Expert Systems with Applications 39 (2012) 11503–11509

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans

Hui Chen [a,*], Jing Zhang [a], Yan Xu [b], Budong Chen [b], Kuan Zhang [a]

[a] School of Biomedical Engineering, Capital Medical University, Beijing 100069, China
[b] Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China

ARTICLE INFO

Keywords:
Artificial neural network
Logistic regression
Lung nodule
Diagnostic performance
Comparison

ABSTRACT

Purpose: To compare the diagnostic performances of artificial neural networks (ANNs) and multivariable logistic regression (LR) analyses for differentiating between malignant and benign lung nodules on computed tomography (CT) scans.

Methods: This study evaluated 135 malignant nodules and 65 benign nodules. For each nodule, morphologic features (size, margins, contour, internal characteristics) on CT images and the patient's age, sex and history of bloody sputum were recorded. Based on 200 bootstrap samples generated from the initial dataset, 200 pairs of ANN and LR models were built and tested. The area under the receiver operating characteristic (ROC) curve, Hosmer–Lemeshow statistic and overall accuracy rate were used for the performance comparison.

Results: ANNs had a higher discriminative performance than LR models (area under the ROC curve: 0.955 ± 0.015 (mean ± standard error) and 0.929 ± 0.017, respectively, $p < 0.05$). The overall accuracy rate for ANNs ($90.0 \pm 2.0\%$) was greater than that for LR models ($86.9 \pm 1.6\%$, $p < 0.05$). The Hosmer–Lemeshow statistic for the ANNs was 8.76 ± 6.59 vs. 6.62 ± 4.03 ($p > 0.05$) for the LR models.

Conclusions: When used to differentiate between malignant and benign lung nodules on CT scans based on both objective and subjective features, ANNs outperformed LR models in both discrimination and clinical usefulness, but did not outperform for the calibration.

1. Introduction

Among all types of cancer, lung cancer is the leading cause of death for both men and women throughout the world. According to data from the American Cancer Society (American Cancer Society, 2010), there were approximately 157,300 lung cancer-related deaths in the United States in 2010, accounting for about 28% of all cancer deaths. Despite advances in treatment and diagnosis of lung cancer, the five-year overall survival rate is currently only 16% but can be improved to 53% if the disease is diagnosed and treated at an early stage (American Cancer Society, 2010). However, only 15% of lung cancers are diagnosed at this early stage, when lesions are bright round-shaped abnormalities, generally referred to as nodules (American Cancer Society, 2010).

Computed tomography (CT) screening has been proven to have strong potential for improving the likelihood of detecting lung cancer at an earlier and more operable stage (The International Early Lung Cancer Action Program Investigators, 2006). Recent advances in multi-slice CT, with spatial resolutions of less than a millimeter, routinely provide images of smaller nodules, making the detection of lung cancer at an earlier stage more possible than ever before (Mulshine & Smith, 2002). However, it is still difficult for radiologists to distinguish between benign and malignant nodules on CT scans, partly owing to the overlapping spectrums of radiographic appearance and clinical presentation (Leef & Klein, 2002; Tsubamoto et al., 2003). It has been reported that as many as 50% of nodules resected at surgery were benign, and only 1–10% of nodules detected by a multi-detector CT were malignant (Stephen et al., 2000; Swensen et al., 2003). In addition, large quantities of image data can be acquired for interpretation by radiologists, increasing the possibility that there will be substantial variability in radiological readings and interpretation (McCulloch et al., 2003). Therefore, a number of investigators are currently developing computer-aided diagnosis (CAD) schemes to facilitate the differentiation between benign and malignant lung nodules on CT images. These CAD schemes play a role of a "second reader" in assisting radiologists in the diagnosis of lung cancer from CT images.

The general approach of CAD for analyzing lung nodules detected on CT scans involves two steps after the nodules have been identified (either designated by radiologists or detected automatically). The first step is to determine the features of the

* Corresponding author.
E-mail address: chenhui@ccmu.edu.cn (H. Chen).

0957-4174/$ - see front matter
http://dx.doi.org/10.1016/j.eswa.2012.04.001

nodules and the patient's information, both of which are used to build the CAD schemes are built. Classification techniques such as linear discriminant analysis (Aoyama et al., 2003; Armato et al., 2003; Mori et al., 2005; Shah et al., 2005a, 2005b; Shiraishi, Abe, Engelmann, Aoyama, & Doi, 2003; Way et al., 2009), support vector machine (Way et al., 2009), logistic regression (LR) analysis (Shah et al., 2005a, 2005b) and artificial neural network (ANN) (Awai, 2006; Lee et al., 2004; Lo, Hsu, Freedman, Lure, & Zhao, 2003; Matsuki et al., 2002; Suzuki, Li, Sone, & Doi, 2005) are then used to develop models for classifying a nodule as either malignant or benign.

Because of the inherent differences in these classification techniques, model selection becomes an important step in the CAD development procedure. Unfortunately, there is no well-accepted theory to guide the model selection based on the complexity and the nature of the diagnostic task (Mangiameli, West, & Rampalc, 2004). As a result, there is an increasing effort to compare the diagnostic performance of different classification models in different CAD schemes for specific medical diagnosis tasks (Behrman, Linder, Assadi, Stacey, & Backonja, 2007; Bourdes et al., 2007; Chen et al., 2009; Lehmann et al., 2007; Moszynski, Szpurek, Smolen, & Sajdak, 2006; Temurtas, Yumusak, & Temurtas, 2009). Among these comparisons, the difference between ANN and LR models is of practical importance to the development of CAD. If the comparative results show limited benefit from using a non-linear ANN model, one should usually turn to the less complicated LR model, which has the advantage of being more readily interpretable and can thus provide more helpful information for making a diagnostic decision.

To our knowledge, no study has compared ANNs and LR models for lung nodule diagnosis using CT images. In this article, we illustrate LR models and ANNs and the application of these models to the classification of lung nodules based on several clinical features and CT signs. In addition, we compare and discuss the performance of the two models from three aspects: calibration, discrimination and clinical usefulness, which are measured by Hosmer–Lemeshow (HL) statistic, the area under the receiver operating characteristic (ROC) curve, and the overall accuracy (ACC), respectively. The unique advantages and disadvantages of the two models may prove complementary in contributing to improved clinical decision making.

2. Materials and methods

2.1. Study database

The database analyzed in this study included 135 malignant and 65 benign solitary pulmonary nodules with diameters not greater than 3 cm. The final diagnosis of malignant nodules was determined by either surgery or biopsy. As well, the diagnosis of benign nodules was confirmed either pathologically or after a 2-year follow-up. The summary of the patients' sex and age can be found in Table 1.

CT scans of 156 patients were obtained with an eight-slice helical CT scanner (Lightspeed Ultra System CT99, General Electric Healthcare, Waukesha, WI, USA), with a slice thickness and reconstruction interval of 7.5 mm. CT scans of the remaining 44 patients were obtained with a single-slice helical CT scanner (PQ 2000, Picker International, Cleveland, OH, USA), with a slice thickness and reconstruction interval of 2 mm. CT images were displayed at a fixed setting (lung window center, −600 HU; lung window width, 1600 HU). Data were reconstructed on a matrix of 512×512 pixels.

2.2. Feature analysis and selection

A lung nodule's morphologic features (size, margins, contour, internal characteristics) and patient's clinical features can be helpful when differentiating benign from malignant nodules (Erasmus, Connolly, & McAdams, 2000; Erasmus, Connolly, McAdams, & Roggli, 2000). In this study, we employed three clinical features and nine radiologic features of lung nodules on CT images. Clinical features included patient age, sex and history of bloody sputum. Radiological features included the nodule's longest and shortest diameters, location, contour definition, spiculation, halo sign, air space, and relationship to blood vessels and pleura. Subjective ratings for all of the radiologic findings except the nodule's diameters, ranging from 1 to 3–5 or Yes–No were determined subjectively. The classification performance strongly depends on the quality of the input data, and the quality of input data using the subjective ratings depends on the ability of radiologists (Matsuki et al., 2002; Nakamura et al., 2000). As such, two attending radiologists, each with at least 20 years' experience in chest radiology and who were unaware of the final diagnosis, quantitatively recorded the subjective radiologic features of all of the lung nodules. Subjective ratings for each radiological finding were independently rated by both radiologists. If they gave a different subjective rating for a radiological finding, the final subjective rating decision was made by negotiation. Descriptions of the final subjective ratings are given in Table 1.

2.3. Logistic regression modeling

The LR model is a kind of generalized linear regression model that is widely used to estimate the probability of a dichotomous outcome event being related to a set of predictors. In this study, a binomial LR model with seven independent variables (one continuous and six categorical in Table 1) corresponding to the seven selected features was developed as $\log it(P) = \beta_0 + \sum_{i=1}^{12} \beta_i x_i$, where P was the probability of malignancy, and β_i ($i = 0, 1 \ldots, 12$) were the unknown parameters that were estimated using a maximum likelihood approach. Using the estimated parameter $\hat{\beta}_i$, the probability of malignancy for a certain sample (nodule) was calculated. The classification threshold was set to 0.5.

2.4. Artificial neural network modeling

ANNs are commonly known as biologically inspired, highly sophisticated analytical techniques, being able to model extremely complex non-linear functions. In this study, a three-layer feed forward ANN with a 12-5-1 nodal architecture was constructed (Fig. 1). The input layer contained twelve neurons for the three clinical features and nine radiologic findings on CT. The hidden layer contained five neurons with the hyperbolic tangent transfer function with a curve slope of 1. The number of hidden neurons was decided experientially, as is generally done in ANN applications. Finally, the output layer had only one neuron with the log-sigmoid transfer function with a curve slope of 1, giving an output value between 0 and 1. The classification threshold was set to 0.5.

The ANNs were trained by using the back propagation algorithm based on the Levenberg–Marquardt rule (Mukherjee & Routroy, 2012; Pham & Sagiroglu, 1995). To prevent non-uniform learning, in which the weights associated with some neurons converged faster than others, all of the continuous inputs were normalized to have zero mean and unit standard deviation. The initial learning rate and the momentum coefficient were set to 0.4 and 0.9, respectively. The training stopped either when the maximum number of iterations (which was set to 100) was reached, or when the performance measured by the mean squared error had increased in six consecutive steps since the last step at which it had decreased.

2.5. Performance comparison

When conducting the performance comparison of different diagnostic models, it is always important to use more than one