# Preparation of breast cancer data using different preprocessing technique

# An efficient preprocessing technique for breast cancer data

Anupama Y.K., Dr. Amutha.S. S, Dr. Ramesh Babu. D. R.

**ABSTRACT**

Preprocessing is a significant data preparation step in Data mining/Machine Learning techniques, to improve the performance of the classification models and to obtain better results. Exploring efficient preprocessing technique plays a vital role. In this studies, Principle Component Analysis (PCA) preprocessing technique compared with the proposed Preprocess\_Integration(PI) preprocessing technique. The principal component analysis(PCA) is a preprocessing technique that reduces the dimension of the original data. The algorithm proposed in this paper integration of the PCA, correlation matrix, covariance matrix, and chi-square test. Both the PCA and proposed methodology are applied on the Breast cancer data set. Results obtained with PCA are better than the PI technique.

**Keywords: principal component analysis, correlation matrix, covariance matrix, chi-square test**

**INTRODUCTION**

Extracting useful relevant information and hidden knowledge from the dataset helps in developing an efficient algorithm to determine and analyze trends. This phenomenon refers to Data mining. Real-world data is incomplete and can contain many errors. Data preprocessing technique transforms raw data into an understandable format. In Machine Learning Process, Data Preprocessing is a very important step as it is responsible for converting the data into a format that can be understood by the machine learning (ML) algorithm.

Sensors are used to collect the data for any ML application. However, there can be errors generated due to instrumental failures, human errors, noise, redundant data, and incorrect information being collected by these sensors. This erroneous information can influence the performance of the model.

Different pre-processing techniques can be used to remove these anomalies from the collected data. Principal Component Analysis(PCA) is one such technique used to analyze the high dimensional data. The main idea of PCA is to reduce the dimensionality of a highly correlated data set and extract important information from the collected data. This is achieved by converting the largely correlated variables into a small set of variables, known as principal components.

Aim of the paper is to explore an efficient preprocessing technique for breast cancer data. So, the performance of the machine learning model can be improved. The rest of the paper is categorized as follows. Division 2 is about literature survey. Division 3 introduces breast cancer datasets. Division 4 describes proposed methodology consisting of data preparation step, classification step and model evaluation step. Visualization of output is in Division 5.

**2.LITERATURE SURVEY**

Shruti Sehgal[1] et al., proposed a method to separate two different objects having different combinations for four sensors and identified similar sensors or different sensors using the association between variables. Using the PCA technique, it transformed the original dataset to a reduced format using relevant vectors. The main objective is to consider the significant related information and ignoring insignificant information. The extension of the work is to find new patterns by using PCA in neural network representations.

Asma Saleem[2] et al., proposed strategies to accommodate the values of mislaid variables and to identify an instance in a set of results that is very much vast or minor than the nearby existing instances. 2-Sigma (2S) and box-plot (BP) techniques are used to find exceptional instances. The weighted average strategy is used to accommodate the values for mislaid variables. The method proposed is specific to the weather forecast dataset. Accordingly, data preprocessing techniques depend on the domain. As a future scope, it suggests extending this method to general applications.

Trupti A. Aneyrao and R. A. Fadnavis[3] T explored a strategy to eliminate bias and anticipation from data. Mainly focused on direct bias elimination from the dataset. Direct discrimination measure and the rule is used to eliminate and to avoid discrimination.

Lin Rui et al.[4], explored an ensemble strategy for data preprocessing of the digital information. The Ensemble data preprocessing technique is a fusion of PCA and WA (Weighted Average) algorithm. PCA reduces the processing cost of volume information. Accordingly, 73% prediction accuracy was obtained by applying PCA.

Raksha Upadhyay et al.,[5] proposed a methodology to reduce the high dimensionality of databases. In this paper PCA with covariance matrix is used to check the correlated input dimensions. These input dimensions are ordered to uncorrelated PCs (principal components) with low variants and few of the first PCs represented nearly the whole data set. This method is used for analyzing the PC in the simulated RSS pattern.

R. R. Janghel et al.[6], proposed a novel methodology for the prediction of breast cancer. PCA is used to reduce the features of the data set and the results are predicted using different neural network architectures Best results are obtained for the SANE(Symbiotic, Adaptive Neuro-evolution) strategy after applying PCA to the dataset.

Duygu Kaya and MustafaTurk[7] proposed a methodology for breast cancer diagnosis. PCA-Support Vector Machine (SVM) model classifier implemented in parallel using graphical language programming LABVIEW. Better classification results were obtained using this approach

Muduli, D., Dash, R., & Majhi, B[8] explored a methodology to classify breast mass of a mammogram image as normal or abnormal. Features extraction is done using PCA and LDA methods. Moth flame efficient optimization technique with EMI is used for classification. Classification accuracy results obtained as 97% and 98% for datasets MIAS and DDSM respectively.

**PROPOSED METHODOLOGY**

1. **Data set:** The Wisconsin Diagnostic Breast Cancer dataset is used which gives information about the cell nuclei and determines whether the cancer is Benign or Malignant. It consists of 29 predictors which include features like radius, area, perimeter, concavity, texture, etc. The response variable is whether the cancer diagnosis is B(Benign) or M(Malignant). A total of 569 instances are present in the dataset out of which 357 are Benign and 212 are malignant

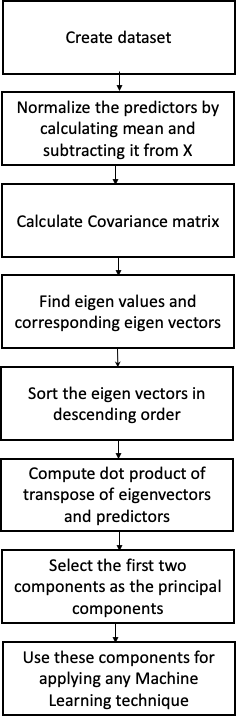
Ten real-valued features are computed for each cell-nucleus:

|  |  |
| --- | --- |
| Radius : | Mean of distances from center to points on the perimeter |
| Texture : | Standard deviation of gray scale values |
| Perimeter | Length of the continuous line forming the boundary of the nucleus |
| Area | Measurement of the surface of nucleus |
| Smoothness | Local variation in radius length |
| Compactness | Perimeter^2 /area -1.0 |
| Concavitiy | Severity of concave portions of the contour |
| Concave Points | Number of concave portions of the contour |
| Symmetry | The orientation of nucleus symmetric axis present in the nucleus. |
| Fractal Dimension | Coastline approximation -1 |

1. **Related Work**

Existing Methodology: Principal Component Analysis

Principal Component Analysis is a dimensionality reduction preprocessing technique. It analyzes a large amount of data and tries to convert its high dimensionality to low dimensionality so that it becomes easier to understand. It converts the correlated data into uncorrelated variables and then stores the most important information as principal components.



The PCA algorithm works as follows:

Step 1: Creating/Collecting a dataset and deciding the predictor and the response variables.

Step 2: Normalizing the predictors by calculating the mean of the data and subtracting it from itself.

X = X - mean(X)

Step 3: Calculating the covariance matrix for the data to determine the covariance present in the dataset.

Step 4: Finding eigenvalues and the associated eigenvectors of the matrix returned from step 3.

Step5: Sorting the eigenvectors in descending order to obtain important information present in the initial components

Step6: Determining the score matrix by calculating the dot product of eigenvectors and predictors obtained from step 2.

Step7: Selecting the first two components, known as the principal components and further apply any machine learning algorithm required.

In this way, the dimensionality of the dataset is reduced from n to 2, where n is the number of features present in the dataset. It is a very useful technique as it not only decreases the dimensionality of the dataset but also tends to improve the accuracy of the predictions made.

Proposed Methodology: Preprocess Integration algorithm works as follows:

Step1: Creating/Collecting a dataset and deciding the predictor and the response variables.

Step2: Calculate contingency table (X’).

(X’)2 = Σ (m – N )2 / N  
m = observed value

n = expected value

Step3 : normalize the data present in the contingency table by calculating the mean of the rows present in the contingency table and then subtracting them from the table.

X’ = X’-mean(X’)

Step4: determine the covariance matrix and different types of correlation matrix(Spearman, Kendall, Pearson)

Step5: Use the original dataset and then normalize the predictors.

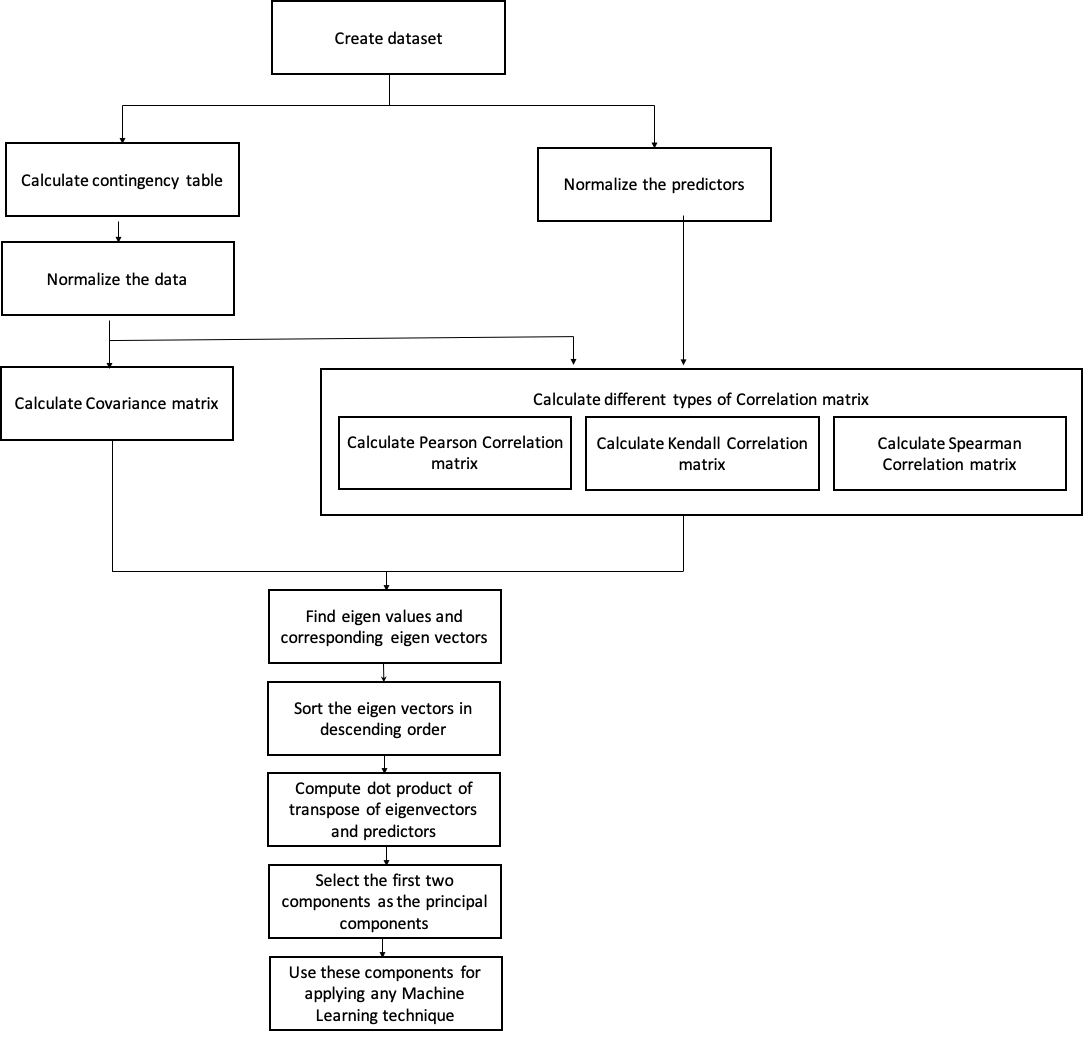
Step6: Calculate different types of correlation matrix (Spearman, Kendall, Pearson)

Step 7: Find eigenvalues and the associated eigenvectors of the matrix returned from steps 4 and 6.

Step8: Sort the eigenvectors in descending order to obtain important information present in the initial components

Step9: Determining the score matrix by calculating the dot product of eigenvectors and predictors obtained from step 2 and step 5

Step10: Selecting the first two components, known as the principal components and further apply any machine learning algorithm required.



The correlation matrix is a matrix where a cell depicts the correlation between two variables. The correlation matrix is of 3 types:

1. Kendall correlation matrix : Determines strength of dependence between two variables.

The following formula is used to calculate the Kendall rank correlation:

A= number of concordant variables

B= Number of discordant variables

M= number of observations

1. Pearson correlation matrix: Determines the relationship between linearly related variables.

The following formula is used to calculate the Pearson rank correlation:

= Pearson R correlation coefficient between a and b

N = number of observations

ai = value of a (for ith observation)

bi = value of b (for ith observation)

1. Spearman correlation matrix: Determines the degree of a monotonic relationship between variables.

The following formula is used to calculate the Spearman rank correlation:

= Spearman rank correlation

= the difference between the ranks of corresponding variables

N= number of observations

Chi-square test is a test used to check whether there is independence, that is no relationship between the categorical variables or not. A contingency table is a table used to check whether a variable depends on the other variable or not.

The test data and training data are divided in the ratio of 25:75 and Logistic regression is used as a linear classification machine learning technique to classify the cancer into benign or malignant.

**Accuracy:**

Accuracy is the result obtained by dividing the total number of correct predictions made by the model and total number of predictions made my the model.

Accuracy =

Accuracy can also be defined as the result obtained by dividing the sum of True negatives and true positives by sum of all the predictions made.

A = True positives

B= False positives

C = True negatives

D = False negatives

**Confusion Matrix:**

Confusion matrix represents the summary of all the predictions made by the model. The confusion matrix consists the information about

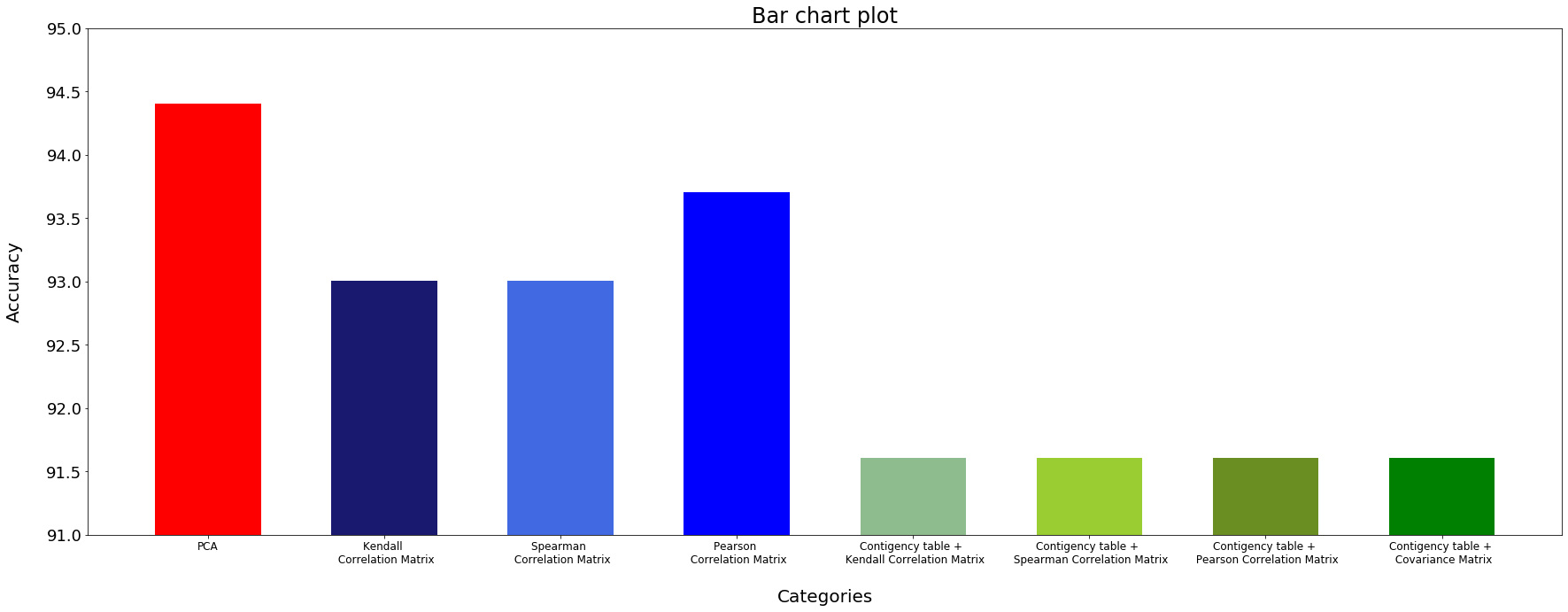
* The number of tuples that were observed as X and predicted as X
* The number of tuples that were observed as Y and predicted as Y
* The number of tuples that were observed as X and predicted as Y
* The number of tuples that were observed as Y and predicted as X

|  |  |  |
| --- | --- | --- |
|  | Predicted as A | Predicted as B |
| Observed as A | A | D |
| Observed as B | B | C |

**RESULTS**

We have considered categorical data, i.e. Breast Cancer dataset which determines whether the cancer is malignant or benign based on various predictors. The following accuracy and confusion matrix were obtained :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CASES** | PCA | Kendall Correlation matrix | Spearman Correlation matrix | Pearson Correlation matrix | Chi square test+ Kendall Correlation matrix | Chi square test+ Spearman Correlation matrix | Chi square test+ Pearson Correlation matrix | Chi square test+ Covariance Matrix |
| ACCURACY | 0.9440559440559441 | 0.9300699300699301 | 0.9300699300699301 | 0.9370629370629371 | 0.916083916083916 | 0.916083916083916 | 0.916083916083916 | 0.916083916083916 |
| CONFUSION MATRIX | [[87 3]  [ 5 48]] | [[83 7]  [ 3 50]] | [[83 7]  [ 3 50]] | [[83 7]  [ 2 51]] | [[86 4]  [ 8 45]] | [[86 4]  [ 8 45]] | [[86 4]  [ 8 45]] | [[86 4]  [ 8 45]] |



**CONCLUSION & FUTURE WORK**

In the proposed methodology PCA is used in combination with Chi Square, Kendall, Pearson, and Spearman correlation matrix. Using PCA as a data preprocessing technique for breast Cancer dataset obtained an accuracy of 94% higher than the other methods.

Work can be extended to experiment PCA in combination with different discriminant Analysis techniques

**REFERENCES**

[1] Shruti Sehgal, Harpreet Singh, Mohit Agarwal, V. Bhasker and Shantanu,”Data Analysis Using Principal Component Analysis”, 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 978-1-4799-5097-3/14/$31.00 ©2014 IEEE,pp45-48

[2] Asma S aleem, Ahmad Ali ,Khadim Hussain Asif and Shahid Mahmood Awan, “Pre-Processing Methods of Data Mining”, 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, 978-1-4799-7881-6/14 $31.00 © 2014 IEEE, pp451-456

[3] Trupti A. Aneyrao and R. A. Fadnavis T ,“Analysis for Data Preprocessing To Prevent Direct Discrimination in Data Mining”, 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR’16), 978-1-4673-9214-3/16/$31.00 © 2016 IEEE

[4] Lin Rui, Wang Jiamei, Li Bingze, and Yuan Changsen,,” Digital Information Mining Technology Based on Principal Component Analysis”, 2017 IEEE 3rd International Conference on Control Science and Systems Engineering, 978-1-5386-0484-7/17/$31.00 ©2017 IEEE

[5]Upadhyay, Raksha and Panse, Purvi and Soni, Ankit and Rathore Bhatt, Uma, Principal Component Analysis as a Dimensionality Reduction and Data Preprocessing Technique (April 2, 2019). Proceedings of Recent Advances in Interdisciplinary Trends in Engineering & Applications (RAITEA) 2019. Available at SSRN :  <https://ssrn.com/abstract=3364221> or [http://dx.doi.org/10.2139/ssrn.3364221](https://dx.doi.org/10.2139/ssrn.3364221)

[6] Rekh Ram Janghel , Rekh Ram Janghel , Rahul Kala and A. Shukla ,“Breast Cancer Data Prediction by Dimensionality Reduction Using PCA and Adaptive Neuro Evolution”, International Journal of Information Systems and Social Change, 3(1), 1-9, January-March 2012, DOI: 10.4018/jissc.2012010101

[7] Kaya, D., & Turk, M. (2020). LabVIEW based robust cascade predictive model for evaluating cancer prognosis. Physica A: Statistical Mechanics and Its Applications, 123978. doi:10.1016/j.physa.2019.123978

[8] Muduli, D., Dash, R., & Majhi, B. (2020). Automated breast cancer detection in digital mammograms: A moth flame optimization based ELM approach. Biomedical Signal Processing and Control, 59, 101912. doi:10.1016/j.bspc.2020.101912

[9] Abdullah Farid, A.; Selim, G.; Khater, H. A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm For Breast Cancer Detection. Preprints 2020, 2020030298 (doi: 10.20944/preprints202003.0298.v1).