International Journal of Medical Informatics 115 (2018) 10–17

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Mortality prediction system for heart failure with orthogonal relief and dynamic radius means

Zhe Wang[a,b,*], Lijuan Yao[b], Dongdong Li[b], Tong Ruan[b,*], Min Liu[c], Ju Gao[c,*]

[a] *Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai, 200237, China*
[b] *Department of Computer Science & Engineering East China University of Science and Technology, Shanghai, 200237, China*
[c] *Shanghai Shuguang Hospital, Shanghai, 200021, China*

ARTICLE INFO

Keywords:
Mortality prediction
Heart failure
Feature selection
Dynamic radius
Machine learning

ABSTRACT

Objective: This paper constructs a mortality prediction system based on a real-world dataset. This mortality prediction system aims to predict mortality in heart failure (HF) patients. Effective mortality prediction can improve resources allocation and clinical outcomes, avoiding inappropriate overtreatment of low-mortality patients and discharging of high-mortality patients. This system covers three mortality prediction targets: prediction of in-hospital mortality, prediction of 30-day mortality and prediction of 1-year mortality.
Materials and methods: HF data are collected from the Shanghai Shuguang hospital. 10,203 in-patients records are extracted from encounters occurring between March 2009 and April 2016. The records involve 4682 patients, including 539 death cases. A feature selection method called Orthogonal Relief (OR) algorithm is first used to reduce the dimensionality. Then, a classification algorithm named Dynamic Radius Means (DRM) is proposed to predict the mortality in HF patients.
Results and discussions: The comparative experimental results demonstrate that mortality prediction system achieves high performance in all targets by DRM. It is noteworthy that the performance of in-hospital mortality prediction achieves 87.3% in AUC (35.07% improvement). Moreover, the AUC of 30-day and 1-year mortality prediction reach to 88.45% and 84.84%, respectively. Especially, the system could keep itself effective and not deteriorate when the dimension of samples is sharply reduced.
Conclusions: The proposed system with its own method DRM can predict mortality in HF patients and achieve high performance in all three mortality targets. Furthermore, effective feature selection strategy can boost the system. This system shows its importance in real-world applications, assisting clinicians in HF treatment by providing crucial decision information.

1. Introduction

Heart failure (HF) is a physiological state that the cardiac output is insufficient to meet the needs of body. Clinically, HF could be characterized by some typical symptoms and signs. The former includes the shortness of breath, ankle swelling and fatigue [1,2]. While the latter contains the elevated jugular venous pressure, pulmonary crackles and peripheral edema [3,4]. Meanwhile, HF may result in a reduced cardiac output and/or elevated intracardiac pressures [5–8]. As an important part of the global chronic cardiovascular disease, HF could be caused by many severe cardiovascular diseases. HF also tends to be the first cause of death that the cardiovascular disease leads to. Epidemiological data shows that the probability of adults suffering from HF is 1%–2%. But with the increase of age, the probability of the elder over 70 years old is even more than 10%. Furthermore, the number of patients suffering from HF is growing [9,10]. Therefore, predicting the mortality for HF patients is definitely necessary for assisting clinicians to make the optimal decision during the therapeutic process. Considering the mortality prediction reports, clinicians can take special treatments for patients with various degrees of HF. Consequently, effective mortality prediction can improve clinical outcomes and resources allocation, avoiding the inappropriate overtreatment of low-mortality patients and discharging of high-mortality patients [11].

Nevertheless, it is not easy to predict the mortality in HF patients. Because the mortality varies from individual patients and the factors that affect mortality are diverse. Recently, medical institutions have accumulated immense amounts of clinical data of HF, such as specialist diagnosis, laboratory tests, medications, operations. The abundant and

* Corresponding authors at: Department of Computer Science & Engineering East China University of Science and Technology, Shanghai, 200237, China.
E-mail addresses: wangzhe@ecust.edu.cn (Z. Wang), ruantong@ecust.edu.cn (T. Ruan), gaoju@smmail.cn (J. Gao).

https://doi.org/10.1016/j.ijmedinf.2018.04.003
Received 18 September 2017; Received in revised form 3 April 2018; Accepted 8 April 2018
1386-5056/

important information about patients is kept in Electronic Health Records (EHRs) [12]. The information about EHRs has potential practical value because of its comprehensiveness and authenticity, which deserve our exploration. For instance, Connelly D P, Park Y T, Du J, et al. [13] evaluated if EHR has observable effects on care outcomes. The results pointed in the direction of some impact of an EHR on some measures of healthcare utilization and outcomes and laid the groundwork for further, more definitive studies in the area. Moreover, Maryam Panahiazar et al. [14] developed a risk prediction model by machine learning techniques. Meanwhile, results showed that the model which was built using EHR data was more accurate (11% improvement in AUC). In this paper, we acquire a real-world data from the EHR of HF patients in Shanghai Shuguang hospital.

The last two decades have witnessed the rapid development of machine learning for medical care. Researches introducing techniques into HF diagnosis have achieved positive results. For instance, Wu, Jionglin et al. [15] presented that HF was predicted more than 6 months before clinical diagnosis and AUC can reach 77%. Banerjee D, Thompson C, Kell C, et al. [16] instituted EMR-based measures designed to improve cohort identification, intervention tracking, and readmission analysis. They significantly reduced 30-day index hospital all-cause HF readmission rates from 18.2% at baseline to 14%. Moreover, Evans R S, Benuzillo J, Horne B D, et al. [17] developed an automated identification and predictive tools to help identify high-risk heart failure patients with significant reduction in 30-day mortality and a significant increase in patient discharges to home care. Choi E, Schuetz A, Stewart W F, et al. [18] applied recurrent neural network to early detection of heart failure onset. They used a 12-month observation window, the area under the curve (AUC) for the RNN model was 77.7%. In addition, Deekshatulu B L, Chandra P [19] combined KNN with genetic algorithm for effective classification. And experimental results showed the enhancement of accuracy in diagnosis disease.

As stated above, scholars and researchers did a lot of work about HF disease, and machine learning played a great role. By their inspirations, we apply the machine learning into mortality prediction for HF patients. The neighbor method is selected as classifier in our paper owing to its simple, popular and highly efficient. Moreover, the neighbor method takes less time to predict mortality without training. Meanwhile, we propose a novel neighbor method called Dynamic Radius Means (DRM) that is more excellent than the traditional neighbor methods. In addition, considering that the dataset contains the redundant and irrelevant attributes, a selection feature method called Orthogonal Relief (OR) is applied into feature processing. Effective feature selection strategy can boost the system. We build a mortality prediction system depending on the demand of hospital. This prediction system covers three mortality prediction targets: predication of in-hospital mortality, 30-day mortality and 1-year mortality. Clinicians and specialists can select different prediction target they need to assist diagnosis depend on crucial decision information the system provides.

2. Materials and methods

2.1. System architecture

The mortality prediction system architecture is shown in Fig. 1. The processing pipeline consists of four parts: data description, prediction targets, feature selection and classification.

First, HF dataset is collected from Shanghai Shuguang hospital. Most of the acquired information is stored in the categorical form, such as gender, medication. We transfer the text or nonnumerical data into numerical format in order that the dataset can be utilized by machine learning. Then there are three mortality prediction targets: prediction of in-hospital mortality, 30-day mortality and 1-year mortality. Clinicians and specialists can select different prediction target they need to assist diagnosis depending on crucial decision information system provides. Afterwards, we deal with the features of data at feature selection stage. Effective feature selection is beneficial to reduce computation time and improve system performance [20–23]. In this paper, a classical feature selection algorithm OR [23] is employed to remove the irrelevant and redundant features. Finally, we introduce a novel algorithm DRM to predict the mortality for HF patients at the classification stage.

2.2. Data description

HF dataset is collected from Shanghai Shuguang hospital which is a large hospital of integrated with Chinese and Western Medicine. We name this dataset with Shanghai Shuguang Heart Failure (SSHF) dataset for convenience. SSHF are extracted from encounters occurring between March 2009 and April 2016.

We identify a cohort of patients based on meeting the following criteria: (1) Having a principle ICD-10-CM diagnosis of HF (codes: I11.0, I13.0, I13.2, I50, I50.1, I50.2, I50.3, I50.4 and I50.9). (2) A minimum of 1 HF therapy is initiated within the first 2 days of hospitalization. Thus, 10,203 in-patients records are left. The records involve 4682 patients with a median follow-up of 0.96 years, during which 539 patients are dead in hospital.

There are 1302 features in SSHF, and the explanations of the features are shown in Table 1. Meanwhile Table 1 shows the range values of features in SSHF. The value of Laboratory tests are categorical values and consist of three attributes. Features (1, 0, 0), (0, 1, 0) and (0, 0, 1) indicate the value of test to represent higher, normal, lower than normal value.

For each patient in SSHF, we hold some general demographic details (i.e., age, and gender), and common clinical descriptors are available in a structured format: heart rate, diagnoses, medications and laboratory tests. For age and heart rate, we keep the original numerical value. The male is converted to digit 1 and the female is converted to digit 0 for gender. With regard to common clinical descriptors, we use a vector based representation in order that clinical events in CDR data can be represented as computable event sequences. Thus we construct a vector to represent the feature appeared in the CDR for every HF case. The value of vector on each dimension is in the presence of the corresponding medical event in a specific time period.

The handling of the normalized records is differed by event type and is described as follows. For diagnoses, 1222 distinct ICD-10-CM diagnosis codes are discovered from the selected HF cases. For medications, 61 widely used medications in China are selected by our specialists manually, which are further grouped into 11 classes as seen in Table 2. In addition, 22 related laboratory tests are chosen for this research via our specialists, and Table 2 describes the tests in detail. On basis of the reference value of each lab test, a flag, including High (H), Low (L) and Normal (N) was used to denote the result.

2.3. Feature selection with OR

Feature selection is an indispensable component in proposed mortality prediction system since valid feature selection can boost the system [20–23]. The mortality prediction system adopts a classical feature selection algorithm OR [23]. The OR has two main steps: removing the irrelevant features and removing the redundant features.

In the first step, each feature is calculated to a feature score that denotes the relevance to classification. The high scores are assigned to all discriminative features. The score s of feature F_i is defined as:

$$s = \frac{1}{\mathrm{N}}\sum_{k=1}^{N}\left[\left|\mathrm{F}_i(k) - \mathrm{F}_i^{\mathrm{M}}(k)\right| - \left|\mathrm{F}_i(k) - \mathrm{F}_i^{\mathrm{H}}(k)\right|\right] \tag{1}$$

Where, N denotes the number of samples. $F_i^M(k)$ and $F_i^H(k)$ denote the values of feature of the nearest-miss and the nearest-hit samples of sample k, respectively [23]. The nearest-hit sample is defined as the nearest neighboring sample of the same class, while the nearest-miss