Indonesian Journal of Electrical Engineering and Computer Science
Vol. 13, No. 1, January 2019, pp. 405~410
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i1.pp405-410

Classification enhancement of breast cancer histopathological image using penalized logistic regression

Mohammed Abdulrazaq Kahya
Department of Computer science, Education College for Pure Science, University of Mosul, Mosul, Iraq

Article Info

Article history:

Received Jun 19, 2018
Revised Aug 21, 2018
Accepted Nov 18, 2018

Keywords:

Breast cancer
Histopathological image
L1-norm
Penalized logistic regression
Smoothing

ABSTRACT

Classification of breast cancer histopathological images plays a significant role in computer-aided diagnosis system. Features matrix was extracted in order to classify those images and they may contain outlier values adversely that affect the classification performance. Smoothing of features matrix has been proved to be an effective way to improve the classification result via eliminating of outlier values. In this paper, an adaptive penalized logistic regression is proposed, with the aim of smoothing features and provides high classification accuracy of histopathological images, by combining the penalized logistic regression with the smoothed features matrix. Experimental results based on a publicly recent breast cancer histopathological image datasets show that the proposed method significantly outperforms penalized logistic regression in terms of classification accuracy and area under the curve. Thus, the proposed method can be useful for histopathological images classification and other classification of diseases types using DNA gene expression data in the real clinical practice.

Corresponding Author:

Mohammed Abdulrazaq Kahya,
Department of Computer science,
Education College for Pure Science,
University of Mosul, Mosul, Iraq.
Email: mohammedkahya@uomosul.edu.iq

1. INTRODUCTION

Nowadays, cancer is the second leading cause of death worldwide. On the other hand, the World Health Organization (WHO) confirmed that 8.2 million deaths were caused by cancer in 2012 and 8.8 million in 2015. Moreover, it expected 27 million of new cases of this disease before 2030 [1]. In particular, breast cancer is one of the leading causes of women's death in the world. A recent study confirmed that breast cancer accounts for 18% of all types of women cancers and the fifth reason of death in the worldwide [2].

However, the early stage diagnosis and therapy can increase the survival rates to 98% [3]. There are many noninvasive imaging techniques for breast cancer such as magnetic resonance imaging (MRI), mammograms (X-rays), ultrasonography and histopathological image [4-7]. Diagnosis using histological images has become a powerful gold standard for deadly diseases such as breast and lung cancers, which gives a satisfactory diagnosis compared with other methods such as mammography and ultrasonography [8].

On the other hand, machine learning techniques have been used to enhance the diagnostic accuracy for breast cancer through a computer-assisted system [9]. In general, breast cancer is classified into benign and malignant types and this diagnosis is very important in drug discovery and treatment [10-11].

Logistic regression (LR) is considered one of the famous machine learning techniques of classification such as support vector machines (SVM), random forests (RF), and neural networks (NNet) [12]. Logistic regression is an extensive classification technique and has many applied fields like gene expression data [13], prediction of therapy outcome [14] and protein function [15].

The classification performance improvement is the core of the breast cancer histopathological image classification to increase the diagnostic accuracy through the features selection process [16-17], pre-processed image [18-19], or any other techniques. However, the proposed method differs from previous techniques in the preprocessing action of the features matrix which aims to eliminating the outlier values in these features to increase the classification accuracy through smoothing features matrix data process.

2. THE PROPOSED METHOD

2.1. Penalized Logistic Regression

Logistic regression is one of the powerful classification algorithms that is comparatively easy and robust for classification between two classes. In this paper, logistic regression technique was used to illustrate the relationship between independent variables (breast cancer histopathological image features) and the variable of response (1 for the benign class or 0 for the malignant class).

Let we have n independent observations $\{y_i, x_i; i = 1, 2, ..., n\}$ where $y_i \in \{1, 0\}$ are response variables, and $x_i = \left(x_{i1}, ..., x_{ip}\right)^T$ is a vector of image features. Consequently, the logistic regression model is explained as

$$\text{Prob}(y_i = 1 : x_i) = \pi(x_i), \tag{1}$$

$$\text{Prob}(y_i = 0 : x_i) = 1 - \pi(x_i), \tag{2}$$

This probability can be explained as follows:

$$\pi(x_i) = \frac{\exp\left(\alpha + x_i^T\beta\right)}{1 + \exp\left(\alpha + x_i^T\beta\right)}, \qquad \log\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] = \alpha + x_i^T\beta, \tag{3}$$

where $\alpha + x_i^T\beta = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$.

The log-likelihood function for response variables y_i can be written as:

$$l(\beta) = \sum_{i=1}^{n}\left[y_i\left(\alpha + x_i^T\beta\right) - \log\left\{1 + \exp\left(\alpha + x_i^T\beta\right)\right\}\right], \tag{4}$$

The penalized logistic regression model (PLR) adds a nonnegative penalty term to Equation (4), and is defined as follows:

$$l(\beta) = \sum_{i=1}^{n}\left[-y_i\left(\alpha + x_i^T\beta\right) + \log\left\{1 + \exp\left(\alpha + x_i^T\beta\right)\right\} + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right], \tag{5}$$

Minimizing the PLR function gives us the parameters α and β [20-21].

2.2. Histopathological Images Features Extraction

In this paper, discrete wavelet transform was used to decompose histopathological images of breast cancer [22]. Precisely, each image was decomposed to level VII based on Haar discrete wavelet transform to extract the features [23]. Level I of image decomposition gives four equal size of sub-images, namely A1 (approximation coefficients), H1 (horizontal coefficients), V1 (vertical coefficient) and D1 (diagonal coefficient). Then, the next level decomposition is based only on the previous A of the previous decomposition. Therefore, level II of image decomposition gives another four equal size of sub-images, namely A2, H2, V2 and D2 result of decomposition A1. The decomposition continues until reaching the level VII. Thus, twenty-eight sub-images are decomposed. Next, three new sub-images are generated from the color channels (red, green, blue) of each sub-image. Thus, the original image is decomposed to the 28 x 3 from sub-images. Then, nine of the traditional statistical standards (mean, mean absolute deviation, median absolute deviation, standard

deviation, entropy, energy, skewness, kurtosis, root mean square) are extracted from every sub-image. As a result, 756 features have been obtained from each histopathological image.

2.3. Smoothing Data of Features Matrix

Suppose a sequence of data points (i.e.: a feature in pattern recognition, a gene in gene expression data or a variable in statistics) are given which represent the characteristic features for kinds of classes. This data is often contained outlier (extreme) values as noise for the classification problem in data mining field. To reduce this noise data (Rough data), we can consider the sequence of data points as a discrete signal in time domain using the digital filters in signal processing.

Digital filters techniques are used to extract beneficial parts of the signal or to clear out unwelcome parts of the signal [24-26]. Figure 1 clarifies the basic idea of the filter.

Figure 1. Filter Block Diagram

In general, there are several digital filters techniques to smooth data such as moving average, local regression (lowess and loess), and robust local regression (rlowess and rloess) and Savitzky-Golay [24-28]. This paper uses the moving average technique which considered as the most common digital filter in signal processing to ease the calculation and understanding. In a simple way the work of moving average technique can be summarized as follow, if we have an array of raw (Rough) data $[x(1), x(2), ..., x(N)]$, it can be refined to a new array of smoothed data $[\tilde{x}(1), \tilde{x}(2), ..., \tilde{x}(N)]$. The smoothed point $\tilde{x}(k)$ equal the average number of an odd neighbor points for the current point. The following formula represents the equation of the moving average filter.

$$\tilde{x}(k) = \frac{1}{2m+1} \sum_{i=-m}^{m} x(k+i) \qquad , m = 1, 2, 3, ... \tag{6}$$

The odd number $2m+1$ is always named filter span. Subsequently, the smoothed features matrix data (Figure 2) is used for classification problem.

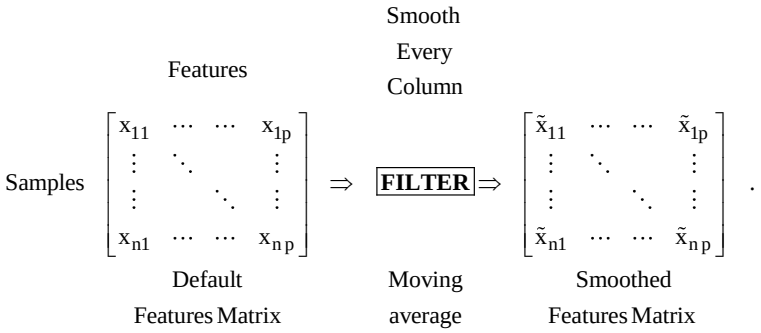

Figure 2. Features Matrix Data Smoothing Process

2.4. Breast Cancer Classification

After the preprocessing for the features matrix, the PLR with smooth features matrix is utilized to get high classification accuracy. The detailed of the Adaptive PLR (APLR) computation is described in Algorithm 1.

Algorithm 1: The computation of APLR

Step 1: Extract features matrix via wavelet transform.
Step 2: Smooth features matrix via moving average technique (Figure 2).
Step 3: Solve the APLR,

$$\hat{\beta}_{ALR} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left[\log\left\{1+\exp\left(\alpha+\tilde{x}_i^T\beta\right)\right\} - y_i\left(\alpha+\tilde{x}_i^T\beta\right)\right] + \lambda\sum_{j=1}^{p}|\beta_j| \right\}. \quad (7)$$

3. RESEARCH METHOD

3.1. Datasets Description

The database that has been used is the BreaKHis (The Breast Cancer Histopathological Images), BreaKHis database supplies us with 7,909 of microscopic biopsy images which have included two types of benign and malignant tumors that had collected from 82 patients using different magnifying factors: 40X, 100X, 200X, and 400X [4]. The available histopathological images of true colors in Portable Network Graphics (PNG) format with 700 × 460 pixels' resolution are the raw images of neither normalization nor color standardization. These images are acquired in RGB channels. A summary of this database is listed in Table 1 and samples of these images in Figure 3.

Table 1. Summary of the BreaKHis database

Magnification	Benign	Malignant	Total
40X	625	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
Total	2480	5429	7909
# Patients	24	58	82

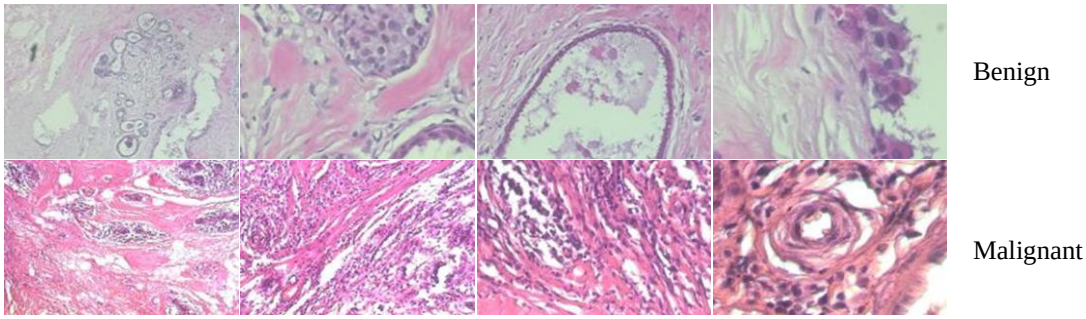

Figure 3. Samples Breast Cancer Histopathological Images

3.2. Performance Evaluation

In order to evaluate the proposed method, two performance metrics were used: the patient classification rate (PCR) and the overall classification accuracy (OCA)[4]. The standard PCR is the rate of the number of images classified correctly to the number of all images for each patient, the PCR can explain as:

$$PCR = \frac{n_{correct}}{n_{all}} x100\%, \quad (8)$$

where $n_{correct}$ is the number of histopathological images classified correctly for the patient i and n_{all} is the number of histopathological images of the patient i .

The OCA can be explained as:

$$OCA = \frac{\sum_{i=1}^{n_{patients}} PCR_i}{n_{patients}}, \quad (9)$$

where $n_{patients}$ is the number of patients.

3.3. Experimental Setting

To confirm the usefulness of the proposed method, comprehensive experiment with LR is conducted. To do so, the smooth features matrix data is partitioned into the training set and the test set, where 70% of the samples are selected randomly for the training set and the rest 30% are selected for testing set. To mitigate the effects of the features matrix data partition, all the results obtained were the average of five trials for partitions.

4. RESULTS AND DISCUSSION

4.1. Classification Performance

Table 2 reports, on average, the OCA for the training and testing datasets of applying the APLR and PLR methods. The number in parenthesis is the corresponding standard deviation. In addition, the last column represents the filter span value.

At the beginning with the magnification 40X, regarding the overall classification accuracy and based on the training dataset, the proposed method, APLR, achieves 100.00%, defeating PLR, by 3.098% whether the filter span value equal five or three. Depending on the testing dataset, the APLR which depends on the filter span five is better than PLR of overall classification accuracy because it achieved 91.922 %, which is 6.962 better than PLR.

While the magnification 100X, the APLR also provides enhancement over the PLR by 6.608% for the training dataset regardless of the filter span value. Moreover, the proposed method beats PLR in terms of overall classification accuracy based on the testing dataset.

Looking at the magnification 200X, the OCA of the proposed method performance is better than the non-smoothed data of PLR. In terms of OCA, the OCA obtained from the proposed method was 100.00% for the training dataset and 92.46% that depends on the filter span five as well 93.496% that depends on the filter span three for the testing dataset. This indicates the superiority of the proposed method. Eventually, regarding the magnification 400X, the APLR shows a considerable dominance against non-smoothed data PLR. It achieved the higher overall classification accuracy for both the training and testing datasets.

Table 2. Classification performance of the APLR and PLR

Methods	Training dataset OCA %	Testing dataset OCA %	Filter Span
40X			
APLR	100.00 (0.000)	**91.922 (4.413)**	5
APLR	100.00 (0.000)	91.544 (4.830)	3
PLR	96.902 (0.949)	84.960 (4.602)	Non Smoothed
100X			
APLR	100.00 (0.000)	**92.822 (1.431)**	5
APLR	100.00 (0.000)	90.996 (0.898)	3
PLR	96.018 (1.010)	86.214 (0.544)	Non Smoothed
200X			
APLR	100.00 (0.000)	92.460 (3.169)	5
APLR	100.00 (0.000)	**93.496 (2.532)**	3
PLR	96.862 (1.147)	86.858 (2.323)	Non Smoothed
400X			
APLR	100.00 (0.000)	**88.364 (1.985)**	5
APLR	100.00 (0.000)	87.246 (2.076)	3
PLR	94.870 (1.035)	82.572 (1.919)	Non Smoothed

4.2. Statistical Significance Test

To confirm the utility of the proposed method in high classification performance, a pairwise comparison between the proposed method and each competitor method was used using Mann–Whitney U test. The area under the curve (AUC) for the training dataset was used for this test. Table 3 shows the Mann–Whitney U test results at significance level $\alpha = 0.05$. As highlighted in Table 3, the AUC of the proposed method is statistically significantly better than PLR.

Table 3. P-values for the Mann–Whitney U test of the proposed method results with competitor method. (*) means that the two methods have significant differences

Dataset	APLR vs PLR
40X	0.0068(*)
100X	0.0054(*)
200X	0.0060(*)
400X	0.0011(*)

5. CONCLUSION

This paper presents an adaptive penalized logistic regression by means of smoothing of features matrix to increase overall classification accuracy of breast cancer histopathological images. The superior classification performance of the proposed method was shown through two aspects: high overall classification accuracy and the Mann–Whitney U test for the AUC. Consequently, the results confirm that APLR is a promising method for medical image classification, medical diagnosis of tumors and very useful in other types of high-dimensional classification data related to the biological field.

REFERENCES

[1] Boyle, P. and B. Levin, *World cancer report 2008.* 2008: IARC Press, International Agency for Research on Cancer.
[2] Zhang, Y., B. Zhang, and W. Lu, *Breast cancer histological image classification with multiple features and random subspace classifier ensemble*, in *Knowledge-Based Systems in Biomedicine and Computational Life Science*. 2013, Springer. p. 27-42.
[3] Center, M., R. Siegel, and A. Jemal, *Global cancer facts & figures.* Atlanta: American Cancer Society, 2011: p. 1-52.
[4] Spanhol, F.A., et al., *A dataset for breast cancer histopathological image classification.* IEEE Transactions on Biomedical Engineering, 2016. 63(7): p. 1455-1462.
[5] Vu, T.H., et al., *Histopathological image classification using discriminative feature-oriented dictionary learning.* IEEE Transactions on Medical Imaging, 2016. 35(3): p. 738-751.
[6] Acharya, U.R., et al., *Automated characterization of fatty liver disease and cirrhosis using curvelet transform and entropy features extracted from ultrasound images.* Computers in Biology and Medicine, 2016. 79: p. 250-258.
[7] Uyun, S. and L. Choridah, *Feature Selection Mammogram based on Breast Cancer Mining.* International Journal of Electrical and Computer Engineering (IJECE), 2018. 8(1): p. 60-69.
[8] Huang, C.-H., et al., *Time-efficient sparse analysis of histopathological whole slide images.* Computerized medical imaging and graphics, 2011. 35(7-8): p. 579-591.
[9] Belsare, A.D., et al. *Classification of breast cancer histopathology images using texture feature analysis*. in *TENCON 2015 - 2015 IEEE Region 10 Conference*. 2015.
[10] Korkmaz, S., G. Zararsiz, and D. Goksuluk, *Drug/nondrug classification using support vector machines with various feature selection strategies.* Computer Methods and Programs in Biomedicine, 2014. 117(2): p. 51-60.
[11] Zhang, L., et al., *Similarity-balanced discriminant neighbor embedding and its application to cancer classification based on gene expression data.* Computers in biology and medicine, 2015. 64: p. 236-245.
[12] Abu-Nimeh, S., et al. *A comparison of machine learning techniques for phishing detection*. in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. 2007. ACM.
[13] Algamal, Z., *Classification of gene expression autism data based on adaptive penalized logistic regression.* Electronic Journal of Applied Statistical Analysis, 2017. 10(2): p. 561-571.
[14] Park, H., et al., *A novel adaptive penalized logistic regression for uncovering biomarker associated with anti-cancer drug sensitivity.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016.
[15] Lee, H., et al., *Diffusion kernel-based logistic regression models for protein function prediction.* Omics: a journal of integrative biology, 2006. 10(1): p. 40-55.
[16] Kahya, M.A., W. Al-Hayani, and Z.Y. Algamal, *Classification of breast cancer histopathology images based on adaptive sparse support vector machine.* Journal of Applied Mathematics and Bioinformatics, 2017. 7(1): p. 49.
[17] Hlaing, T., *Feature selection and fuzzy decision tree for network intrusion detection.* International Journal of Informatics and Communication Technology (IJ-ICT), 2012. 1(2): p. 109-118.
[18] Spanhol, F.A., et al. *Breast cancer histopathological image classification using Convolutional Neural Networks*. in *International Joint Conference on Neural Networks (IJCNN)*. 2016. Vancouver, Canada: IEEE.
[19] Motlagh, N.H., et al., *Breast Cancer Histopathological Image Classification: A Deep Learning Approach.* bioRxiv, 2018: p. 242818.
[20] Algamal, Z.Y. and M.H. Lee, *A novel molecular descriptor selection method in qsar classification model based on weighted penalized logistic regression.* Journal of Chemometrics, 2017. 31(10).
[21] Algamal, Z., *An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression.* Electronic Journal of Applied Statistical Analysis, 2017. 10(1): p. 242-256.
[22] Gonzalez, R. and R. Woods, *Digital Image Processing*. 2008, Upper Saddle River, New Jersey, USA: Pearson Prentice Hall.
[23] El Ayachi, R., B. Bouikhalene, and M. Fakir, *New Image Compression Algorithm using Haar Wavelet Transform.* International Journal of Informatics and Communication Technology (IJ-ICT), 2017. 6(1): p. 43-48.
[24] Antoniou, A., *Digital signal processing*. 2016: McGraw-Hill.
[25] Rabiner, L.R. and B. Gold, *Theory and application of digital signal processing.* Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p., 1975.
[26] Quinquis, A., *Digital signal processing using MATLAB*. 2008: Wiley Online Library.
[27] Champagne, B. and F. Labeau, *Discrete time signal processing.* Class Notes for the Course ECSE-412, Department of Electrical & Computer Engineering, McGill University, 2004: p. 190-194.
[28] Terrell, T.J., *Introduction to digital filters*. 1988: Springer.