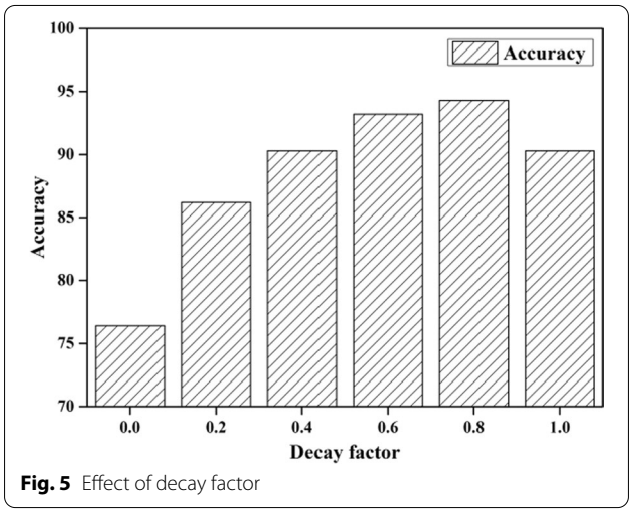

Fig. 5 Effect of decay factor

result" as follows: The doctors select 10 cases from the case base that are most similar to the target case in their mind. When the calculation result is one of these five cases, the calculation result is marked as "correct". If not, it's marked as "incorrect". On this basis, we randomly select 100 community residents, and then we can obtain the accuracy of CBR-based algorithms by comparing the calculation results and the doctors. In this experiment, the decay factor is set to 0.8. Figure 4 shows the experiment results.

As shown in Fig. 4, when $\theta=0.6$, the accuracy is the highest of various situations. As the value of the influence factor increases or decreases, the accuracies are significantly decline. When $\theta=0$, the actual distance between curves is not considered, the accuracy is lowest. When $\theta=1$, the derived function distance has no effect on the similarity calculation. The accuracy is just a little higher than $\theta=0$. The experiment results indicate that both actual distance and derived function distance have a significant impact on similarity calculation. θ is set to 0.6 in the following experiments.

Decay factor

In this experiment, we consider the effect of decay factor μ on similarity calculation. Figure 5 shows the experiment results.

As shown in Fig. 5, when $\mu=0$, the accuracy is the lowest of various situations. The main reason is that the similarity calculation method ignores the history records and the variation tendency. Only the latest history record can be considered. As the decay factor increases, the accuracy increases as well. The experiment results illustrate that the delay factor can effectively reflect the influence of time on similarity calculation. In the meantime, when μ is greater than 0.8, the accuracy is a little lower. This experiment results indicate that the recent records have more impact than the old records, and the old records gradually lose their reference value. μ is set to 0.8 in the following experiment.

Performance evaluation of proposed algorithm

In this experiment, we evaluate the performance of the proposed FDA-CBR algorithm against the GRA-CBR algorithm [16] and CS-CBR algorithm [14]. GRA-CBR algorithm is a grey relational analysis (GRA) based similarity calculation algorithm, which enables the CBR to quantify the similarity with heterogeneous multi-attributes and makes the attribute weights assignment more reasonable by considering information correlation. CS-CBR algorithm is a decision model based on Cosine similarity and Euclidean distance. Figure 6 shows the experiment results.

In Fig. 6, with the number of medical examinations of the new case increases, the accuracy of FDA-CBR increases as well. However, GRA-CBR and CS-CBR are designed without consideration of the variation tendency of the risk value. Therefore, the change in the number of

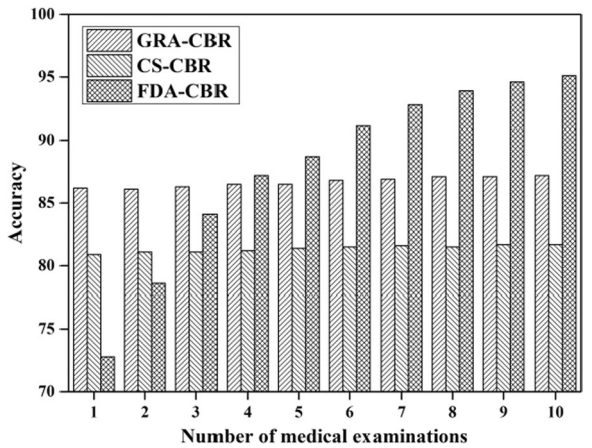

Fig. 6 Comparison of accuracy of FDA-CBR, CS-CBR, and GRA-CBR under varying number of medical examinations

medical examinations has little impact on the accuracy of GRA-CBR and CS-CBR.

In the meantime, it's worth mentioning that the accuracy of proposed algorithm is lower than two other algorithms when the number is less than or equal 2. However, when the number is greater than 4, the accuracy of FDA-CBR is significantly higher than GRA-CBR and CS-CBR. The accuracy of FDA-CBR is 9.94 and 16.41% higher than GRA-CBR and CS-CBR when the number is equal to 10. The main reason is that it is hard to identify the most similar cases just by only one or two recent medical examination. The variation tendency of the risk value is difficult to describe when there is no enough data. At this time, reducing the value of influence factor is beneficial to improve the accuracy. In addition, the experiment results indicate that the proposed FDA-CBR algorithm can effectively reveal the internal characteristics of the medical examination data and find the most similar case. It provides an effective method for the establishment of personalized intervention model for hypertension and other chronic diseases.

Conclusions

Hypertension has posed a severe threat to public health, and it creates a lot of chain problems. In this paper, we present a novel FDA-based CBR model. Firstly, the weights of input attributes are calculated by random forest algorithm. Then, the risk value of hypertension from each medical examination is evaluated according to the input data and the attribute weights. By fitting the risk values into a risk curve of hypertension, we calculate the similarity between two curves and obtain the most similar case according to the similarity. The experiment results show that the accuracy of FDA-CBR algorithm is higher than GRA-CBR and CS-CBR, increasing by 9.94% and 16.41% respectively. It provides an effective method for the establishment of personalized intervention model for hypertension and other chronic diseases.

However, as mentioned above, when the number of medical examinations of the new case is less than 3, the accuracy of FDA-CBR is a little lower than GRA-CBR. Therefore, how to adjust the similarity calculation process with the lack of input data is our future work.

Abbreviations

HER: Electronic health record; CBR: Case-based reasoning; FDA: Functional data analysis; DT: Decision tree; GRA: Grey relational analysis.

Acknowledgements

We would like to acknowledge the hard and dedicated work of all the staff that implemented the intervention and evaluation components of the study.

Author contributions

The authors contributed equally to this manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by key research and development program of Anhui province, under Grant No. 202004a05020010; Natural Science Foundation of Universities of Anhui Province, under Grant No.KJ2020A0694; Key program in the youth elite support plan in universities of Anhui province, Grant No. gxyqZD2020043.

Availability of data and materials

The data that support the findings of this study are available from Tongling Municipal Hospital but restrictions apply to the availability of these data, which were under license for the current study. Data are available from the corresponding author upon reasonable request and with permission of the Ethics Committee of Tongling Municipal Hospital and the Ethics Committee of Anhui Medical University.

Declarations

Ethics approval and consent to participate

This study received ethical approval from Ethics Committee of Tongling Municipal Hospital and Anhui Medical University. The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and research regulations of the country. Considering retrospective nature of the study, Informed consent was waived by the Ethics Committee of Tongling Municipal Hospital.

Consent for publication

Not applicable.

Competing interests

All other authors report no conflicts of interest.

Author details

[1]Department of Mathematics and Computer Science, Tongling University, Tongling 244061, China. [2]School of Public Health, Anhui Medical University, Hefei 230032, China. [3]Tianqiao Community Health Service Station, Tongling Municipal Hospital, Tongling 244061, China.

Received: 9 March 2022 Accepted: 31 May 2022
Published online: 06 June 2022

References

1. Su M, Zhang Q, Bai X, et al. Availability, cost, and prescription patterns of antihypertensive medications in primary health care in China: a nationwide cross-sectional survey. Lancet. 2017. https://doi.org/10.1016/S0140-6736(17)32476-5.
2. Gou J, Wu H. Secular trends of population attributable risk of overweight and obesity for hypertension among Chinese adults from 1991 to 2011. Sci Rep. 2021. https://doi.org/10.1038/s41598-021-85794-2.
3. Yin R, Yin L, Li L, et al. Hypertension in China: burdens, guidelines, and policy responses: a state-of-the-art review. J Hum Hypertens. 2020. https://doi.org/10.1038/s41371-021-00570-z.
4. Briggs FBS, Hill E, Abboud H. The prevalence of hypertension in multiple sclerosis based on 37 million electronic health records from the United States. European J Neurol. 2020. https://doi.org/10.1111/ene.14557.
5. Liang J, Li Y, Zhang Z, et al. Adoption of electronic health records (EHRs) in China during the past 10 years: consecutive survey data analysis and comparison of sino-american challenges and experiences. J Med Internet Res. 2021. https://doi.org/10.2196/24813.
6. Fu A, Dck AB. Using large aggregated de-identified electronic health record data to determine the prevalence of common chronic diseases in pediatric patients who visited primary care clinics. Acad Pediatr. 2021. https://doi.org/10.1016/j.acap.2021.05.007.
7. Choi YG, Hanrahan LP, Norton D, et al. Simultaneous spatial smoothing and outlier detection using penalized regression, with application to childhood obesity surveillance from electronic health records. Biometrics. 2020. https://doi.org/10.1111/biom.13404.
8. Alanazi TA, Dalia KA. Data analysis and computational methods for assessing knowledge of obesity risk factors among Saudi citizens. Comput Math Methods Med. 2021. https://doi.org/10.1155/2021/1371336.
9. Liu L, Li H , Hu Z, et al. Learning hierarchical representations of electronic health records for clinical outcome prediction. In: AMIA. Annual Symposium proceedings/AMIA Symposium. AMIA Symposium. 2020.
10. Duan J, Jiao F. Novel case-based reasoning system for public health emergencies. Risk Manag Healthc Policy. 2021. https://doi.org/10.2147/RMHP.S291441.
11. Bentaiba-Lagrid MB, Bouzar-Benlabiod L, Rubin SH, et al. A case-based reasoning system for supervised classification problems in the medical field. Expert Syst Appl. 2020. https://doi.org/10.1016/j.eswa.2020.113335.
12. Xu Z, Li S, Li H, Li Q. Modeling and problem solving of building defects using point clouds and enhanced case-based reasoning. Autom Constr. 2018. https://doi.org/10.1016/j.autcon.2018.09.003.
13. Leys C, Klein O, Dominicy Y, et al. Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. J Exp Soc Psychol. 2017. https://doi.org/10.1016/j.jesp.2017.09.011.
14. Zhang Lin, Qi Ping. Research on key technologies of personalized intervention for chronic diseases based on case-based reasoning. Comput Math Methods Med. 2021. https://doi.org/10.1155/2021/8924293.
15. Baharav TZ, Kamath GM, David NT, et al. Spectral Jaccard similarity: a new approach to estimating pairwise sequence alignments. Patterns. 2020. https://doi.org/10.1016/j.patter.2020.100081.
16. Chen W, Wang X, Wang W, et al. A heterogeneous GRA-CBR-based multi-attribute emergency decision-making model considering weight optimization with dual information correlation. Expert Syst Appl. 2021. https://doi.org/10.1016/j.eswa.2021.115208.
17. Chinedu SN, Iheagwam FN, Onuoha MK, et al. Stage 2 hypertension and electrocardiogram abnormality: evaluating the risk factors of cardiovascular diseases in Nigeria. High Blood Press Cardiovasc Prev. 2022. https://doi.org/10.1007/s40292-022-00509-6.
18. Sun JY, Ma YX, Liu HL, et al. High waist circumference is a risk factor of new-onset hypertension: evidence from the China health and retirement longitudinal study. J Clin Hypertens. 2022. https://doi.org/10.1111/jch.14446.
19. Janjua ZH, Kerins D, O'Flynn B, et al. Knowledge-driven feature engineering to detect multiple symptoms using ambulatory blood pressure monitoring data. Comput Methods Programs Biomed. 2022. https://doi.org/10.1016/j.cmpb.2022.106638.
20. Nour M, Polat K, Torres JM. Automatic classification of hypertension types based on personal features by machine learning algorithms. Math Probl Eng. 2020. https://doi.org/10.1155/2020/2742781.
21. Zhang C, Cai Y, Lin G, et al. Deepemd: few-shot image classification with differentiable earth mover's distance and structured classifiers. In: CVPR. 2020.
22. Vij S, Tayal D, Jain A. A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs. Wirel Pers Commun. 2020;111(2):1271–82.
23. Kumar V, Sood A, Gupta S, et al. Prevention-versus promotion-focus regulatory efforts on the disease incidence and mortality of COVID-19: a multinational diffusion study using functional data analysis. J Int Mark. 2021. https://doi.org/10.1177/1069031X20966563.

Publisher's Note