A multiple measurements case-based reasoning method for predicting recurrent status of liver cancer patients

Xiao-Ou Ping [a], Yi-Ju Tseng [b,e,*], Yan-Po Lin [b], Hsiang-Ju Chiu [b], Feipei Lai [a,b,c], Ja-Der Liang [d], Guan-Tarn Huang [d], Pei-Ming Yang [d]

[a] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[b] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
[c] Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[d] Department of Internal Medicine, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan
[e] Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Children's Hospital Boston, Boston, MA, USA

ARTICLE INFO

ABSTRACT

In general, the studies introducing the medical predictive models which frequently handle time series data by direct matching between pairs of features within sequences during calculation of similarity may have following limitations: (1) direct matching may not be a suitable matching because these paired cases by a fixed order may not be with the most similar temporal information, and (2) when two patients have different numbers of multiple cases, some cases may be ignored. For example, one patient with four cases and another one with five cases, only first four cases of these two patients are paired and the left one case may be ignored. In this paper, in order to dynamically determine matching pairs among cases and pair all cases between two patients, we propose a multiple measurements case-based reasoning (MMCBR) to be used for building liver cancer recurrence predictive models. MMCBR and single measurement case-based reasoning (SingleCBR) are evaluated and compared. According to experiment results in this study, the performance of MMCBR models is better than that of SingleCBR models. Multiple measurements accumulated during a period of time do have benefits for building predictive models with improved performance based on this proposed MMCBR method.

1. Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver cancer [1–3]. In a research in the United States from 1975 to 2005, both incidence and mortality rates of HCC continue to increase [4]. Furthermore, in Taiwan, HCC has been the leading cause of cancer death [5]. Radiofrequency ablation (RFA) is one of local ablative therapies [6,7] and RFA is a more effective therapy than other local ablative therapies [8].

However, RFA may have high rate of recurrence after treatment [9]. Recently, most of proposed studies are for analyzing the risk factors for recurrence of HCC patients after RFA [10–13] but not for introducing the predictive recurrence models with evaluated results such as sensitivity, specificity, and accuracy. When the patients are predicted as potential HCC recurrence by the

predictive recurrence model, frequent follow-up examinations may be applied on these patients for early detecting and treating the recurrence status.

In this study, clinical data from the RFA treatments to 180 days before these treatments were collected for developing recurrence predictive models. Static features (e.g., gender) and dynamic features (e.g., laboratory tests) are both included in data sets. According to literature review, approaches based on case-based reasoning (CBR) which can well handle time series data have been proposed [14–18]. However, the time series data in this study are not measured as frequently as signal data (which may have multiple measurements within a day). In this study, serum laboratory tests may change over time and different laboratory tests may have different measured frequencies (e.g., from several weeks to several months among 180 days). The amount of each patient's data was also different. Some patients may have more measurements within 180 days and some patients may have fewer measurements within 180 days. Because of above characteristics of these data, direct matching between pairs of features within sequences during calculation of similarity may have the following

limitations: (1) the pairs of features within sequences are matched according to a fixed order. The nth value of a feature from one sequence is matched with the nth value of a feature from another sequence. Sometimes, this matching may not be a suitable matching. For example, from a sequence, the nth value is measured 3 weeks before the RFA and from another sequence, the nth value is measure 2 months before the RFA and $(n-1)$th value is measure 4 weeks before the RFA. Then, matching the nth value from the former sequence with the $(n-1)$th value from the latter sequence is more suitable than matching pairs of features within sequences with a fixed order, and (2) when two patients have different numbers of multiple cases, some cases may be ignored. For example, one patient with four cases and another one with five cases, only first four cases of these two patients are paired and the left one case may be ignored. In this paper, in order to dynamically determine matching pairs among cases and pair all cases between two patients, we propose a multiple measurements case-based reasoning (MMCBR). In this study, CBR was used for building liver cancer recurrence predictive models and these models were used for predicting whether a patient has recurrent HCC in 1 year after the patient received RFA treatment.

Because different laboratory tests may have different measured frequencies, multiple measurements should be processed before performing classification. Before performing MMCBR, multiple measurements are merged based on different time periods for constructing multiple cases of a patient. Several different periods are tested.

Classification results of getting only one value of each feature named as single measurement case-based reasoning (SingleCBR) is regarded as baseline in this study for evaluating performance of predictive models based on multiple measurements with different merging time periods.

2. Literature review

In recent survey studies, researchers have reviewed more than 30 case-based reasoning (CBR) systems/projects [19,20] and these studies reveal that CBRs have been widely employed in medical domain, including disease diagnosis, classification, treatment, and management.

Numerous studies are proposed for handling time series data based on CBR. Temporal abstraction is widely used in medical informatics for handling time series data which can be used for reducing the dimensionality of data, abstracting variations of data, and further summarizing information of data. Schmidt and Gierl [14] proposed a prognostic model which combined temporal abstractions with case-based reasoning and was applied on the prognosis of kidney function and the infectious diseases (e.g., influenza). In prognosis of kidney function courses, the data were abstracted by states of the renal function and temporal abstraction was performed for generating three trend descriptions. In prognosis of influenza, temporal abstraction was performed and several assessments were used for describing trends, including enormous decrease, sharp decrease, decrease, steady, increase, sharp increase, and enormous increase. Leonardi et al. [15] implemented a case-based architecture with temporal abstraction in the domain of renal disease. A temporal abstraction processing module was developed for abstracting states, trends, and significant combinations of both. They took haematic volume (HV) as an example. Expected trends may contain exponential decrease, linear decrease, fall decrease, stable, and increase. Expected states may contain positive state and negative state.

Some CBR-based studies process time series data from signals. Montani et al. [16] proposed a CBR system to support the treatment of end stage renal failure patients. They used discrete Fourier transform for reducing the dimensionality of time series data.

Temporal constraint network (TCN) considers temporal relations between temporal events. Juarez et al. [17] proposed T-CARE for temporal case retrieval in an intensive care burn unit which included TCN-based function [18]. Temporal cases were modeled based on possibilistic temporal constraint network (PTCN). PTCN regards each pair of matched events as a single merged point. The temporal relation between two merged points is described by three values (before, at the same time, or after) which are the degrees of possibility.

3. Method

In MMCBR, the temporal information is the number of days before treatment of a case in this study. MMCBR would pair patients' cases based on temporal information (i.e., *Treatment Distance* and *Case Distance*, which are calculated by temporal information and are introduced in Section 3.2). An example of case pairing is shown in Fig. 1. Each patient may have multiple cases. Patient A has three cases (C_{A1}, C_{A2}, and C_{A3}) and patient B has two cases (C_{B1} and C_{B2}) from the RFA treatment time points to 180 days before their treatments. These cases are ordered by their temporal information. Cases are paired based on smaller time distance between their temporal information. For example, C_{A1} is paired with C_{B1} because the distance between C_{A1} and C_{B1} (25 days) is smaller than that of C_{A1} and C_{B2} (110 days). C_{A2} is paired with C_{B1} because the distance between C_{A2} and C_{B1} (15 days) is smaller than that of C_{A2} and C_{B2} (70 days). C_{A3} is paired with C_{B2} because the distance between C_{A3} and C_{B2} (30 days) is smaller than that of C_{A3} and C_{B1} (115 days). *Case similarity* denotes similarity of two cases and is measured based on two paired cases. *Patient similarity* denotes the similarity of two patients and is measured based on case similarities of two patients. When SingleCBR is performed, each patient only takes one case which is closest to RFA treatment. For example, in SingleCBR, only patient A's case C_{A1} and patient B's case C_{B1} are included and patient similarity between two patients is equal to case similarity of C_{A1} and C_{B1}. To pair multiple cases between two patients, *case pairing* is necessary. After cases between two patients are paired, a similarity of two paired cases, named case similarity, can be calculated based on feature weights and feature similarities. A similarity of two patients, named patient similarity, can be calculated based on case similarities and *case weights* in case weight mode, or case similarities and *pair weights* in pair weight mode.

Relevant algorithms are introduced in the following content, including data merging, case pairing, calculation of feature weights, feature similarities, case similarities, case weights, and pair weights, and classification.

3.1. Merging data for constructing multiple cases

One clinical feature may have different measurement results at various time points. In this study, a patient's *case* means a set of measurement results (i.e., measurement results of 16 features in this study) and the temporal information which is the distance between the date of measuring these laboratory features and the RFA treatment. Clinical data from the RFA treatment to 180 days before this treatment were collected. Fig. 2 illustrates data merging based on a specific time period. There are 16 features (from F_1 to F_N and N is 16 in this study). In a specific time period, there are several laboratory test results which are measured at different time points. These laboratory test results are merged as a patient's case [21]. Patient's cases are formed by merging feature values based on different time periods. For example, in Fig. 2, the first value of F_1 (V_1) and the first value of F_N (V_1) are measured at different time points in the 1st time period. They are collected (merged) and regarded as feature values of the 1st case because they are in the