COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 119 (2015) 142–162

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Probabilistic modeling of short survivability in patients with brain metastasis from lung cancer

Bunjira Makond[a], *Kung-Jeng Wang*[b], *Kung-Min Wang*[c,*]

[a] *Faculty of Commerce and Management, Prince of Songkla University, Trang, Thailand*
[b] *Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, ROC*
[c] *Department of Surgery, Shin-Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan, ROC*

ARTICLE INFO

Article history:
Received 3 May 2014
Received in revised form
7 February 2015
Accepted 10 February 2015

Keywords:
Bayesian network
Brain metastasis
Lung cancer
Survivability

ABSTRACT

The prediction of substantially short survivability in patients is extremely risky. In this study, we proposed a probabilistic model using Bayesian network (BN) to predict the short survivability of patients with brain metastasis from lung cancer. A nationwide cancer patient database from 1996 to 2010 in Taiwan was used. The cohort consisted of 438 patients with brain metastasis from lung cancer. We utilized synthetic minority over-sampling technique (SMOTE) to solve the imbalanced property embedded in the problem. The proposed BN was compared with three competitive models, namely, naive Bayes (NB), logistic regression (LR), and support vector machine (SVM). Statistical analysis showed that performances of BN, LR, NB, and SVM were statistically the same in terms of all indices with low sensitivity when these models were applied on an imbalanced data set. Results also showed that SMOTE can improve the performance of the four models in terms of sensitivity, while keeping high accuracy and specificity. Further, the proposed BN is more effective as compared with NB, LR, and SVM from two perspectives: the transparency and ability to show the relation of factors affecting brain metastasis from lung cancer; it allows decision makers to find the probability despite incomplete evidence and information; and the sensitivity of the proposed BN is the highest among all standard machine learning methods.

1. Introduction

Physicians generally predict the survivability of patients from the combination of clinical symptoms and signs, as well as from laboratory data, based on their experience and judgment. The prediction precision that depends on the levels of experience and length of patient–physician relationship may vary in each individual [1]. Although regression models have been developed to predict survivability [1–5], these models still have restrictions in applying medical data that are uncertain, complex, and have nonlinear variables with implicit interactions between the variables themselves [6]. Predicting the short survivability of cancer metastases patients (i.e., survival of less than two months) is thus challenging, and attempts for precise prediction were not completely successful.

* *Corresponding author at*: Department of Surgery, Shin-Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan, ROC.
E-mail addresses: bunjira_m@hotmail.com (B. Makond), kjwang@mail.ntust.edu.tw (K.-J. Wang), albert.hua@msa.hinet.net (K.-M. Wang).

http://dx.doi.org/10.1016/j.cmpb.2015.02.005
0169-2607/

Lung cancer often spreads to the brain, given that 65% of patients diagnosed with primary tumor in their lungs have brain metastases [7]. Therefore, modeling and predicting the development of brain metastasis from lung cancer is necessary. Whole brain radiotherapy (WBRT) and corticosteroids are the standard recommended treatments to effectively control symptoms for patients with brain metastases [8] and survivability prognosis is crucial to administer such treatment. A patient with short survivability may not gain any advantage from additional radiation [9–11]. A patient who is expected to have short survivability can be recommended by treatment using steroids alone. Therefore, correct classification of the short survival group of patients with brain metastasis is required to maximize the benefit of the treatment and minimize the suffering of patients. Given the abovementioned properties inherent in lung cancer patients with brain metastases with extremely short survivability, current prognostic models failed to identify patients who will not gain any major benefit from WBRT. The goal of this study is to develop a more reliable survival prediction model.

Traditional statistical models for survivability prediction have rigid assumptions in which the variables are independent [12,13] and difficult to use in calculating for posterior probabilities. By contrast, Bayesian network (BN) is a powerful tool for representing probabilistic events in a simple graphically readable manner and for efficiently predicting tasks. BN can be viewed as a knowledge-representation method with an explicit structure and an associated semantic and reasoning method [14]. Each node can be computed for the posterior probability, which is useful for decision makers. BN can approximate complex multivariate probability distributions of heterogeneous variables as interpretable local probabilities to incorporate prior clinical and biological knowledge as well as to visualize and interpret interactions among variables of interest for clinical use [15]. In addition, BN can be applied in both linear and nonlinear relation problems, including interaction problems, such as a parent–child relationship.

The remainder of this paper is organized as follows: in Section 2, we briefly review the BN along with other standard machine learning methods; in addition, a technique to solve imbalanced problem is stated in this section as well. Variables and data, including model evaluation criteria, are described in Section 3. The experiments and results are presented in Section 4. We provide the conclusions in Section 5.

2. Methods

2.1. Bayesian network

The BN is a probability graphical model that has capability to encode a joint probability distribution over a set of random variables that is either discrete or continuous. Officially, a BN builds a directed-acyclic graph (DAG) by using a set of nodes representing the variables and a set of directed edges representing the relationships between the variables [15]. Given $\mathbf{X} = \{X_1, \ldots, X_n\}$ as the set of random variables, each variable X_i is independent of its non-descendants given its parents in the graph. The joint probability distribution over $\mathbf{X}$ is given by $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|P_a(X_i))$, where $P_a(X_i)$ is the set of parents of X_i.

There are three important steps for the BN construction [12,16]. (i) Identify the set of correlated variables and their possible values, as well as details for variable identification. (ii) Find the network structure by linking nodes that represent variables with arcs with DAG, and constructing the graph using expert knowledge or by applying the algorithm obtained from the data. (iii) Define the conditional probability table (CPT) for each node in the graph.

In this study, the network structure was rigorously examined by domain experts/doctors and constructed using data from related medical literature. After the structure of a BN was known, we quantified the relationship between connected nodes by CPT for discrete variables. We used the Maximum a Posteriori (MAP) estimation method to obtain the values for CPT.

2.1.1. Parameter learning

Given that $_D$ represents a set of random variables $\{X_1, X_2, \ldots, X_n\}$ where X_i is an element of the BN variables, θ represents a set of probability distribution parameter best explain D. The purpose of parameter learning is to find the most probable values for vector θ. The maximum a posterior estimate for the parameter set θ denoted by $\hat{\theta}_{MAP} = arg\max_{\theta} P(\theta|D)$.

From the Bayes' rule, the posterior distribution over θ:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where $P(D|\theta)$ represents the likelihood function $L(\theta{:}D)$, $P(\theta)$ represents the prior over parameters, and $P(D)$ represents a normalizing factor.

The MAP estimation for a binary variable can be explained as follows. Assume D is the set of i.i.d samples $D = \{X_1, \ldots, X_n\}$, X_i is a discrete binary variable with real values 0 and 1(X_i has Bernoulli distribution), then $P(D|\theta) = \theta$, $P(X_i = \theta)$. Thus, $L(\theta{:}D)$ representing the likelihood function for Bernoulli distribution has the form: $P(D|\theta) = L(\theta : D) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{N_1}(1-\theta)^{N_2}$.

A conjugate prior for a binomial is Beta distribution, then $P(\theta) = \gamma\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$, $\gamma = \Gamma(\alpha_1 + \alpha_2)/\Gamma(\alpha_1)\Gamma(\alpha_2)$, where $\Gamma(\cdot)$ is the Gamma function. For any integer x, $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 1$, $\Gamma(x) = (x-1)!$.

From Bayes' rule in Eq. (1), $P(\theta|D) \propto \theta^{N_1}(1-\theta)^{N_2} \cdot \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} \propto \theta^{\alpha_1 + N_1 - 1}(1-\theta)^{\alpha_2 + N_2 - 1}$. Therefore, $\hat{\theta}_{MAP} = (\alpha_1 + N_1 - 1)/(\alpha_1 + \alpha_2 + N_1 + N_2 - 2)$.

The MAP estimation for a multinomial variable can be extended as follows. Assume D is a multinomial experiment consisted of n trials, i.e. $D = \{X_1, \ldots, X_n\}$ where X_i has k possible outcomes; for X_i, the k outcomes can occur with the probability $\theta_1, \ldots, \theta_k$ when $\sum_j^k \theta_j = 1$. Given n_j represents number of the jth outcome, then

$$P(n_1, \ldots, n_k) = \frac{n!}{n_1! \cdots n_k!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k} \quad \text{where} \quad \sum_i n_i = n$$

Therefore $L(\theta{:}D)$ representing the multinomial likelihood function has the form:

$$P(D|\theta) = L(\theta : D) = \prod_j^k \theta_j^{n_j}.$$