Breast cancer prediction model based on logical programming algorithm

Zhao Chunxue
School of Mathematics and Statistics, Anyang Normal University, Anyang, China
Yu Zhaojiang
Human Resources Office, Anyang Normal University, Anyang, China

Abstract
The breast cancer is a malignant tumor occurring in the epithelial tissue of the breast gland, which seriously threatens the health and life of women. In this paper, 698 samples were mined by logical programming algorithm to provide a method for breast cancer prediction.

Keywords
logical programming algorithm; machine learning; breast cancer prediction

INTRODUCTION
A woman’s breast is made up of skin, fibrous tissue, glands and fat. It is a common malignant tumor disease in Chinese women. The risk is higher in women between the ages of 40 and 55. In recent years, there is a trend of getting younger. The incidence is also increasing year by year. The mammary gland is not an important organ for maintaining human life, and in situ breast cancer is not fatal. However, metastasis can occur in the lungs, liver, bone, brain and other organs, and destroy their normal tissue, if not timely treatment can endanger life. Therefore, the early detection and treatment of breast cancer has become an important measure to save lives, which is an important research direction of many researchers. For example, Jonathan et al. [1] built a breast cancer prediction model incorporating familial and personal risk factors. William et al. [2] built a prospective breast cancer risk prediction model for women undergoing screening mammography. Richard et al. [3] studied critical assessment of new risk factors for breast cancer: considerations for development of an improved risk prediction model. Huang et al. [4] carried out the feature selection with support vector machines in breast cancer diagnosis. Catherine et. al. [5] made a systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Anothaisintawee et al. [6] did a systematic review of model performances. Kim et al. [7] did the development of novel breast cancer recurrence prediction model using support vector machine. Lee et al. [8] provided BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Veerle et al. [9] did the validation and modification of a prediction model for acute cardiac events in patients with breast cancer treated with radiotherapy based on three-dimensional dose distributions to cardiac substructures. Candido et al. [10] did an updated PREDICT breast cancer prognostication and built a treatment benefit prediction model with independent validation. Huang et al. [11] studied the SVM and SVM ensembles in the breast cancer prediction.

In this paper, machine learning algorithms (logical programming algorithm) were used to mine 698 breast cancer data sets collected in Wisconsin for breast cancer prediction.

MODELING
We first review the logical programming algorithm. Logistic regression, also known as log-probability regression, is used to deal with regression problems where the dependent variable is a categorical variable. The logical programming algorithm is often used in many practical problems and extensively studied by many scholars[12]-[18]. For m samples and each sample has n characteristics. The steps of the logical programming algorithm are as follows. First, we choose the logistic function \( g(z) = \frac{1}{1 + e^{-z}} \). For the case of linear boundary, the boundary form is as follows:

\[
z = \theta^T x = \theta_0x_0 + \theta_1x_1 + \cdots + \theta_nx_n = \sum_{i=0}^{n} \theta_i x_i
\]

Where, \( x \) is the training data vector, \( \theta \) is the parameter, The predictive function is constructed as follows:
The value of the function $h(x)$ has a special meaning, it represents the probability that the result is 1, so the probability that the input $x$ classification result is category 1 and category 0 is:

$$P(y = 1 | x; \theta) = h_\theta(x)$$
$$P(y = 0 | x; \theta) = 1 - h_\theta(x)$$

The integration form is

$$P(y | x; \theta) = (h_\theta(x))^y(1-h_\theta(x))^{1-y}$$

The likelihood function and logarithm function are respectively as follows:

$$L(\theta) = \prod_{i=1}^{m} P(y_i | x_i; \theta) = \prod_{i=1}^{m} (h_\theta(x_i))^y(1-h_\theta(x_i))^{1-y_i}$$
$$l(\theta) = \log L(\theta) = \sum_{i=1}^{m} (y_i \log h_\theta(x_i) + (1-y_i) \log(1-h_\theta(x_i)))$$

The maximum likelihood estimation is used to find the optimal value of the parameter $\theta$ when the function $l(\theta)$ is maximized.

The optimal value of the parameter $\theta$ is also calculated by the Cost function and the J function. The Cost function and the J function are respectively

$$Cost(h_\theta(x), y) = \begin{cases} 
-\log(h_\theta(x)) & \text{if } y = 1 \\
-\log(1-h_\theta(x)) & \text{if } y = 0
\end{cases}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x_i), y_i) = -\frac{1}{m} \left[ \sum_{i=1}^{m} (y_i \log h_\theta(x_i) + (1-y_i) \log(1-h_\theta(x_i))) \right]$$

i.e.

$$J(\theta) = -\frac{1}{m} l(\theta)$$

The method of gradient descent is used to find the optimal value of the parameter $\theta$ when the function $J(\theta)$ is the minimum value.

In this study, we use the logistic regression to study breast cancer prediction. The used data is a Wisconsin breast cancer data set with a total of 698 fine-needle aspiration biopsy sample units, of which 457 were benign and 241 were malignant. The data set contained 11 variables, i.e., the data set had 11 columns: ID, tumor thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, size of single epithelial cell, bare nucleus, boring chromosome, normal nucleus, mitosis, category.

Firstly, 488 samples were randomly selected from 698 sample units as the training set, and the results of logistic regression for these 488 samples by R language were shown in table 1.

**Table 1. The results of the parameter**

|            | Estimate  | Std. Error | z value | Pr(>|z|) |
|------------|-----------|------------|---------|----------|
| (Intercept)| -7.970e+01| 7.762e+04  | -0.001  | 0.999    |
| $\theta_1$| 1.813e-10 | 8.622e+03  | 0.000   | 1.000    |
| $\theta_2$| -3.110e-10| 1.517e+04  | 0.000   | 1.000    |
| $\theta_3$| -3.499e-10| 1.463e+04  | 0.000   | 1.000    |
In this paper, the logistic regression method was used to detect 698 samples of breast cancer. The results showed that the method was effective and provided a theoretical basis for the detection of breast cancer.

**ACKNOWLEDGMENTS**

The research was supported by National Natural Science Foundation of China(11801012) and National Social Science Foundation of China(18CTJ007).

**REFERENCES**


