Journal of Biomedical Informatics 45 (2012) 999–1008

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches

Chang-Sik Son[a], Yoon-Nyun Kim[a,b,c,*], Hyung-Seop Kim[b], Hyoung-Seob Park[b], Min-Soo Kim[c]

[a] Dept. of Medical Informatics, School of Medicine, Keimyung University, Daegu, Republic of Korea
[b] Division of Cardiology, Dept. of Internal Medicine, Keimyung University Dongsan Medical Center, Daegu, Republic of Korea
[c] Biomedical Information Technology Center, School of Medicine, Keimyung University, Daegu, Republic of Korea

ARTICLE INFO

Article history:
Received 14 October 2011
Accepted 26 April 2012
Available online 4 May 2012

Keywords:
Congestive heart failure
Decision-making model
Rough set
Discernibility matrix and function
Maximum entropy principle
Decision tree

ABSTRACT

The accurate diagnosis of heart failure in emergency room patients is quite important, but can also be quite difficult due to our insufficient understanding of the characteristics of heart failure. The purpose of this study is to design a decision-making model that provides critical factors and knowledge associated with congestive heart failure (CHF) using an approach that makes use of rough sets (RSs) and decision trees. Among 72 laboratory findings, it was determined that two subsets (RBC, EOS, Protein, O2SAT, Pro BNP) in an RS-based model, and one subset (Gender, MCHC, Direct bilirubin, and Pro BNP) in a logistic regression (LR)-based model were indispensable factors for differentiating CHF patients from those with dyspnea, and the risk factor Pro BNP was particularly so. To demonstrate the usefulness of the proposed model, we compared the discriminatory power of decision-making models that utilize RS- and LR-based decision models by conducting 10-fold cross-validation. The experimental results showed that the RS-based decision-making model (accuracy: 97.5%, sensitivity: 97.2%, specificity: 97.7%, positive predictive value: 97.2%, negative predictive value: 97.7%, and area under ROC curve: 97.5%) consistently outperformed the LR-based decision-making model (accuracy: 88.7%, sensitivity: 90.1%, specificity: 87.5%, positive predictive value: 85.3%, negative predictive value: 91.7%, and area under ROC curve: 88.8%). In addition, a pairwise comparison of the ROC curves of the two models showed a statistically significant difference ($p < 0.01$; 95% CI: 2.63–14.6).

1. Introduction

The syndrome of heart failure is most commonly defined as a state in which cardiac abnormalities cause cardiac dysfunction such that the heart is unable to meet the circulatory demands of the body, or does so with elevated filling pressures [1]. Given that there is no definitive diagnostic test for heart failure, clinical diagnosis is largely based on a careful history and physical examination that are supported by ancillary tests such as chest radiography, electrocardiogram, and echocardiography. Despite advances related to the complex pathophysiology of heart failure, both its diagnosis and the assessment of therapeutic approaches remain difficult. A timely and accurate diagnosis by a physician is important in order to avoid unnecessary diagnostic procedures and to identify appropriate therapeutic measures and clinical management strategies. However, the search for meaningful sets among critical factors that can affect the early diagnosis of heart failure is difficult, due to the numerous clinical features of routinely available tests, echocardiography, etc.

Data-mining techniques can be applied to overcome effectively these limitations by using large data sets with many predictive factors in order to identify not just linear relationships, but non-linear relationships as well. In particular, the rough set theory (RST) [2] can be used as a tool to discover data dependencies [3–5] and to reduce the number of attributes contained within a data set, using the data alone, and no additional information. In RST, this attribute reduction removes superfluous features and makes it possible to

Abbreviations: SG, specific gravity; OB, occult blood; WBC, white blood cell; RBC, red blood cell; Ep. Cell, epithelial cell; HGB (Hb), hemoglobin; HCT, hematocrit; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; PLT, platelet count; NEUT, neutrophil; LYMP, lymphocyte; MONO, monocyte; EOS, eosinophil; BASO, basophil; LUC, large unstained cell; MPV, mean platelet volume; APTT, activated partial thromboplastin time; PT, prothrombin time; Na, sodium; K, potassium; Cl, chloride; LDH, lactate dehydrogenase; CK, creatine kinase; CK-MB, creatine kinase MB fraction; BUN, blood urea nitrogen; ALP, alkaline phosphatase; AST, aspartate aminotransferase; ALT, alanine aminotransferase; Ca, actual calcium; Mg, magnesium; ABGA, arterial blood gas analysis; BE, base excess; O2CT, oxygen content; O2SAT, oxyhemoglobin saturation; TCO2, total carbon dioxide; CRP, c-reactive protein.

[*] Corresponding author at: Dept. of Medical Informatics, School of Medicine, Keimyung University, 2800 Dalgubeoldaero, Dalseo-Gu, Daegu 704-701, Republic of Korea. Fax: +82 53 580 3745.

E-mail address: ynkim@dsmc.or.kr (Y.-N. Kim).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.04.013

select a feature subset that has the same discernibility as the original set of features. From the medical viewpoint, this approach aims to identify subsets of the most informative attributes that would influence the treatment of patients. Such rule induction methods generate decision rules, which may potentially reveal profound medical knowledge and provide new medical insight. These rules are more useful for medical experts who seek to analyze and gain understanding of the problem at hand [6]. Since this pioneering study was introduced, various related studies have been performed. Bazan [7] compared RST-based methods with statistical methods, neural networks, decision trees, and decision rules using medical data on several pathologies such as lymphography, breast cancer, and primary tumors. He found that the error rates for RS are not only completely comparable, but are also often significantly lower than those obtained using other techniques. Tsumoto [8] investigated the characteristics of such medical reasoning, and showed that the RS representation of diagnostic models is a useful approach for extracting insightful information from medical databases. Carlin et al. [9], who described the application of RS to diagnose suspected acute appendicitis, found that while the difference between a logistic regression (LR) model and RS was quite small, RS offered the advantage of more explicit decision rules. Komorowski and Øhrn [10] discussed the use of an RS framework to identify a patient group in need of a scintigraphic scan for subsequent modeling. They showed that the identification of such patients has the potential to lower the cost of medical care and to improve its quality because, virtually without any loss of information, fewer patients may be referred for this procedure.

These models offer a common advantage, in that their results are directly interpretable, and the decisions obtained from uncertain, incomplete, or approximate data become explainable for unknown phenomena [11]. However, most of these studies have focused on the issues, the discriminatory power of decision models, or the decision rules derived from different algorithms, but without performing a comparison of the significance of their results. In this study, we describe the scheme of a decision-making model based on rough set and decision tree approaches in order to extract the most relevant factors and knowledge from high-dimensional clinical data that typically incur great extra expense and impose an increased workload on clinicians, and we then apply it to the widespread problem of congestive heart failure (CHF).

2. Methods

2.1. Congestive heart failure data

We retrospectively collected the medical records of all patients who went to the emergency medical center of Keimyung University Dongsan Hospital complaining mainly of dyspnea, between July 2006 and June 2007. Only complete medical records with no missing values were included, i.e., demographic characteristics (age and gender), and clinical laboratory findings, such as urinalysis, common blood cell and differential counts, serum electrolytes, routine admission tests, and arterial blood gas analysis. Patients diagnosed with complaints other than CHF were excluded, such as those who presented evidence such as coronary heart disease, including left ventricular (LV) asynergy or a history of previous coronary bypass surgery, significant congenital or valvular disease, or known cardiomyopathy. Eligibility for the study group ($n = 71$) was defined according to the International Classification of Diseases – 10 codes, I50.0. Discharged patients ($n = 88$), i.e., non-cardiogenic dyspnea (Non-CD) patients who were admitted to the emergency medical center complaining mainly of dyspnea, were defined as the control group. All data collected was reconfirmed by three cardiovascular specialists.

2.2. Statistical analysis

Univariate correlations between clinical features were evaluated using the Chi-square test or Fisher's exact test, which are appropriate for categorical variables, and using the Student *t*-test or Mann-Whitney *U*-test with continuous variables, after first checking for normality using the Kolmogorov-Smirnov test. The collected data was expressed as a percentage or mean ± standard deviation. A two-tailed $p < 0.05$ was selected as the level of statistical significance. Following a univariate analysis, a logistic regression (LR) model with Wald's forward feature selection was used for multivariate analysis to identify the independent predictors of CHF, with entry and removal criteria of 0.05 and 0.10 as the default settings. The results are shown as odds ratios (OR) with 95% confidence intervals (95% CI). All statistical analyses were performed using SPSS 12.0 for Windows (SPSS Inc., Chicago, IL, USA).

2.3. Selection of reference intervals

Most of the clinical findings, such as laboratory tests and electrocardiogram results are numerical. For a more accurate discriminative diagnosis, evaluation of therapeutic effects, and prognosis, it is necessary to provide an appropriate reference. In laboratory medicine, even though there has been no definition of normal subjects, a normal value is considered to be observed in subjects under normal conditions. Since clinicians evaluate laboratory data or disease conditions in terms of normal values or ranges, normal values are necessary for the interpretation of numerical laboratory data. Most of the normal values or ranges that are commonly used have been statistically calculated from data obtained from a sample population consisting of individuals who are considered normal, or not abnormal, based upon certain criteria [12]. Generally, the standard limits of normal values are between 2.5 and 97.5 centiles, thus defining a 95% reference interval. Normal ranges are used instead of the reference interval, based upon the logic that values outside the range are abnormal. One flaw in this rationale, however, is that by definition, 5% of normal individuals will have values outside the normal range. There is also possible confusion within the normal distribution; modeling the data under the assumption of normality is a common approach but is not always appropriate for the estimation of reference limits [13]. To address these issues, we describe a method for extracting appropriate reference intervals from clinical laboratory data, using the maximum entropy principle (MEP).

The MEP [14] is an unsupervised learning method for determining the classification boundaries, i.e., cut-off points, in various application areas such as pattern classification [15,16] and image processing [17,18]. If we suppose that X is a discrete random variable and the range $R = \{x_1, x_2, \ldots, x_n\}$ is finite or countable and $p_i = P[X = x_i]$, $i = 1, 2, \ldots, n$, then the Shannon entropy of X is defined as

$$H(X) \cong \sum_{i=1}^{n} p_i \log_a \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_a p_i \tag{1}$$

Eq. (1) defines the entropy $H(X)$ of the random variable X so that it represents the amount of information contained in X, and includes as a measure of uncertainty the stochastic field of X. $H(X)$ represents its function of probability distribution $p_1, p_2, \ldots, p_n$, and is defined as

$$H(X) = H(p_1, p_2, \ldots, p_n) \cong -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

If K is a set of the data point values of a random variable, p_i is the probability of the ith histogram level, N is the number of partitions or subspaces of K. Each subspace is denoted as K_j ($j = 1, 2, \ldots, N$), and $p(K_j)$ is denoted as the probability from the cumulative probability $\sum_{j \in K_j} p_j$ of K_j. A thresholding function is then defined as follows.