TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Principal Component Analysis and Correlation Analysis on Wisconsin Breast Cancer Dataset

[1]P. R. Anisha, [2]Dr. B. Vijaya Babu

[1]Research Scholar, Dept. of CSE, KLEF, Guntur
[2] Professor, Dept. of CSE, KLEF, Guntur

Article Info
Volume 82
Page Number: 8316 - 8324
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 07 February 2020*

Abstract:
Breast cancer is considered to beone of the serious malignant tumorthat originates from the cells present in the breast. The disease arises typically in women, but additionally men can also be rarelyget effected. During the diagnosis of breast cancer, odd growth of cells in breast takes vicinity and this increase may be in two sorts which are benign (non-cancerous) and malignant (cancerous). For data preparation tools such as IBM SPSSModeler 14.2, Access 2003 and Excel 2003 and IBM SPSSStatistics 16 was used to calculate Principal Component Analysis to find the adequacy of the dataset attributes for the prediction of the nature of Breast Cancer Disease. Further correlation analysis is also taken up to figure the dependencies among attributes. The paper focus on the foresaid experimentation and the results are justified to generated the appropriate and the sufficiency of the attributes for the prediction.

Keywords: *Breast Cancer, Correlation analysis, IBM SPSS, Principal Component Analysis.*

INTRODUCTION

Threatening development of disease cells is a critical issue with growing example of occasion far and wide. The amount of as of late broke down dangerous development cases in 2008 were generally 12.7 million, with about5.6 million in financially created nations and of about 7.1 million in other nations. Breast malignant growth is a significant worry in the United States and the one of the subsequent disease reason for death following lung cancer. The American Cancer Society extends that 246,660 new bosom malignancy cases have been analyzed in 2016[2]. In spite of the fact that bosom malignancy is considerably more typical among ladies, it is accounted for that 500 men lost their lives because of bosom disease in the year 2015.

New cases in female having Breast Cancer were 1,383,523 with a mortality of about of 458,367. In various developed countries, the frequencies of bosom disease was 692,242 and number of death was 1, 89,488. Similarly in different countries, the rates and mortality of such cases remained to be around 691,281 and 268,879 separately, age homogenous occurrence and passing rates for bosom malignant growth were 39 and 12.5 per 100,000 independently [1].Although there are numerous general wellbeing non transferable maladies bosom disease has been a main source of death.

The worldwide problem of bosom malignancy broke down was 1.38 million, which remained as the second reason for death. Bosom Cancer remained as one of the most concerning malady for each one of every 9 females, other than the mental effect that it makes.It's a sickness inside the tissues of the bosom where the threatening cells are being created. Bosom malignant growth is of two sorts, lobular disease that starts in a few little sacks inside the bosom that produce milk and Ductal malignant growth that

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

creates inside the cylinders that convey milk from the lobules to the areola among these the later happens the most than the lobular cancer[3] [4]. Shockingly, the frequency of carcinoma in China is expanding doubly as fast as that of the overall rates, outstandingly in urban territories [6]. In this way, decreasing the frequency of bosom disease has become an indispensable world medical problem. the exact reasons for bosom canceris still not completely prepared, however a few hazard factors identified with bosom malignant growth, similar to case history, menarche, adiponectin levels , and life vogue, are known. Finding a great deal of hazard factors and uncovering an exactitude forecast model can benefit the compelling bar and intercession of bosom malignancy.

In this examination paper Principal segment Analysis for include extraction. It's an element extraction system that takes partner symmetrical change in order to change over a lot of perceptions of apparently connect parameters into a lot of estimations of directly non associated parameters known as head components[5].The paper is sorted out with the related work followed by a brief about the datasets which are additionally registered for the information investigation utilizing PCA , Correlation examination and finished up with an end driven from the experimentation.

RELATED WORK

The Wisconsin Breast Cancer Dataset was created by Dr. Wolberg in 1995. From there on, a ton of work is finished by different restorative and united specialists to disentangle the bosom malignant growths identification and anticipation. In 1999 Xin Yao et al. for bosom disease determination hasimplemented artificial neural system utilizing Negative connection preparing calculation usingtwo approaches viz.evolutionary and group approach. In 2004 Tuba Kiyan et al. has applied Neural Network on WBCD to appraise the conclusion exactness of different systems. In 2007 Dr. Sumathi et al.[7] have utilized hereditary calculations way to deal with WBDC and found that hereditary calculation not justimprovised the level of exactness yet in addition diminish the time taken to prepare the system. In 2010Val'erie[8] have relatively considered different measurable models to get an alluring outcome from WBCD. In 2009, Y. Iraneus Anna Rejani[9] and Dr. S. ThamaraiSelvi utilized SVM for the early distinguishing proof of Breast Cancer. In 2012, Muhammad Rafi et al. utilized SVM and RVM methods for report classification without utilizing least precision limit and find that anticipating exactness of RVM is a lot higher than SVM. In 2012, Z Qinli et al. [10]apply SVM approach and its application to bosom malignant growth finding. It is seen that it is able to lessen the speculation blunder and computational expense. In 2013, D. Kishore [11]used Big Data ways to deal with Medical field and cleared a route for its execution to recognize Breast Cancer.

DATASET BRIEF

The dataset used in the said research work is freely reachable and was made by Dr. William H. Wolberg, doctor at the University of Wisconsin Hospital at Madison, Wisconsin, USA, and was given by OlviMangasarian on July fifteenth, 1992. To shape the dataset Dr. Wolberg utilized uid tests, taken from patients having strong bosom masses related a simple touse graphical bug known as Xcyt, that is equipped for play out the investigation of cytologic alternatives upheld a computerized sweep. Bosom disease analysis is arranged into classifications Benign and Malignant. Kind bumps zone unit unusual irregularities anyway not destructive while Malignant knots region unit carcinogenic protuberances in female body. Each element is assessed in a scope of 1 to 10, with 1 being the closest to considerate and 10 the closest to dangerous. Measurable examination indicated that the ensuing 9 attributes disagree impressively among kind and harmful were tests comprising of : consistency of cell form(A), consistency of cell size(B), cluster thickness(C), uncovered nuclei(D), cell size(E), typical nucleoli(F), bunch cohesiveness(G), atomic chromatin(H) and mitoses(I) individually.

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Table 1: Dataset Attribute description

S.No	Attribute Description	Representation
1	Clump Thickness: 1 – 10	A
2	Uniformity of Cell Size: 1 - 10	B
3	Uniformity of Cell Shape: 1 - 10	C
4	Marginal Adhesion: 1 - 10	D
5	Single Epithelial Cell Size: 1 - 10	E
6	Bare Nuclei: 1 – 10	F
7	Bland Chromatin: 1 – 10	G
8	Normal Nucleoli: 1 – 10	H
9	Mitoses: 1 – 10	I
10	Class benign, malignant	Class

Table 2: Sample dataset

A	B	C	D	E	F	G	H	I	Class
5	4	4	5	7	10	3	2	1	Benign
3	1	1	1	2	2	3	1	1	Benign
6	8	8	1	3	4	3	7	1	Benign
4	1	1	3	2	1	3	1	1	Benign
8	10	10	8	7	10	9	7	1	Malignant
1	1	1	1	2	10	3	1	1	Benign
2	1	2	1	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign
1	1	1	1	1	1	3	1	1	benign
2	1	1	1	2	1	2	1	1	benign
5	3	3	3	2	3	4	4	1	malignant
1	1	1	1	2	3	3	1	1	benign
8	7	5	10	7	9	5	5	4	malignant
7	4	6	4	6	1	4	3	1	malignant
4	1	1	1	2	1	2	1	1	benign
4	1	1	1	2	1	3	1	1	benign
10	7	7	6	4	10	4	1	2	malignant
6	1	1	1	2	1	3	1	1	benign
7	3	2	10	5	10	5	4	4	malignant

DATA ANALYSIS

The data used in this research was initiallyanalyzed using Chi Square tests, descriptive statistics, and Factor Analysis using SPSS 17.0. The obtained data are presented in the tables and discussed. In order to perform the factor analysis over the datasets all the records are segmented into two class 1 and class 2 i.e benign and malignant respectively.Factor analysis is a crucial measure to spot common dimension factors. It's a data reduction technique that may facilitate to work out a lesser range of underlying dimensions of an outsized set of inter-correlated variables.On individual analysis of each of the class the results are as follows :

Table 3: The Mean, standard deviation calculation for Class 1

Descriptive Statistics			
	Mean	Std. Deviation	Analysis N
Clump Thickness(1-10)	2.96	1.672	443
Uniformity of Cell Size (1-10)	1.31	.857	443

TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Uniformity of Cell Shape(1-10)	1.42	.958	443
Marginal Adhesion	1.35	.918	443
Single Epithelial cell size	2.11	.878	443
Bare Nuclei (1-10)	1.35	1.179	443
Bland Chromatin(1-10)	2.08	1.063	443
Normal Nucleoi(1-10)	1.26	.956	443

The Kaiser-Meyer-Olkin (KMO) paradigm and Bartlett's checks were placed in usein request to discover whether factor examination is reasonable for this information. KMO measures inspecting ampleness Associate andBartlett's tests the invalid speculation that the principal network is a character grid.

Table 4 represents that, for the given information the KMO score is zero.790. This KMO cost portrays that the example that was considered was satisfactory and accordingly worthy, and in this manner the conveyance of significant worth is adequate enough for performing factor investigation. The Bartlett's check of sphericity esteem was critical (Chi sq. = 476.583, p < 0.001), thus factor investigation is satisfactory.

Table 4: KMO and Bartletts test to check the dataset sufficiency for Class 1

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.790
Bartlett's Test of Sphericity	Approx. Chi-Square	889.882
	Df	36
	Sig.	.000

Table 5explains the Eigen values which relates to every linear component (factor) before performing the extraction, after extraction rotation. The Eigen values connected to every issue depicts the variance explained by the particular linear component. It additionally gives the Eigen values in terms of the percentage of variance. Here all factors with Eigen values larger than 1 is considered. The first component gave a variance of 34.23% and is given in table 5.

Table 5: Total variance, Extraction Method: PRINCIPAL COMPONENT ANALYSIS – Class 1

Component		Initial Eigenvalues			Rotation Sums of Squared Loadings		
		Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
	1	3.201	35.567	35.567	3.098	34.423	34.423
	2	1.082	12.020	47.587	1.185	13.164	47.587
	3	.988	10.978	58.565			
	4	.859	9.549	68.114			
	5	.842	9.352	77.466			
	6	.732	8.139	85.604			
	7	.532	5.906	91.511			
	8	.481	5.343	96.853			
	9	.283	3.147	100.000			

By factor analysis, 2 major components there were 9 attributes that were extracted.Table 6 depicts the components obtained on the class 1. It shows that for class 1 the components from attribute 1 to attribute 8 shows maximum correlation with a marginal difference with attribute 9.

Table 6: Rotated Component Matrix[a] for Class 1

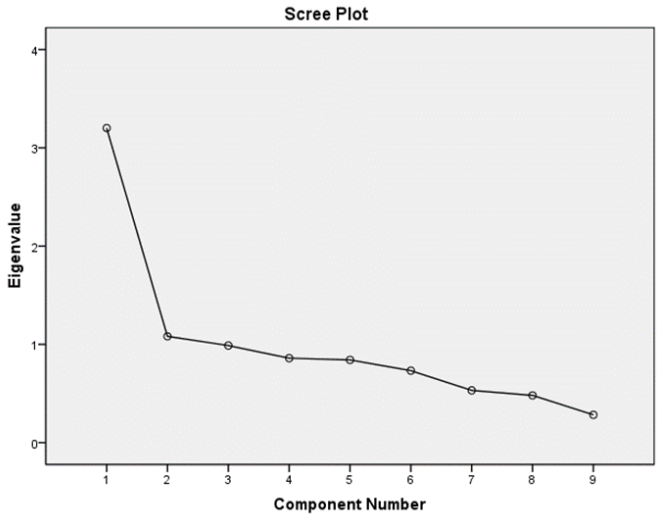
	Component	
	1	2
Uniformity of Cell Size (1-10)	.808	.146
Uniformity of Cell Shape(1-10)	.752	.048
Normal Nucleoi(1-10)	.698	.096
Single Epithelial cell size	.624	.132

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Bare Nuclei (1-10)	.558	.458
Clump Thickness(1-10)	.478	-.157
Bland Chromatin(1-10)	.477	-.177
Marginal Adhesion	.469	.354
Mitoses(1-10)	-.128	.862
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.		

The graphical representation of the result generated on PCA is given in Figure 1.

Figure 1: Screen plot of components on rotation matrix for Class 1



Similar procedure is followed on Class 2, so as to identify the number of components that could be generated and the attributes that fall correlation for the identification of class 2. The Descriptive Statistics for class 2 is shown in Table 7.

Table 7: The Mean, standard deviation calculation for class 2

Descriptive Statistics			
	Mean	Std. Deviation	Analysis N
Clump Thickness(1-10)	7.19	2.438	239
Uniformity of Cell Size (1-10)	6.58	2.724	239
Uniformity of Cell Shape(1-10)	6.56	2.569	239
Marginal Adhesion	5.59	3.197	239
Single Epithelial cell size	5.33	2.443	239
Bare Nuclei (1-10)	7.63	3.117	239
Bland Chromatin(1-10)	5.97	2.282	239
Normal Nucleoi(1-10)	5.86	3.349	239
Mitoses(1-10)	2.60	2.564	239

As in the previous class 1 case the Kaiser-Meyer-Olkin (KMO) criterion and Bartlett's tests were made use of so as to test whether factor analysis is appropriatefor these data considered, the same results were generated for class 2 as well as shown in the Table 4.As the Table 5 explains the Eigen values related to every linear component (factor) before extraction, once extraction rotation. The Eigen values related to every factor represent the variance explained by that particular linear component, Table 8 depicts constant for class 2. the primary factor component within the table 8 explains 27.626% of the variance".

Table 8: Total variance, Extraction Method: PRINCIPAL COMPONENT ANALYSIS – Class 2

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.850	31.662	31.662	2.486	27.626	27.626
2	1.301	14.452	46.114	1.597	17.740	45.367
3	1.023	11.365	57.479	1.090	12.112	57.479
4	.943	10.476	67.955			
5	.802	8.910	76.865			
6	.681	7.567	84.432			
7	.626	6.958	91.389			
8	.516	5.730	97.120			
9	.259	2.880	100.000			

Through factor analysis, 3 major components were extracted from the 9 attributes. Table 9 depicts the components obtained on the class 2. It shows that for class 2 the components from attribute 1 to attribute 7 shows maximum correlation with a marginal difference with attribute 8 and attribute 9.

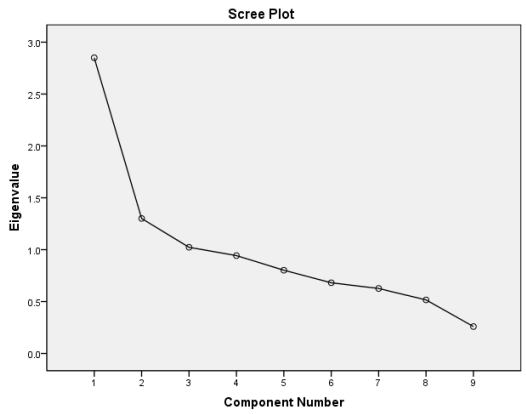
January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Table 9: Rotated Component Matrix[a] for Class 2

	Component		
	1	2	3
Uniformity of Cell Size (1-10)	.748	.359	.156
Single Epithelial cell size	.680	.092	.004
Uniformity of Cell Shape(1-10)	.675	.402	.220
Normal Nucleoi(1-10)	.602	.009	-.257
Mitoses(1-10)	.567	-.090	.058
Bare Nuclei (1-10)	-.338	.748	.117
Bland Chromatin(1-10)	.337	.607	-.091
Marginal Adhesion	.291	.597	-.336
Clump Thickness(1-10)	.109	-.070	.901
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.			

The graphical representation of the result generated on PCA for Class 2 is given in Figure 2.

Figure 1: Screen plot of components on rotation matrix for Class 1



The basic code of the PCA method is to pick out the lowest number of components to gain the maximum amount of total information contained in primary data. This technique provides a graphic visualization of the map of individuals within the study consistent with similarities between them and therefore the map of variables consistent with their correlations. Although this technique relies on an equivalent principle as in the case of factor analysis ,the most component analysis differs from it by the means of definition of parts associated with initial data table and therefore the intention means of the gap between points.

On performing the Descriptive statistics, calculating the total variance and generating the rotatedcomponent matrix using the Principal Component Analysis method of extraction the entire dataset inclusive of both the classes are further analysis to find the intercorrelation between the attributes and the data sufficiency for the analysis. For doing the same the descriptive statistics inclusive of both the classes is performed and is shown in Table 10.

"To calculate the data adequacy and to calculate the null hypothesis the KMO measures and Bartlett's test is run. The results obtained is shown in Table 11. Table 11 explains that, for the given data the KMO score obtained is 0.935. This KMO worth shows that the sample was sufficient and is acceptable, and also the distribution of value is sufficient enough for performing the factor analysis".

Table 11: KMO and Bartletts test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.935
Bartlett's Test of Sphericity	Approx. Chi-Square	4760.785
	Df	36
	Sig.	.000

Table 10: The Mean, standard deviation calculation for entire dataset

	Mean	Std. Deviation[a]	Analysis N[a]
Clump Thickness(1-10)	4.44	2.823	682
Uniformity of Cell Size (1-10)	3.15	3.066	682
Uniformity of Cell Shape(1-10)	3.22	2.990	682
Marginal Adhesion	2.83	2.866	682
Single Epithelial cell size	3.24	2.224	682
Bare Nuclei (1-10)	3.55	3.645	682
Bland Chromatin(1-10)	3.45	2.451	682
Normal Nucleoi(1-10)	2.87	3.054	682
Mitoses(1-10)	1.60	1.734	682

The Eigen values related with each factor signify the variance clarified by that specific linear component, Table 12 depicts the first factor component in the table 12 explains 65.548% of the variance.

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Table 12: Total variance, Extraction Method: PRINCIPAL COMPONENT ANALYSIS

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.899	65.548	65.548	5.899	65.548	65.548
2	.776	8.623	74.171			
3	.539	5.987	80.158			
4	.460	5.110	85.268			
5	.380	4.227	89.495			
6	.302	3.356	92.851			
7	.294	3.270	96.121			
8	.261	2.896	99.017			
9	.088	.983	100.000			

The rotated component matrix is performed on the entire dataset which resulted in generating only one component due to the intercorrelation between all the attributes in the entire dataset and is shown in Table 13.

Table 13: Rotated Component Matrix[a]

	Component
	1
Clump Thickness(1-10)	.734
Uniformity of Cell Size (1-10)	.925
Uniformity of Cell Shape(1-10)	.917
Marginal Adhesion	.808
Single Epithelial cell size	.817
Bare Nuclei (1-10)	.814
Bland Chromatin(1-10)	.840
Normal Nucleoi(1-10)	.815
Mitoses(1-10)	.559

Extraction Method: Principal Component Analysis.

Correlation is a statistical technique that may showswhether or not and how strongly pairs of variables are connected..To find the correlation between the attributes taken for the breast cancer data set table 14 shows the result that is generated through SPSS.

Table 14 : Correlations matrix among all the attributes in the breast cancer dataset

Sig. (2-tailed)		Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial cell size	Bare Nuclei	Bland Chromatin	Normal Nucleoi	Mitoses
Clump Thickness	Correlation	1	.643**	.654**	.488**	.524**	.594**	.554**	.534**	.351**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000
	N	682	682	682	682	682	682	682	682	682
Uniformity of Cell Size	Correlation	.643**	1	.907**	.707**	.753**	.691**	.756**	.719**	.461**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000
	N	682	682	682	682	682	682	682	682	682
Uniformity of Cell Shape	Correlation	.654**	.907**	1	.686**	.722**	.714**	.735**	.718**	.441**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000
	N	682	682	682	682	682	682	682	682	682
Marginal Adhesion	Correlation	.488**	.707**	.686**	1	.594**	.670**	.669**	.603**	.419**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000
	N	682	682	682	682	682	682	682	682	682

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

Single Epithelial cell size	Correlation	.524**	.753**	.722**	.594**	1	.585**	.618**	.629**	.480**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000	.000
	N	682	682	682	682	682	682	682	682	682
Bare Nuclei	Correlation	.594**	.691**	.714**	.670**	.585**	1	.681**	.584**	.339**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.000
	N	682	682	682	682	682	682	682	682	682
Bland Chromatin	Correlation	.554**	.756**	.735**	.669**	.618**	.681**	1	.666**	.346**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000	.000
	N	682	682	682	682	682	682	682	682	682
Normal Nucleoi	Correlation	.534**	.719**	.718**	.603**	.629**	.584**	.666**	1	.434**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000
	N	682	682	682	682	682	682	682	682	682
Mitoses	Correlation	.351**	.461**	.441**	.419**	.480**	.339**	.346**	.434**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	
	N	682	682	682	682	682	682	682	682	682

*. Correlation is significant at the 0.01 level (2-tailed).

The graphical representation of the matrix is shown as in the Figure 3.

Figure 3: Graph for correlation matrix

TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 8316 - 8324

CONCLUSION

Diagnosis or prognosis of any serious disease such as breast cancer is a very challenging problem and it requires many preprocesses experiments and significant dataset. The fore said datasets the are gathered from UCI machine repository. In this study, in order to identify the breast cancer dataset authenticity the datasets are preprocessed using IBM SPSS though which PCA is performed to identify the best fitting attributes and correlation analysis is taken up to analyze the correlation between the attributes using which the nature of the breast cancer could be predicted.

REFERENCES

1. Thun MJ, Jemal A, Ward E (2011) Global cancer prevalence and mortality. In: DeVita VT, Lawrence TS and Rosenberg SA, Editors, Cancer Principles & Practice of Oncology, ninth Ed. USA. 241-sixty six.
2. American Cancer Society. Breast Cancer Facts. Atlanta: American Cancer Society, Inc
3. Dollinger M, Rosenbaum EH, Cable G (1991) Understanding Cancer in Everyone's guide to cancer remedy. Universal press syndicate employer, USA, p. 1-9.
4. Isa AB, Evren H (2016) Cancer Incidence in Nigeria and Border Countries. Malays j med biol res three: 7-12.
5. https://en.Wikipedia.Org/wiki/Principal_component_analysis.
6. Fan L, Zheng Y, Yu KD et al. Breast cancer in atransitional society over 18 years: Trends andpresent popularity in Shanghai, China. Breast CancerRes Treat 2009;117:409–416.
7. C.P. Sumathi, T.Santhanam and A. Punitha 2007, 琯Combination of genetic algorithm and ART neural network for breast most cancers diagnosis恊, Asian Journal of statistics generation, Medwell journals.
8. Val´erieBourd`es, St´ephaneBonnevay, Paolo Lisboa, R´emyDefrance, David P´erol, Sylvie Chabaud, Thomas Bachelot, Th´er`ese Gargi, and Sylvie N´egrier 2010, Comparison of Artificial Neural International Journal of Distributed and Parallel Systems (IJDPS) four(three), Network with Logistic regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient consequences.
9. Y.Ireaneus Anna Rejani, Dr.S.ThamaraiSelviNoorul 妾Early Detection Of Breast Cancer Using SVM Classifier Technique椒, International Journal on Computer Science and Engineering Vol.1(3), 2009, 127-a hundred thirty.
10. M. Rafi and M.S. Shaikh, 2013, 琯A Comparison of SVM and RVM for Document Classification恊 Proceedings of Computer Science, pp 3-8.
11. D. Kishor, 2013, 琯Big Data: The New Challenges in Data Mining恊. International Journal of Innovative Research in Computer Science & Technology, 1(2), pp. 39-42.