International Core Journal of Engineering Volume 8 Issue 7, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

Breast Cancer Evaluation Model based on Principal Component Analysis

Shiquan Zhu, Yining An, Wenjie Sun

School of Artificial Intelligence, North China University of Science and Technology, Tangshan, 063200, China

Abstract

In this paper, the eigenvalues of breast cancer cell nuclei are used for Principal Component Analysis, and a comprehensive evaluation model of breast cancer is established based on the obtained principal components. Based on this model, the comprehensive score of the patient's breast cancer is obtained. The larger the value, the higher the risk of the tumor, the score greater than zero is a malignant tumor, and the score less than zero is a benign tumor. After data test, the correct rate of breast cancer evaluation model is as high as 86.64%.

Keywords

Principal Component Analysis; Breast Cancer; Comprehensive Evaluation Model.

1. Introduction

The breast cancer database used in this article was compiled by the University of Wisconsin. The dataset has a total of 569 patient data, which includes patient IDs, diagnostic labels for breast cancer (M: malignant B: benign), and 30 feature values. Eigenvalues include radius (distance from nucleus center to peripheral points), texture, nucleus perimeter, nucleus area, smoothness (local variation of radius length), compactness (perimeter*perimeter/area-1), concavity (degree of contour concavity), concave points (number of contour concavities), symmetry, fractal dimension (shoreline approximation-1) mean, standard deviation and maximum. The mean value describes the overall characteristics of the sample nuclei; the standard deviation describes the fluctuation of each eigenvalue of the sample; the maximum value is not the maximum value of the sample, but the average value of the top three eigenvalues of the sample, which reduces the generation of error. These features are calculated from digital images of fine needle extracts of breast masses, and the eigenvalues in the dataset describe the morphological characteristics of the nuclei in the sample images [1]. This data describes the magnitude of nuclear eigenvalues associated with benign and malignant tumors. Through this data set, scientists can automatically detect the nature of tumors through data analysis, assist doctors to make correct clinical judgments, avoid missed diagnosis and misdiagnosis, and then improve the hospital's treatment level for tumor patients [2].

2. Model Building and Solving

Principal Component Analysis (PCA) is an unsupervised machine learning algorithm that can achieve dimensionality reduction of data. The basic principle of PCA is to transform a set of correlated data into a linearly uncorrelated data set by means of orthogonal transformation, and perform dimensionality reduction to a certain extent [3]. In this paper, the principal component analysis method is used to reduce the dimension of 30 eigenvalues, and the principal components obtained by dimension reduction are used to obtain a comprehensive scoring model based on the principal components, thereby obtaining a comprehensive evaluation model of breast cancer for patients [4].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

2.1 Data Preprocessing

Firstly, the original data is normalized, and then the correlation coefficient matrix of 30 eigenvalues is obtained. In the 30*30 correlation matrix, the correlation coefficients between the variables are basically greater than 0.3. From a statistical point of view, it can be considered that there is a good correlation between the variables. Therefore, principal component analysis can be used for this dataset.

Correlation	Zscore(radius_mean)	Zscore(texture_mean)	Zscore(perimeter_mean)	Zscore(area_mean)	Zscore(smoothness_mean)	Zscore(compactness_mean)	Zscore(concavity_mean)	Zscore(concavepoints_mean)	Zscore(symmetry_mean)	Zscore(fractal_dimension_mean)
Zscore(radius_mean)	1	0.324	0.998	0.987	0.171	0.506	0.677	0.823	0.148	-0.312
Zscore(texture_mean)	0.324	1	0.33	0.321	-0.023	0.237	0.302	0.293	0.071	-0.076
Zscore(perimeter_mean)	0.998	0.33	1	0.987	0.207	0.557	0.716	0.851	0.183	-0.261
Zscore(area_mean)	0.987	0.321	0.987	1	0.177	0.499	0.686	0.823	0.151	-0.283
Zscore(smoothness_mean)	0.171	-0.023	0.207	0.177	1	0.659	0.522	0.554	0.558	0.585
Zscore(compactness_mean)	0.506	0.237	0.557	0.499	0.659	1	0.883	0.831	0.603	0.565
Zscore(concavity_mean)	0.677	0.302	0.716	0.686	0.522	0.883	1	0.921	0.501	0.337
Zscore(concavepoints_mean)	0.823	0.293	0.851	0.823	0.554	0.831	0.921	1	0.462	0.167
Zscore(symmetry_mean)	0.148	0.071	0.183	0.151	0.558	0.603	0.501	0.462	1	0.48
Zscore(fractal_dimension_mean)	-0.312	-0.076	-0.261	-0.283	0.585	0.565	0.337	0.167	0.48	1

Fig. 1 Heatmap of correlations among the top 10 variables

As shown in Figure 1, this paper selects the correlation of the top 10 variables to make a heat map. On the diagonal line of the heat map is the correlation coefficient of the variable itself, red represents a strong positive correlation of the variable, blue represents a strong negative correlation, and the light-colored area has a weak correlation.

Using SPSS software, principal component analysis was performed from the standardized data set. As shown in Table 1, the KMO test coefficient of this model is 0.883. The KMO test coefficient is distributed between 0 and 1. When the KMO test coefficient is greater than 0.8, it can be considered that the principal component analysis method can be used for the data, and the principal components can be better extracted. The premise of applying principal component analysis is that there is a strong correlation between variables, so the null hypothesis should be rejected. The null hypothesis of Bartlett's test is that there is no correlation between the variables, that is, the diagonal value of the correlation matrix is 1, and the remaining values are 0. The approximate chi-square value of the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

Bartlett test is 39362.121, the degree of freedom is 435, and the significance is 0, that is, the data set rejects the null hypothesis, and the principal components can be extracted from the data set.

Table 1. KMO and Bartlett's test

KMO Sampling Suitability Quantity		0.832
Bartlett's sphericity test	Approximate chi-square	39362.121
	Degrees of freedom	435
	Salience	0

2.2 Principal Component Analysis

Table 2. Common factor variance

Variable name	Extract	Variable name	Extract
Zscore(radius_mean)	0.954	Zscore(compactness_se)	0.898
Zscore(texture_mean)	0.901	Zscore(concavity_se)	0.833
Zscore(perimeter_mean)	0.958	Zscore(concavepoints_se)	0.761
Zscore(area_mean)	0.961	Zscore(symmetry_se)	0.850
Zscore(smoothness_mean)	0.867	Zscore(fractal_dimension_se)	0.830
Zscore(compactness_mean)	0.910	Zscore(radius_worst)	0.972
Zscore(concavity_mean)	0.921	Zscore(texture_worst)	0.972
Zscore(concavepoints_mean)	0.927	Zscore(perimeter_worst)	0.978
Zscore(symmetry_mean)	0.781	Zscore(area_worst)	0.949
Zscore(fractal_dimension_mean)	0.846	Zscore(smoothness_worst)	0.915
Zscore(radius_se)	0.889	Zscore(compactness_worst	0.904
Zscore(texture_se)	0.764	Zscore(concavity_worst)	0.905
Zscore(perimeter_se)	0.879	Zscore(concave_points_worst)	0.922
Zscore(area_se)	0.862	Zscore(symmetry_worst)	0.924
Zscore(smoothness_se)	0.744	Zscore(fractal_dimension_worst)	0.849

As shown in Table 2, the model has a total of 30 variables, corresponding to 30 components. The principal component analysis method extracts the principal components from the 30 components and extracts most of the information of the components to optimize the data set and simplify the calculation. For example, the Zscore(radius_mean) component, the principal component can explain 95.4% of the information of this component.

In this paper, components with eigenvalues greater than 1 are selected as principal components, and a total of 6 principal components are obtained. The relationship between the eigenvalue and the contribution rate is:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

$$Contribution\ rate = \frac{Eigenvalues}{30} \times 100\%$$

Table 3. Principal component table

principal component	Eigenvalues	Contribution rate %	Cumulative contribution rate %
1	13.282	44.272	44.272
2	5.691	18.971	63.243
3	2.818	9.393	72.636
4	1.981	6.602	79.239
5	1.649	5.496	84.734
6	1.207	4.025	88.759

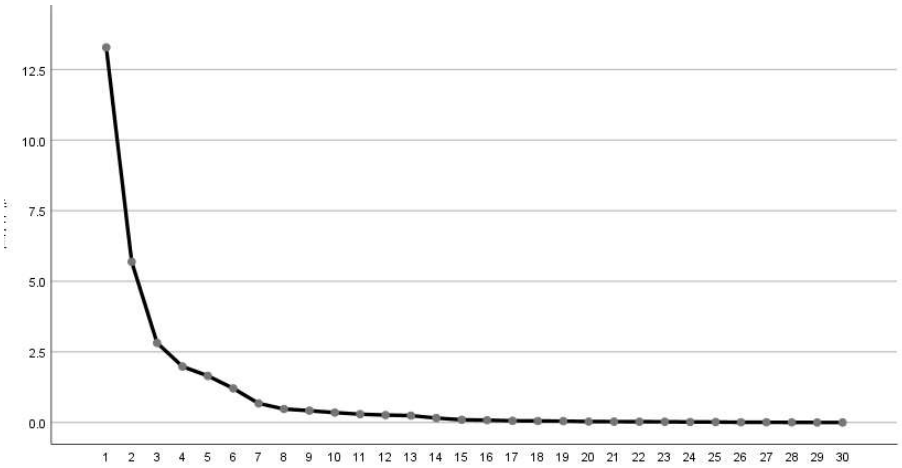

Fig. 2 Gravel diagram of principal components

Zscore(radius_mean)	0.798	Zscore(radius_se)	0.751	Zscore(radius_worst)	0.831
Zscore(texture_mean)	0.378	Zscore(texture_se)	0.064	Zscore(texture_worst)	0.381
Zscore(perimeter_mean)	0.829	Zscore(perimeter_se)	0.770	Zscore(perimeter_worst)	0.862
Zscore(area_mean)	0.805	Zscore(area_se)	0.739	Zscore(area_worst)	0.820
Zscore(smoothness_mean)	0.520	Zscore(smoothness_se)	0.053	Zscore(smoothness_worst)	0.466
Zscore(compactness_mean)	0.872	Zscore(compactness_se)	0.621	Zscore(compactness_worst)	0.766
Zscore(concavity_mean)	0.942	Zscore(concavity_se)	0.560	Zscore(concavity_worst)	0.834
Zscore(concavepoints_mean)	0.951	Zscore(concavepoints_se)	0.668	Zscore(concave_points_worst)	0.914
Zscore(symmetry_mean)	0.504	Zscore(symmetry_se)	0.155	Zscore(symmetry_worst)	0.448
Zscore(fractal_dimension_mean)	0.235	Zscore(fractal_dimension_se)	0.374	Zscore(fractal_dimension_worst)	0.480

Fig. 3 first principal component

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

The contribution rate of the first principal component F1 is 44.272%. F1 is a positive loading on all variables. F1 has approximately higher positive loadings on the mean, standard deviation, and maximum variables of radius, perimeter, area, compactness, concavity, and number of concavities. The remaining variables have approximately lower positive loadings. It reflects the comprehensive morphological characteristics of tumor cell nuclei, so the first principal component can be called the comprehensive nuclear morphological component.

Zscore(radius_mean)	-0.558	Zscore(radius_se)	-0.252	Zscore(radius_worst)	-0.525
Zscore(texture_mean)	-0.142	Zscore(texture_se)	0.215	Zscore(texture_worst)	-0.108
Zscore(perimeter_mean)	-0.513	Zscore(perimeter_se)	-0.213	Zscore(perimeter_worst)	-0.477
Zscore(area_mean)	-0.551	Zscore(area_se)	-0.363	Zscore(area_worst)	-0.523
Zscore(smoothness_mean)	0.444	Zscore(smoothness_se)	0.488	Zscore(smoothness_worst)	0.411
Zscore(compactness_mean)	0.362	Zscore(compactness_se)	0.555	Zscore(compactness_worst)	0.343
Zscore(concavity_mean)	0.144	Zscore(concavity_se)	0.470	Zscore(concavity_worst)	0.234
Zscore(concavepoints_mean)	-0.083	Zscore(concavepoints_se)	0.311	Zscore(concave_points_worst)	-0.020
Zscore(symmetry_mean)	0.454	Zscore(symmetry_se)	0.439	Zscore(symmetry_worst)	0.338
Zscore(fractal_dimension_mean)	0.875	Zscore(fractal_dimension_se)	0.668	Zscore(fractal_dimension_worst)	0.657

Fig.4 Second principal component

The contribution rate of the second principal component F2 is 18.971%. F2 has an approximately higher load on the mean, standard deviation and maximum variables of radius, perimeter, area. F2 has approximately higher positive loadings in smoothness, compactness, symmetry, and fractal dimension variables. By observing the raw data, it can be found that the radius, perimeter and area of malignant tumor cell nuclei have increased significantly compared with benign ones. Smoothness describes the smoothness of the edges of the nucleus. Symmetry describes the regularity of the morphology of the nucleus. The fractal dimension reflects the shape characteristics of the image of the nucleus. Compared with benign eigenvalues, the smoothness and symmetry of malignant tumor cells have obvious overlapping intervals, and the distinguishability is poor. The eigenvalues of the fractal dimension of malignant and benign tumor cells have obvious overlapping intervals and a large number of outliers. Therefore, the second principal component can be called the nuclear integrated radius component.

The contribution rate of the third principal component F3 is 9.393%. F3 has higher positive loadings on the standard deviation variables of radius, texture, nucleus perimeter, nucleus area, smoothness, compactness, concavity, concavity, symmetry, fractal dimension. There is a larger load on its maximum variable. The standard deviation describes the volatility of nuclei in the eigenvalues. The

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

maximum value describes the variability of nuclei in eigenvalues. Therefore, the third principal component can be called the nuclear fluctuation variation component.

The contribution rate of the fourth principal component F4 is 6.602%. F4 has approximately higher positive loadings on the mean, standard deviation, and max variables of the texture. Texture feature is one of the important basis for clinical diagnosis of tumor. The borders of benign tumor nuclei are smooth and oval, and the borders of malignant tumor cells are rough and crab-like. By observing the original data, it can be found that the texture feature values of the two types of nuclei have great differences in the mean and maximum values, and the standard deviations have overlapping intervals. Therefore, the fourth principal component can be called the comprehensive texture component of the nucleus.

The contribution rate of the fifth principal component F5 is 5.496%. F5 has higher positive loadings on the mean, standard deviation, and max variables of smoothness. The smoothness feature is one of the important basis for clinical diagnosis of benign and malignant tumors. The nuclei of benign tumors are approximately round and smoother than those of malignant tumors. Therefore, the fifth principal component can be called the integrated smoothing component of the nucleus.

The contribution rate of the sixth principal component F6 is 4.025%. F6 has higher loadings on the mean, standard deviation, and max variables of symmetry. The nuclei of benign tumors have less atypia and better symmetry than those of malignant tumors. Therefore, the sixth principal component can be called the comprehensive symmetry component of the nucleus.

2.3 Comprehensive Evaluation Model

1) Find the coefficient of each component in the linear combination of each principal component. The general formula is:

$$u_{ij} = \frac{Number\ of\ loads}{Eigenvalues^{\frac{1}{2}}}$$

2) For example, the first coefficient of the principal component F1 is calculated as follows:

$$u_{11} = \frac{0.798}{13.282^{\frac{1}{2}}} = 0.218963$$

Find the score for the comprehensive evaluation model:

Principal component:

$$F1 = u_{11}{*}Zradius\_mean + u_{12}{*}Ztexture\_mean + \cdots + u_{130}{*}Zfractal\_dimension\_worst$$
$$\vdots$$
$$\vdots$$
$$F6 = u_{61}{*}Zradius\_mean + u_{62}{*}Ztexture\_mean + \cdots + u_{630}{*}Zfractal\_dimension\_worst$$

Comprehensive score:

$$F = F1{*}\frac{44.272}{88.759} + F2{*}\frac{18.971}{88.759} + \cdots + F6{*}\frac{4.025}{88.759}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

3) Standardize the comprehensive score obtained:

Table 4. Normalized composite score for selected patients

id	diagnosis	Zscore
8510426	B	-0.41938
8510653	B	-0.47716
8510824	B	-0.7696
842302	M	2.40314
842517	M	0.0961
84300903	M	1.2985

The normalized composite scores for the intercepted part are shown in the table. By observing the comprehensive score of each patient, it is not difficult to find that the comprehensive score of most benign tumors is less than zero, and the comprehensive score of most malignant tumors is greater than zero. Therefore, this characteristic can be used to evaluate the benign and malignant of breast cancer in patients. Therefore, the comprehensive score can be called the tumor risk score. The larger the value, the higher the tumor risk. The score is greater than zero for malignant tumors, and the score is less than zero for benign tumors.

Table 5. Comprehensive evaluation results analysis table

Analysis of evaluation results	Number of benign tumors	Number of malignant tumors	Total
Number of false reviews	48	28	76
Number of correct evaluations	309	184	493
Total	357	212	569

As shown in the table, the correct rate of the evaluation model for breast cancer is as high as 86.64%. The evaluation model can assist doctors to make correct clinical judgments on breast cancer, avoid the occurrence of delayed treatment timing due to misdiagnosis, and help improve the hospital's treatment level for breast cancer patients.

3. Summary

In this paper, the comprehensive evaluation model of breast cancer based on the principal component analysis method provides scientific and powerful data support for the identification of breast cancer, and further improves the hospital's treatment level for breast cancer patients. Under the condition of known eigenvalues of breast cancer cell nuclei, the established evaluation model of breast cancer has good promotion value. This model can play a certain role in the diagnosis of other diseases.

References

[1] Yang Chenxue. Research on Image Feature Learning Method and Application [D]. University of Electronic Science and Technology of China, 2016.

[2] Jin Weizhe, Wang Dajiang, Zhuang Rong, Wang Su, Cheng Qiwei, Wang Wei, Pan Qi. Computer-aided digital image analysis of benign and malignant breast lesions [J]. Advances in Anatomy Science, 1997(04):75- 78. DOI: 10.16695/j.cnki.1006-2947.1997.04.019.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0038

[3] Lin Haiming, Zhang Wenlin. Similarities and Differences Between Principal Component Analysis and Factor Analysis and SPSS Software: Discussion with Liu Yumei, Lu Wendai and Other Comrades [J]. Statistical Research, 2005(03): 65-69. DOI: 10.19343 /j.cnki.11-1302/c.2005.03.015.

[4] Lin Haiming, Du Zifang. Problems that should be paid attention to in the comprehensive evaluation of principal component analysis [J]. Statistical Research, 2013, 30(08): 25-31. DOI: 10.19343/j.cnki.11-1302/c.2013.08 .004.