Aided diagnosis methods of breast cancer based on machine learning

Yue Zhao, Nian Wang and Xiaoyu Cui
Sino-Dutch Biomedical and Information Engineering School, Northeastern University.
Email: zhaoyue@bmie.neu.edu.cn

Abstract. In the field of medicine, quickly and accurately determining whether the patient is malignant or benign is the key to treatment. In this paper, K-Nearest Neighbor, Linear Discriminant Analysis, Logistic Regression were applied to predict the classification of thyroid, Her-2, PR, ER, Ki67, metastasis and lymph nodes in breast cancer, in order to recognize the benign and malignant breast tumors and achieve the purpose of aided diagnosis of breast cancer. The results showed that the highest classification accuracy of LDA was 88.56%, while the classification effect of KNN and Logistic Regression were better than that of LDA, the best accuracy reached 96.30%.

1. Introduction
Breast cancer, ranking the first in the incidence of malignancy in women, is a major malignant tumors seriously influencing women's physical and mental health and even life-threatening. Breast cancer is the leading cause of death among women between 40 and 55 years of age and is the second overall cause of death among women exceeded only by lung cancer. Fortunately, the mortality rate from breast cancer has decreased in recent years with an increased emphasis on diagnostic techniques and more effective treatments. A key factor in this trend is the early detection and accurate diagnosis of this disease. So CAD systems are designed to support radiologist in the processing of visually screening mammograms to avoid miss-diagnosis. Classification systems can help minimizing possible errors that can be done because of inexperienced experts, and also provide medical data to be examined in shorter time and more detailed.

There have been a lot of research on medical aided diagnosis of breast cancer, and most of them have reported high classification accuracy. Quinlan reached 94.74% classification accuracy using 10-fold cross validation with C4.5 decision tree method [1]. The researcher, H. J. Hamiton, and N. Shan, reached 94.99% accuracy with RIAC technique [2]. While B. Ster and A. Dobnikar have got to 96.8% with linear discreet analysis method [3]. Using neuro-fuzzy techniques, the accuracy of method proposed by D. Nauck and R. Kruse was 95.06% [4]. In Abonyi and Szeifert, an accuracy of 95.57% was obtained with the application of supervised fuzzy clustering technique [5]. The fuzzy-GA method was introduced and a classification accuracy of 97.36% was achieved [6]. Three different methods, optimized learning vector quantization (LVQ), big LVQ, and artificial immune recognition system (AIRS) were applied and the obtained accuracies were 96.7%, 96.8%, and 97.2% respectively [7]. Multilayer perceptron neural network, four different methods, combined neural network, probabilistic neural network, recurrent neural network and SVM were used respectively by U beyli; and the highest classification accuracy of 97.36% was achieved by SVM [8]. Two different methods, Bayesian classifiers and artificial neural networks were applied and the obtained accuracies were 92.80% and 97.90%, respectively [9]. Murat Karabatak and M. Cevdet Ince, using the method combined with association rules and neural network, obtained the accuracy of 95.60% [10].
In the paper, K-Nearest Neighbor, Linear Discriminant Analysis, Logistic Regression combination with normalization, principal component analysis, 5-fold cross validation are applied to predict the classification of thyroid, Her-2, PR, ER, Ki67, metastasis and lymph nodes in breast cancer, in order to recognize the benign and malignant breast tumors and achieve the purpose of auxiliary diagnosis of breast cancer.

2. Basic Theories

2.1. K-Nearest Neighbor
K-Nearest Neighbor, abbreviated to KNN, is a simple and effective non-parametric classifier in the field of pattern recognition. The core idea of K-Nearest Neighbor is that there is a sample in the feature space, if the majority of the samples belong to a certain category, the sample also belongs to the same category. In the KNN algorithm, the similarity between objects is usually calculated by the distance between objects, generally using Euclidean distance or Manhattan distance, showed as in equation (1) and equation (2), respectively.

\[
d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}
\]

\[
d(x, y) = \sqrt{\sum_{k=1}^{n} |x_k - y_k|}
\]

2.2. Linear Discriminant Analysis
Linear Discriminant Analysis, abbreviated to LDA, is one of the commonly used classification methods in statistics. The basic theory of the algorithm is to project a high-dimensional pattern of samples into the best discriminant vector space to guarantee model samples have the largest inter-class distance and the minimum inner-class distance.

2.3. Logistic Regression
Logistic Regression is actually a classification method for two classification problems. Its basic flow is to find a suitable predictor function firstly, generally expressed as h function, which is a classification function, used to predict the input data to determine the results, as in equation (3).

\[
h(x) = Q_0 + Q_1 x_1 + Q_2 x_2 + \cdots + Q_m x_m = Q^T X
\]

And constructing a Cost function which represents the deviation between the predicted output (h) and the training data class (y), taking the "loss" of all training data into consideration, and summing or averaging Cost as J (θ) function, which indicates the deviation of the predicted value of all training data from the actual category, as in equation (4). Obviously, the smaller the value of the J (θ) function is, the more accurate the predictor function is. So this step needs to find the minimum value of the J (θ).

\[
J(Q) = \frac{1}{2} \sum_{i=1}^{n} (h_Q(x^{(i)}) - y^{(i)})^2
\]

3. Methodology and Experiment

3.1. Experimental Data
In this experiment, the data was taken from the First Affiliated Hospital of China Medical University, consisting of 77 samples. And every sample had 13 attributes and 7 classes corresponding to each record. These 7 classes represented 7 pathological indicators. Among them, Her-2, PR, ER, Ki67 these 4 indicators were important clinical data to guide doctors to make comprehensive treatment. They
were also very important to decide the patient's treatment schedule. In the experiment, the 77 samples were all used to train to construct a classification model, and then randomly choose 20 samples for test.

3.2. Data Processing

Data normalization processing is a basic step of data mining. Different evaluation indicators often have different dimension units, which will affect the results of data analysis. In order to eliminate the impact of indicators, data must be standardized. After the raw data is normalized, the indicators are in the same order of magnitude, suitable for comprehensive evaluation. In this paper, the equation of data normalization processing is defined as in equation (5).

\[ X_{\text{norm}} = \frac{2*(X - X_{\text{min}})}{X_{\text{max}} - X_{\text{min}} - 1} \]  

(5)

Principal component analysis is also an important issue in building classification systems. It refers to the export of a few principal components from the original variable so that they retain as much information of the original variable as possible and have as much good prediction as possible.

K-fold cross validation is shown to be efficient for small data sample and gives more reliable results than common test sample estimates. In the processing, the whole set is divided randomly into k nearly equally large subsets. The k-1 of subsets will be used to train, and the last will be used to test, which guarantee that each of subsets is involved exactly once in testing. K times results are obtained and then averaged as the accuracy of the algorithm.

3.3. Experimental Scheme

In the experiment, the 77 samples were all used to train to construct a classification model, and then randomly choose 20 samples for test. Normalization, principal component analysis, 5-fold cross validation were applied to processing the original data respectively. Based on the processed data and raw data, KNN, LDA, Logistic Regression three methods are applied to predict the classification of thyroid, Her-2, PR, ER, Ki67, metastasis and lymph nodes in breast cancer, respectively.

The performance of the classifier was evaluated using Sen, Spe and Accuracy. Accuracy indicated the percentage of all samples identified correctly. They could be written as in equations (6).

\[ \text{Sen}(\%) = \frac{TP}{TP + FN} \times 100 \]
\[ \text{Spe}(\%) = \frac{TN}{TN + FP} \times 100 \]
\[ \text{Accuracy}(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \]  

(6)

3.4. Results and Discussion

Based on the original features, KNN, LDA, Logistic Regression three methods were used to obtain the values of Sensitivity, Specificity and Classification accuracy in Table 1.

From the table 1, we could see clearly that in the method of logistic regression, metastasis, thyroid, ki-67 had relatively high accuracy, respectively 92.60%, 88.89%, 85.19%. And in LDA, the classification accuracy of these seven indicators were all relatively lower. However, in the algorithm of KNN, the classification accuracy of metastasis reached the highest, about 96.30%. Considering the results above, this showed that for the same data set, the classification effect varied from classifier to classifier. So in order to get a good classification effect, the choice of classifier was the key.

In addition, the sensitivity of thyroid, Her-2, metastasis in KNN, Her-2, metastasis in logistic regression and the specificity of ER, Ki-67 in KNN and logistic regression were low, and even some was zero. After analyzing, that may be due to the reason that the sample data set was not large enough, and positive and negative samples were uneven distribution, so the learning and master of characteristics was not enough. Therefore, in the next time, the sample data would be collected further.
Table 1. The classification results of KNN, LDA, Logistic Regression based on raw features

| Methods          | indicators | TP   | TN   | FP   | FN   | Sen   | Spe   | Accuracy |
|------------------|------------|------|------|------|------|-------|-------|----------|
| KNN              | thyroid    | 1    | 23   | 1    | 3    | 25.00 | 1     | 88.89    |
|                  | Her-2      | 2    | 17   | 3    | 5    | 28.57 | 5     | 85.00    |
|                  | PR         | 16   | 6    | 5    | 0    | 1     | 54.55 | 81.48    |
|                  | ER         | 18   | 2    | 5    | 2    | 90.00 | 28.57 | 74.07    |
|                  | ki-67      | 22   | 0    | 4    | 1    | 95.65 | 0     | 81.48    |
|                  | metastasis | 0    | 26   | 0    | 1    | 0     | 1     | 96.30    |
|                  | lymph nodes| 14   | 9    | 1    | 3    | 82.35 | 90.00 | 85.19    |
|                  | thyroid    | 3    | 20   | 1    | 3    | 75.00 | 86.96 | 71.24    |
|                  | Her-2      | 5    | 16   | 4    | 2    | 71.43 | 80.00 | 70.83    |
|                  | PR         | 12   | 8    | 3    | 4    | 75.00 | 72.73 | 79.08    |
| LDA              | ER         | 15   | 6    | 1    | 5    | 75.00 | 85.71 | 71.84    |
|                  | ki-67      | 12   | 3    | 1    | 11   | 52.17 | 75.00 | 65.48    |
|                  | metastasis | 1    | 23   | 3    | 0    | 1     | 88.46 | 87.54    |
|                  | lymph nodes| 11   | 8    | 2    | 6    | 64.71 | 80.00 | 74.23    |
|                  | thyroid    | 2    | 22   | 1    | 2    | 50.00 | 95.65 | 88.89    |
|                  | Her-2      | 2    | 18   | 2    | 5    | 28.57 | 90.00 | 74.07    |
|                  | PR         | 13   | 6    | 5    | 3    | 81.25 | 54.55 | 70.37    |
|                  | ER         | 19   | 3    | 4    | 1    | 95.00 | 42.86 | 81.48    |
| Logistic         | ki-67      | 23   | 0    | 4    | 0    | 1     | 0     | 85.19    |
| Regression       | metastasis | 0    | 25   | 1    | 1    | 0     | 96.15 | 92.60    |
|                  | lymph nodes| 12   | 6    | 4    | 5    | 70.59 | 60.00 | 66.67    |

Table 2. The classification accuracy of the different processing methods about feature

| Methods          | indicators | raw   | normalization | PCA | 5-fold cross validation | Normalization, PCA, 5-fold cross validation |
|------------------|------------|-------|---------------|-----|-------------------------|------------------------------------------|
| KNN              | thyroid    | 88.89 | 96.30         | 88.89 | 79.17                    | 81.92                                |
|                  | Her-2      | 70.37 | 70.37         | 70.37 | 68.92                    | 69.08                                |
|                  | PR         | 81.48 | 85.19         | 81.48 | 62.42                    | 66.50                                |
|                  | ER         | 74.07 | 81.48         | 74.07 | 68.92                    | 78.08                                |
|                  | ki-67      | 81.48 | 85.19         | 81.48 | 77.75                    | 79.33                                |
|                  | metastasis | 96.30 | 92.59         | 96.30 | 83.08                    | 83.17                                |
|                  | lymph nodes| 85.19 | 81.48         | 85.19 | 68.67                    | 74.17                                |
|                  | thyroid    | 71.24 | 71.24         | 55.81 | 79.47                    | 70.32                                |
|                  | Her-2      | 70.83 | 70.83         | 64.58 | 76.85                    | 68.45                                |
|                  | PR         | 79.08 | 79.08         | 64.03 | 80.13                    | 73.16                                |
| LDA              | ER         | 71.84 | 71.84         | 58.68 | 77.70                    | 66.97                                |
|                  | ki-67      | 65.48 | 65.48         | 55.95 | 71.74                    | 66.87                                |
|                  | metastasis | 87.54 | 87.54         | 64.85 | 88.56                    | 79.46                                |
|                  | lymph nodes| 74.23 | 74.23         | 70.92 | 79.09                    | 76.61                                |
|                  | thyroid    | 88.89 | 88.89         | 85.19 | 76.67                    | 79.33                                |
|                  | Her-2      | 74.07 | 74.07         | 74.07 | 76.67                    | 75.59                                |
|                  | PR         | 70.37 | 70.37         | 66.67 | 72.83                    | 70.17                                |
| Logistic         | ER         | 81.48 | 81.48         | 74.07 | 71.67                    | 75.42                                |
| Regression       | ki-67      | 85.19 | 85.19         | 85.19 | 77.67                    | 80.75                                |
|                  | metastasis | 92.60 | 92.60         | 96.30 | 84.58                    | 88.25                                |
|                  | lymph nodes| 66.67 | 66.67         | 62.96 | 69.00                    | 79.17                                |
After the data processing of normalization, principal component analysis, 5-fold cross validation, respectively, KNN, LDA, Logistic Regression three methods were used to obtain the values of classification accuracy again. We would compare the results with the previous results in Table 2.

As can be seen from Table 2, in the method of KNN, with the data normalization the results of thyroid, PR, ER were higher than the original results. When using the method of LDA, the classification accuracy of all indicators were slightly improved comparing 5-fold cross validation with original results. However, selecting the algorithm of logistic regression, the original features without any processing would achieve the highest accuracy. Experimental results showed that for the same data set, the classification effect was different with the input characteristics processed differently. So in order to get a better classification effect, the choice of way of data processing was the key.

Seen from Table 1 and Table 2, KNN classification efficiency was better, the highest accuracy rate of 96.30%, while in LDA and SVM, the highest accuracy rate was 92.60%, 85.19%. So for different classification samples, to choose a suitable classifier to get a higher accuracy was essential to achieve the purpose of aided diagnosis of breast cancer.

4. Conclusion

Diagnosis of breast cancer requires the comprehensive use of clinical, imaging, pathology, computer-aided diagnosis and other knowledge and technical means. Early detection and early diagnosis of breast cancer are the key to improve curative effect. In this paper, KNN, LDA and logistic regression three methods were used to the identification of breast cancer pathology, to help doctors comprehensive assessment to make specific treatment options and drug selection. Experiment results showed that KNN classification accuracy, better than LDA and logistic regression, was expected to become an effective and practical computer assisted diagnostic tools in breast cancer that could reduce the human subjective misdiagnosis rate and bring the gospel for the breast cancer patients.

Acknowledgments

The authors are grateful for the financial support from the National Natural Science Foundation of China (61501101, 61605025), the Natural Science Foundation of Liaoning Province (201601015).

5. References

[1] Quinlan J R Improved use of continuous attributes in C4.5 1996 Journal of Artificial Intelligence Research 4 pp 77–90
[2] Hamilton H J, Shan N, and Cercone N RIAC: A rule induction algorithm based on approximate classification 1996 Computer Science Department
[3] Ster B and Dobnikar A Neural networks in medical diagnosis: Comparison with other methods 1996 In Proceedings of the international conference on engineering applications of neural networks pp 427–30
[4] Nauck D and Kruse R Obtaining interpretable fuzzy classification rules from medical data 1999 Artificial Intelligence in Medicine 16 pp 149–69
[5] Abonyi J and Szefert F Supervised fuzzy clustering for the identification of fuzzy classifiers 2003 Pattern Recognition Letters 24 pp 2195–2207
[6] Pena-Reyes C A and Sipper M A fuzzy-genetic approach to breast cancer diagnosis 1999 Artificial intelligence in medicine 17 pp 131–55
[7] Goodman D E, Boggess L and Watkins A Artificial immune system classification of multiple-class problems 2002 In Proceedings of the Artificial Neural Networks in Engineering ANNIE 2 pp 179–83
[8] Ubezii E D Implementing automated diagnostic systems for breast cancer detection 2007 Expert Systems with Applications 33 pp 1054–62
[9] Maglogiannis I, Zafiropoulos E and Anagnostopoulos I An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers 2009 Applied Intelligence 30 pp 24–36
[10] Karabatak M and Cevdet Ince M An expert system for detection of breast cancer based on association rules and neural network 2009 Expert Systems with Applications 36 pp 3465–69
[11] Polat K and Gunes S Breast cancer diagnosis using least square support vector machine 2007 Digital Signal Processing 17 pp 694–701
[12] Salama G I, Abdelhalim M B and Zeid M A Breat Cancer diagnosis on three different datasets using multi-classifiers 2012 *International Journal of Computer and Information Technology* 1 pp 36-43