The Classified Prediction of Coronary Heart Disease Based on Patient Similarity Analysis

Xinyao LI

Xinjiang University

Linlin ZHANG (✉ zllnadasha@xju.edu.cn)

Xinjiang University

Xuehua BI

Xinjiang Medical University

Ying ZHANG

First Affiliated Hospital of Xinjiang Medical University

Guanglei YU

Xinjiang Medical University

Kai ZHAO

Xinjiang University

Research Article

Keywords: Patient similarity metrics, Coronary heart disease classification, LightGBM, Data mining

Posted Date: July 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-724235/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

The Classified Prediction of Coronary Heart Disease Based on Patient Similarity Analysis

Xinyao LI¹, Linlin ZHANG (✉)¹, Xuehua BI², Ying ZHANG³, Guanglei YU², Kai ZHAO¹

✉Correspondence: zllnadasa@xju.edu.cn

1 College of Information Science and Engineering (School of Cyber Science and Engineering), Xinjiang University, Urumqi 830046, China

2 College of Biomedical Engineering and Technology, Xinjiang Medical University, Urumqi 830011, China

3 First Affiliated Hospital of Xinjiang Medical University, Urumqi 830054, China

Abstract

Objective: It is important for physicians' clinical decision support to classify the coronary heart disease (CHD). Customizing personalized predictive models for patients requires selecting a patient group from an existing medical database that most closely resembles the indexed patients. In this study, we introduce a new concept that using the patient similarity for the classification of patient with CHD.

Materials and methods: We performed a structured representation of CHD patients. Obtain the multidimensional attribute distance matrix between patient pairs by calculating the multidimensional attribute distance of the patients. Predict similarity between patient pairs using machine learning (ML) models to predict clinical outcomes for indexed patients based on matched similar patients.

Results: The new measure shows marked improvements over the traditional classification measures. LightGBM is the top-performing ML model. The best model achieved 88.52% accuracy.

Conclusion: The medical applications of ML supported by similarity analytics represent a promising solution through which to reduce the physician workload to achieve the goal of “precision medicine”.

Keywords Patient similarity metrics, Coronary heart disease classification, LightGBM, Data mining

Introduction

Coronary heart disease(CHD) is the primary chronic diseases that has emerged as an increasing prevalent worldwide health concern associated with and physiological[1]. There is a great need to improve understanding of CHD and its complex factors to facilitate prevention, early detection, and improvements in clinical management. A more precise characterization of CHD patient populations can enhance our understanding of CHD pathophysiology. Current clinical definitions classify CHD into 4 subtypes: Stable angina

pectoris, Unstable angina pectoris, Ischemic cardiomyopathy, Myocardial infarction. In recent years, many scholars have integrated artificial intelligence techniques for the CHD with exciting results[2]. Therefore, new patients with CHD can be evaluated and predicted in advance.

In order to predict individualized classification prediction of patients with CHD, a detailed analysis of the characteristics of patients with CHD is required. Early studies on CHD prediction mainly applied machine learning or deep learning methods to patients' own characteristics by probabilistic statistics. Through

Received Month dd, yyyy; accepted Month dd, yyyy

E-mail: zllnadasa@xju.edu.cn

clinical manifestations such as disease symptoms, a statistical model between the probability of disease occurrence and the frequency of disease appearance is established, and the model is learned to complete the diagnostic work. However, with the development of precision medicine, this prediction method is actually inconsistent with the actual physician's diagnostic process. The physician uses analogical reasoning in the process of disease diagnosis, where the physician makes a diagnosis for a new patient based on information such as the clinical presentation, laboratory test data and other information, analogous of the patient in whom the diagnosis has already occurred. So in order to build an effective clinical decision support system, we need to quantitatively measure the distance between two patients, which is the core operation behind patient similarity[3].

Patient similarity learning is a fundamental and important task in healthcare domain, which can improve clinical decision making without additional effort to physicians[4-6]. The goal of patient similarity is to learn a meaningful distances, which refers to the selection of clinical concepts (e.g., diagnosis, symptoms, examination tests, family history, past history, drugs, surgery, genes, etc.) as patient characteristics in a specific medical setting. Quantitatively analyzing the distance between concepts will dynamically measuring the distance between patients and filtering out patient similarity groups that are similar to the index patients[7]. In addition patient similarity analysis can leverage the wealth of information in the electronic medical record (EMR) for patient-specific prediction tasks[8]. Appropriate similarity distances enable a variety of downstream applications in medicine, such as personalized medicine, medical diagnostics, trajectory analysis, and cohort studies[9]. The use of patient similarity helps to identify the exact cohort of indexed CHD patients to train a personalized model. This approach has the potential to provide customized predictions compared to traditional models trained for all CHD patients.

Previous prediction method ignored the analogical reasoning work of physicians and focused too much on data without considering empirical knowledge in the prediction process, it needs further improvement.

To address the above issues, we combines the

performance of EMR of CHD patients to form a method for representing CHD patients; proposes a method for distancing patient similarity based on patient attribute information; and examines the performance of the algorithm through systematic experiments.

The following categories of papers can be submitted to the journal:

- Consisting of discharge abstracts with comprehensive descriptions of the patients, 4910 EMRs with the first diagnosis of "coronary atherosclerotic heart disease" from a tertiary hospital in Xinjiang has been collected and labeled. Based on relevant literature and clinician's guidance, we have identified 37 attributes related to patients and made further discrete labels of them;
- We have obtained a structured patient representation based on the vectorization of the labeled data through the CatBoost-Encoder algorithm;
- The distance between patient pairs is represented by the distance metric of multidimensional attributes, the similarity of patient pairs is analyzed quickly and effectively by the application of LightGBM pattern and we have verified effectiveness of the algorithm through experiments.

Related Work

In medical field, many advances have been made in predictive models for outcome prediction, but these innovations are aimed at the general patient population and are not sufficiently tunable for individual patients. One developmental idea in is individualized predictive analysis based on patient similarity[10]. "Precision medicine" states that the clinical outcome of individual patients is determined by their genetic, genomic, physiological and clinical characteristics[11]. Therefore, it's important for "right treatment" to use all available and relevant data to determine all possible diagnoses and their disease trajectories.

For the calculation of patient similarity, the similarity is mainly weighted by traditional methods to optimize the similarity calculation such as Manhattan distance and Euclidean distance. Jia[12] et al. proposed a diagnosis prediction framework based on patient similarity, where they defined patient similarity as the

similarity between two sets of diagnoses. Q Suo[13] et al. used a convolutional neural network (CNN) to capture locally important information in the EHR, and then fed the learned representations into a modified triple-loss neural network for training. L Chan[14] et al. proposed a simSVM patient similarity algorithm with 14 similarity measures as input, the model outputs the predicted similarity or dissimilarity, and the trained model has high accuracy on the test data.

With the advent of medical informatization, EMR have become the main vehicle for recording patients' medical procedures. Therefore, it become the mainstream trend in CHD prediction which used machine learning and data mining algorithms by a large number of EMRs. It's quickly and efficiently for predict in CHD with data mining algorithm. Du[15] et al. constructed a chronic disease prediction model using cardiovascular and cerebrovascular diseases, which was modeled by the random forest algorithm in machine learning, and developed a chronic disease prediction system. Shao[16] et al. proposed a new hybrid modeling scheme for prediction of heart disease. They used logistic regression, multivariate adaptive splines, and rough set for feature dimensionality reduction. The prediction model used is neural network. In their classification tasks, the neural network model has obtained effective results. Compared with traditional prediction methods, data mining algorithm takes advantage of large amount of data, which can be more suitable for real-world disease occurrence and reflect the clinical manifestations of patients.

In the research of precision medicine, patient representation has become a research hotspot, structuring free text electronic medical records has always been the focus of many researchers[17,18]. Good results have been achieved in the research of automatic extraction of electronic medical records. Sun[19] et al. summarized and described the current state of research on EMR information extraction in free text, but also provided a review of automated structured methods that raise new questions: Existing methods tend to target only certain types of closed data sets. Even though in the specific area of EMR information extraction, the medical record data collected from different hospitals have a large variability because there is no unified writing norm for medical record text in different regions and hospitals. Consequently, models that have finished training and have achieved higher accuracy on a certain dataset often fail to achieve satisfactory results when tested with other datasets. Therefore, we choose to manually annotate the dataset to ensure the accuracy of subsequent experiments.

Material and methods

The techniques used in EMR-based similarity platform for patients with CHD, which including can be divided into several parts: data acquisition, patient representation, patient-to-attribute similarity matrix generation, LightGBM-based similarity calculation, and CHD subtype prediction. The schematic view of the method is shown in Figure 1.

Fig. 1 A schematic view of the method. **A.** Before starting the analysis, it is important to conduct a structured presentation of patients with CHD. **B.** Construct patient pairs for indexed patients, where the matched patients are CHD patients already in the database. **C&D.** Represents the process of training with the LightGBM model. **E.** CHD classification of indexed patients by matching patients' diagnoses.

EMR Access

In healthcare systems, EMRs provide the basic data support. Since EMR include admission records, medical procedure records, surgery records, discharge summaries, etc. Which are complex and hard to handle, we select the discharge summaries in EMRs as the experimental data in this paper. The reason for choosing the discharge

出院小结			
姓名: []	床号: []	住院号: []	门诊号: []
入院日期: 2020/12/22 出院日期: 2021/01/15			
入院诊断: 冠状动脉粥样硬化性心脏病, 不稳定型心绞痛, 原发性高血压, 高血压2级 (极高危)			
出院诊断: 冠状动脉粥样硬化性心脏病, 不稳定型心绞痛, 原发性高血压, 高血压2级 (极高危), 腔内动脉粥样硬化性慢性阻塞性肺病			
入院科室: 心血管内科 出院科室: 心血管科 住院天数: 24天 转科: 治愈			
入院情况: [], 男, 48岁, 以“间断胸闷、胸痛 10天”为主诉入院。既往病史: 患者自述 2020年12月10日无明显诱因出现胸闷, 位于心前区, 约……。行冠脉造影显示: 左主干末端可见局限……。诊断考虑“冠心病, 高血压2级”……。			
入院查体: 体温 36.3℃, 呼吸 18次/分, 脉搏 78次/分, 血压 140/90mmHg, ……。			
辅助检查: 2020年12月21日。在我院门诊: 血常规, 新型冠状病毒核酸, 新型冠状病毒抗体等未见异常。			
诊疗经过: 患者以“间断胸闷、胸痛 10天”为主诉入院。入院后完善相关检查及化验, 给予强心、利尿、扩冠、营养心肌等对症支持治疗。症状好转, 饮食良好, ……。术后相关检查及化验未见异常, ……。			
重要检查结果: 术前心脏彩超提示: 冠心病-冠状动脉左前降支近段局限性运动异常, ……。因患者体位及环境因素等影响, 术后经声心动图对患者心脏的结构及血流显示不清, 诊断结果仅供参考。			
出院情况: 心悸不宁、未闻及杂音。切口愈合可, 无渗出, 无红肿。			
出院医嘱: 1. 注意休息, 避免劳累, 避免感冒, 避免劳累。2. 按医嘱口服口服以下药物, 按每次 70mg 一日一次。…… 3. 全休三月, 住院期间陪护一人, 一年内避免重体力劳动。4. 定期复查肾功能。三個月後我院复查心脏彩超。定期门诊随访复查。			
随访建议: 请患者出院3月后来我院心血管内科门诊进行随访, 继续在心内科治疗。			
医生签名: []		签名时间: []	

(a)

(b)

Fig. 2 An example of part of an discharge summary from a Chinese hospital. (a) Original discharge summary record in Chinese, and (b) its' translation version in English[20].

Patient representation

The literature[21] points out that the most significant challenge in patient similarity analysis is currently addressing the problem of patient representation due to data heterogeneity. We represent patients as consisting of 37 concept vectors and discretize them by degree, but the discrete data representation is subject to disagreement after distance calculation. For example, the chest pain range feature pair (0.2,0.4) and (0.4,0.6) distances are calculated with the same result, which cannot reflect the difference and are not fully reflected in the process of model learning. Therefore, we introduced CatBoost Encode to encode the features of this study.

The biggest advantage of this coding method is that it is friendly to class variables. Moreover, CatBoost encode can support the existence of missing values with high accuracy that can be developed in the medical field.

Patient Attribute Similarity Calculation

After obtaining the patient representations, we performed a distance metric on the patient pairs. We choose a ratio of 1:100 for the construction of patient group for index patients. That is, a index patient is randomly paired with 100 matched patients for pairwise distance metrics to obtain a 100×37 patient pairwise attribute similarity matrix D , where the first 36 are feature distances, and the last one is labeled distance. We designed different calculations for the feature distance metric and the distance metric for the labels.

summary in the EMR is that the discharge summary for patients with coronary artery disease contains more detailed patient information. The discharge summary consists of many parts: admission status, treatment history, important examination results, discharge status, discharge medical advice, etc. Follow-up recommendations. As shown in Figure 2.

Discharge Summary	
Name	Bed No.
Admission Date	Outpatient No.
Discharge Date	Hospital No.
<p>Admission Diagnosis: Coronary atherosclerotic heart disease, unstable angina, essential hypertension, hypertensive disease grade 2 (very high risk)</p> <p>Discharge Diagnosis: Coronary atherosclerotic heart disease, unstable angina, primary hypertension, hypertension grade 2 (very high risk) ..</p>	
<p>Admission Department: Heart Failure Department</p> <p>Days of Hospitalization: 24</p>	<p>Discharge Department: Cardiac Surgery</p> <p>Discharge Reason: Healing</p>
<p>Admission: male, 48 years old, admitted with the chief complaint of "intermittent chest tightness and chest pain for 10 days". Brief medical history: The patient reported chest pain on December 10, 2020 with no obvious cause, located in the precordial region, ..</p> <p>On admission: temperature 36.3°C, respiration 18/min, pulse 78/min, blood pressure 140/90 mmHg..</p>	
<p>Auxiliary examinations: On December 21, 2020, at our outpatient clinic no significant abnormalities in blood count, new -coronavirus nucleic acid, and new -coronavirus antibodies.</p>	
<p>Medical treatment: The patient was admitted with the main complaint of "intermittent chest tightness and chest pain for 10 days". After admission, the patient completed relevant examinations and laboratory tests, and was ..</p>	
<p>Important findings: Preoperative cardiac ultrasound suggests: coronary artery disease: post coronary artery bypass grafting. No segmental motion abnormalities in the left ventricular wall at rest ...</p>	
<p>Discharge Status: The heart rhythm was arrhythmic and no murmur was heard. The incision healed well, with no exudation and no symptoms.</p>	
<p>Discharge orders: 1. Pay attention to rest, reasonable diet, avoid colds, avoid straining. 2. Continue to take the following drugs orally as prescribed by the doctor: Isosinadin 70mg once a day; ..</p>	
<p>Doctor Signature _____ Date _____</p>	

Feature distance calculation

From patient representation the patient's characteristics are expressed as $F = \{f_1, f_2, f_3, \dots, f_{36}\}$, then we define the characteristic distance between patient a and patient b $D_F^{a,b}$ can be represented as $D_F^{a,b} = \{d_{f_1}^{a,b}, d_{f_2}^{a,b}, d_{f_3}^{a,b}, \dots, d_{f_{36}}^{a,b}\}$, the $d_{f_i}^{a,b}$ is calculated as follows.

$$d_{f_i}^{a,b} = |a_{f_i} - b_{f_i}| \quad (1)$$

a_{f_i}, b_{f_i} denote the values of characteristics f_i in patient a and patient b, respectively.

Tag distance calculation

We use label consistency as the principle for label distance calculation, where the diagnosis of patient a and patient b are consistent is 1, and label inconsistency is 0, forming the label distance representation of patient a and patient b as in equation (2)

$$\mathbf{d}_{\text{Tag}}^{a,b} = \begin{cases} 1 & T^a = T^b \\ 0 & T^a \neq T^b \end{cases} \quad (2)$$

T^a, T^b denote the label of patient a, patient b, respectively.

LightGBM-based patient similarity prediction

After patient representation and patient pair similarity calculation, the attribute similarity matrix of different patient pairs is obtained. The next step is input to the prediction model for prediction. To compare the model prediction detection effect, seven supervised machine learning methods are used in this bit to predict

patient similarity.

LightGBM (Light Gradient Boosting Machine) is a supervised learning algorithm implemented in the Gradient Boosting Decision Tree (GBDT). The biggest advantage for LightGBM is solve the problems encountered by GBDT in massive data, so that can be used in industrial practice better and faster. Therefore, LightGBM has the advantages of faster training speed, lower memory consumption, better accuracy and the possibility of parallelized learning. It is very suitable for large-scale medical data mining. Among them, LightGBM algorithm principles mainly include histogram algorithm, leaf-wise leaf growth strategy with depth restriction and histogram optimization[22].

Based on the above advantages, using the LightGBM algorithm for patient similarity prediction tasks can achieve better results in the shortest time using fewer computational resources. But that's not all. We also

conducted experiments on the results of different machine learning in patient similarity predicting. Detailed data can be found in the experimental results section.

Experiment Data

For the problem of predicting similarity in patients with CHD, the best data should be those that provide a complete description of the patient's condition. We collected discharge summaries from the cardiology department of a hospital in Xinjiang, all of which were patients who were already the first diagnosis of coronary atherosclerotic heart disease. The second diagnosis of which is used as a label for the data, and the data distribution is shown in Table 1.

Table 1 Data Distribution

Diagnosis	Stable angina pectoris	Unstable angina pectoris	Ischemic cardiomyopathy	Myocardial infarction
Amount	927	1594	1214	1175

Experimental procedure

I. Data pre-processing

We constructed an EMR-based attribute representation of CHD patients, including categories, attribute names and their corresponding attribute values. The attributes with more than 30% missing values in all features were removed, and the obtained categories and attributes of patients with CHD are shown in Table 2.

The processing steps for EMR are as follows:

- 1) Use the annotation tool to annotate the attributes in the discharge summary
- 2) Extracting the values of each attribute of the medical record. There are three main forms of values in

the EMR: binary, numeric and textual features. For binary data, we extract binary features from qualitative text descriptions such as "suffering from XXX disease" and "denying XXX disease" in the medical record. For numeric type data, the process is to extract the text that quantitatively describes the features in the medical record. For example, if the blood pressure is xx mmHg, the value is extracted. For text features, the attribute value descriptions corresponding to the criteria given by the doctor are extracted.

- 3) The different attributes of each patient were organized into a table according to Table 2.

Table 2 Categories and attributes of patients with coronary artery disease

Category	Attribute
Background	Age; Sex
Symptom	Range of chest pain; Quality of chest pain; Duration; Triggering factors
Laboratory	Blood pressure; White blood cell value; Fasting blood glucose value; Hemoglobin value; Glycated hemoglobin values; Lactate dehydrogenase; Glutathione transaminase; Phosphocreatine kinase; Phosphocreatine kinase isoenzyme; Troponin I; Troponin T; D-Dimer; BNP
RiskFactor	Hypertension; Diabetes
Image	Resting cardiogram; Ultrasound of the heart; Carotid artery cardiogram; ST segment; Heart Rhythm; Heart rate; Premature beats

Operation	CTA Number of lesion branches; CTA Narrow situation; PCI
Drug	ACEI/ARB、 β -blockers; Calcium channel inhibitors; Modulating drugs; Antiplatelet therapy
Diagnosis	The second diagnosis

Table 3 Results of different coding methods

Coding method	One-hot	Label	Skip-gram	Original	CatBoost
Accuracy	80.53%	80.46%	80.6%	80.2%	88.52%

II. Model construction and parameter setting

We require a pairwise characteristic distance metric for patient pairs. We selected 400 index patients from each diagnostic group for patient pair construction, using a 1:100 ratio for patient pair selection, matching patients were randomly selected from the total patient pool. Generating a total patient pair similarity matrix of 160,000*37. All random number seeds are set to 2021.

According to tag consistency we could classifies the patients similarity into similar 1 and dissimilar 0. Therefore, we converts the multiclassification problem into a second classification problem using the patient similarity approach, then uses LightGBM for prediction.

III. Evaluation Metrics

Accuracy, P (Precision), R (Recall) and F1 (F1-Measure) are used as evaluation metrics for similarity calculation models.

P denotes the proportion of the calculated similar patient pairs that are actually similar, R denotes the ratio of the predicted correct similar patients to the actual similar patient pairs, and the F1 value is a weighted summed average of precision and recall, and they are calculated as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$P = TP / (TP + FP) \quad (4)$$

$$R = TP / (TP + FN) \quad (5)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (6)$$

Where TP (True Positive) denotes a correctly calculated similar patient pair, TN (True Negative) denotes a correctly calculated dissimilar patient pair, FP (False Positive) denotes a wrongly calculated similar patient pair, and FN (False Negative) denotes a wrongly

calculated dissimilar patient pair.

Results

Patients indicated learning impact analysis

From Section 3.3, we propose the use of CatBoost Encode for vectorized encoding of CHD patients. We analyze and compare the advantages of CatBoost Encode by comparing the encoding method proposed by other research scholars and the non-encoding method for comparison experiments.

We use the category features commonly used as a control group, which are one-hot coding, label coding, using Skip-Gram coding, and no coding treatment. All tests use LightGBM as a prediction model to compare the accuracy. The experimental results are shown in Table 3.

As the table shows, CatBoost Encode far outperforms the other four encoding methods in terms of Accuracy values, and it can be seen that CatBoost Encode is well suited for encoding our data.

Distance calculation method impact analysis

Based on the above experimental parameters, we tests the effect of using the attribute similarity calculation method in the LightGBM model. Referring to the literature[6] using the Jaccard distance with one-hot encoding and the Euclidean distance as comparison experiments, respectively. Considering that using the above two distances require classify the features, we classify the features according to Table 2 as the base vector for the distance metric.

For example, under the Euclidean distance, the similarity of the laboratory tests of patient a and patient b is

$$D_L^{a,b} = \sqrt{(a_{f_1} - b_{f_1})^2 + (a_{f_2} - b_{f_2})^2 + \dots + (a_{f_i} - b_{f_i})^2} \quad (7)$$

For example, under the Jaccard distance, the similarity of risk factors for patient a and patient b is

$$D_R^{a,b} = \frac{|A \cap B|}{|A \cup B|} \quad (8)$$

Where A denotes the set consisting of risk factor class characteristics of patient a coded by one-hot and B denotes the set consisting of risk factor class characteristics of patient b coded by one-hot.

The experimental results are shown in Table 4:

Table 4 Results of different distance metrics

Method	Accuracy	P	R	F1
Euclidean	81.77%	85.16%	93.26%	89.04%
Jaccard	80.20%	75.40%	77.95%	76.65%
Our method	88.52%	92.57%	93.01%	92.79%

By comparing the experiments, it can be seen that the distance calculation method proposed is better than the distance measure which proposed by previous scholars. Our analysis suggests that traditional distance metrics based on Euclidean distance and other traditional distance metrics operate to a certain extent on feature fusion, it actually reduces the dimensionality of a small number of features, so the model loses a lot of information in the learning process, leading to a decrease in accuracy.

Predictive Model Analysis

Based on the evaluation metrics, in order to verify the effectiveness of the LightGBM model in the task of calculating patient similarity prediction, we used the calculated pairwise patient similarity matrices as data with 10-fold cross-validation using K-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF), CatBoost, and XGBoost 6 different models of similarity computation performance, the experimental results of model comparison are shown in Table 5 and Figure 3.

Table 5 Results of different models

Model	Accuracy	P	R	F1
KNN	87.85%	91.99%	92.77%	92.38%
LR	84.53%	89.34%	91.34%	90.33%

SVM	85.79%	93.44%	88.14%	90.71%
RF	88.11%	93.34%	92.13%	92.73%
CatBoost	86.45%	90.78%	92.3%	91.53%
XGBoost	86.34%	92.93%	89.46%	91.16%
LightGBM	88.52%	92.57%	93.01%	92.79%

Fig.3 Comparison of experimental results

From the above table, we know that LightGBM is the best model among all models in terms of Accuracy and F1 values, but RF and KNN have similar calculation results with LightGBM. Considering the large scale of medical data, we compared the running time of the above three models, and the running time comparison of different models is shown in Table 6.

Table 6 Results of model runtime

Model	Runtime
KNN	37.82s
RF	43.56s
LightGBM	15.9s

According to the running time comparison results, it can be seen that LightGBM model has the shortest running time of 15.9s, KNN is the second and RF is the slowest. In terms of running time, LightGBM saves twice as much time as the other models, which shows that LightGBM can significantly improve the accuracy of predicting patient similarity while getting the prediction results more quickly.

Discussion

In this paper, we discussed the use of patient similarity in clinical decision support for CHD. The experiment proved that the similarity analysis of patients

can achieve a good function in the classification prediction of CHD. The process of patient similarity is set according to the diagnostic process of clinicians, and it is very consistent with the needs of modern society to make this process intelligent. Patient similarity analysis, to a large extent, realizes personalized treatment for a single patient, solves the problem of complicated workload in precision medicine, and makes use of valuable data in medical database.

However, the method is still limited to the direction of classification of CHD, we can expand the dataset to improve the generalizability of the algorithm and expand the application in other medical fields as well.

Conclusion

Patient similarity prediction is a fundamental problem in the field of disease health and it has a wide range of applications, such as disease prediction, disease evolution and treatment selection. In our study, we apply patient similarity to perform CHD classification task, develop annotation criteria for CHD patients based on EMR discharge summaries, using CatBoost Encode for structured presentation. After that, we obtain similarity matrices of patients by distances multidimensional features, predictive learning of patient similarity is performed by LightGBM model. Experimental results show that our proposed method achieves better results than the baseline method in the classification and prediction task of CHD.

Abbreviations

CHD: coronary heart disease; ML: machine learning; EMR: electronic medical record; HER: electronic health records; CNN: convolutional neural network; LightGBM: Light Gradient Boosting Machine; GDBT: Gradient Boosting Decision Tree; P :Precision; R: Recall; F1: F1-Measure; TP: True Positive; TN :True Negative; FP: False Positive; FN: False Negative; KNN: K-nearest neighbors; LR: logistic regression; SVM: support vector machine; RF: random forest.

Declarations

Ethics approval and consent to participate

This study has been approved by the Ethics Committee of the First Affiliated Hospital of Xinjiang Medical University (in these studies). All the data are anonymized before being used. All methods are in accordance with relevant guidelines and regulations (such as Helsinki Guidelines).

Consent for publication

No applicable.

Availability of data and materials

The datasets analyzed during the current study are not publicly available as they contain information that are sensitive to the institution, but are available from the corresponding author on reasonable request

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant [2019D01C041], Grant [2019D01C062], Grant [2019D01C205], and Grant [2020D01C028]; in part by the Graduate Research Innovation Project of Xinjiang Uygur Autonomous Region under Grant [XJ2020G065]; in part by the National Natural Science Foundation of China under Grant [12061071] and Grant [61867006]; in part by the Higher Education of Xinjiang Uygur Autonomous Region under Grant [XJEDU2017M005], Grant [XJEDU2020Y003], and Grant [XJEDU2019Y006]; in part by the Funds for Creative Research Groups of Higher Education of Xinjiang Uygur Autonomous Region under Grant [XJEDU2017T002]; in part by Tianshan Innovation Team Plan Project of Xinjiang Uygur Autonomous Region under Grant [202101642]; in part by Xinjiang Uygur Autonomous Region Major Science and Technology Special Project under Grant [2020A02001-1]; and in part by the Major Science and Technology Project of Sichuan Science and Technology Plan Grant [2020YFQ0018].

Authors' contributions

Xinyao LI is the first author, Linlin ZHANG is the corresponding author. Xinyao LI, Linlin ZHANG and Xuehua BI wrote the

main manuscript text and Ying ZHANG provided data support. All authors reviewed the manuscript.

Acknowledgements

We would like to thank Ms. Ying ZHANG of the First Affiliated Hospital of Xinjiang Medical University for her data support, and Shiwei XU of Xinjiang University and other students for their input in data processing.

References

1. Junaaid M J A, Kumar R. Data Science And Its Application In Heart Disease Prediction In the proceedings of the 2020 International Conference on Intelligent Engineering and Management (ICIEM), June 17-19, 2020, London, UK, pp:396-400. DOI: 10.1109/ICIEM48762.2020.9160056.
2. Alizadehsani R, Abdar M, Roshanzamir M, et al. Machine Learning-Based Coronary Artery Disease Diagnosis: A Comprehensive Review. *Computers in Biology and Medicine*, 2019, 111: 103346. arXiv:1902.03376
3. WANG F, SUN J. PSF: A Unified Patient Similarity Evaluation Framework Through Metric Learning With Weak Supervision. *IEEE Journal of Biomedical and Health Informatics*, 2015, 19(3): 1053–1060. DOI:10.1109/JBHI.2015.2425365.
4. Belle A, Thiagarajan R, Soroushmehr S M R, et al. Big Data Analytics in Healthcare. *BioMed Research International*, 2015: 370194. DOI:10.1155/2015/370194.
5. P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: Leveraging patient similarity and drug similarity analytics," In *Proceedings of AMIA Summits Transl. Sci.*, 2014, pp. 132–136. PMID: PMC4333693.
6. Lee J, Maslove D M, Dubin J A. Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric. *PLoS ONE*, 2015, 10(5): 0127428. DOI:10.1371/journal.pone.0127428
7. Jia Z, RuiLong Z, HuiLong D, et al. A review of electronic medical record-based patient similarity analysis(in Chinese). *Chinese Journal of Biomedical Engineering*, 2018, 37(03): 353-366. DOI:https://doi.org/10.1016/j.ijmedinf.2019.104073.
8. Gottlieb A, Stein G Y, Ruppin E, et al. A Method for Inferring Medical Diagnoses from Patient Similarities. *BMC Medicine*, 2013, 11: 194. DOI:10.1186/1741-7015-11-194.
9. Zhu Z, Yin C, Qian B, et al. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. In *Proceedings of the IEEE 16th International Conference on Data Mining (ICDM)*, , 2016, 750-758
10. Sharafoddini A, Dubin J A, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Medical Informatics*, 2017, 5(1): e7. DOI:10.2196/medinform.6730.
11. Pai S, Bader G D. Patient Similarity Networks for Precision Medicine. *Journal of Molecular Biology*, 2018, 430(18): 2924–2938. https://doi.org/10.1016/j.jmb.2018.05.037.
12. Jia Z, Zeng X, Duan H, et al. A Patient-Similarity-Based Model for Diagnostic Prediction. *International Journal of Medical Informatics*, 2020, 135: 104073.https://doi.org/10.1016/j.ijmedinf.2019.104073
13. Suo Q, Ma F, Yuan Y, et al. Deep Patient Similarity Learning for Personalized Healthcare. *IEEE Transactions on Nanobioscience*, 2018, 17(3): 219–227. DOI:10.1109/TNB.2018.2837622.
14. Chan L W C, Chan T, Cheng L F, et al. Machine Learning of Patient Similarity: A Case Study on Predicting Survival in Cancer Patient after Locoregional Chemotherapy. In: *Proceedings of the 2010 IEEE International Conference on Bioinformatics and Biomedicine*. 2010, 467-470.
15. Du Z, Yang Y, Zheng J, et al. Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation. *JMIR Medical Informatics*, 2020, 8(7): e17257. DOI:10.2196/17257
16. Shao Y E, Hou C-D, Chiu C-C. Hybrid Intelligent Modeling Schemes for Heart Disease Classification. *Applied Soft Computing*, 2014, 14: 47–52. https://doi.org/10.1016/j.asoc.2013.09.020.
17. Zhang J, Chang D. Semi-Supervised Patient Similarity Clustering Algorithm Based on Electronic Medical Records. *IEEE Access*, 2019, 7: 90705–90714. DOI:10.1109/ACCESS.2019.
18. Hu F, Li L, Huang X, et al. Symptom Distribution Regularity of Insomnia: Network and Spectral Clustering Analysis. *JMIR Medical Informatics*, 2020, 8(4): e16749. DOI:10.2196/16749
19. Sun W, Cai Z, Li Y, et al. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering*, 2018, 2018: 4302425. DOI:10.1155/2018/4302425

20. Hu D, Huang Z, Chan T-M, et al. Utilizing Chinese Admission Records for MACE Prediction of Acute Coronary Syndrome. *International Journal of Environmental Research and Public Health*, 2016, 13(9), Paper 912. <https://doi.org/10.3390/ijerph13090912>
21. Seligson N D, Warner J L, Dalton W S, et al. Recommendations for Patient Similarity Classes: Results of the AMIA 2019 Workshop on Defining Patient Similarity. *Journal of the American Medical Informatics Association : JAMIA*, 2020, 27(11): 1808–1812.
22. Zhang D, Gong Y. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. *IEEE Access*, 2020, 8: 220990–221003. DOI:10.1109/ACCESS.2020.3042848.