|| Volume 3 || Issue 6 || June 2018 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Reducing Number of Parameters for Identifying Breast Cancer

Ravindra Bachate[1] , Supriya Dumbare[2], Gayatri Ladhe[3], Snehal Binawade[4], Srushti Gujar[5]

Assistant Professor, Department of Computer Engineering, JSCOE, Maharashtra, India[1]

UG Student, Department of Computer Engineering, JSCOE, Maharashtra, India[2,3,4,5]

ABSTRACT: Cancer is the major problem today, breast cancer is one of the leading cause of the death among the women. Here we are proposing a system which will be used by the medical experts, using data analytics techniques. The proposed system reduces the number of parameters of the fine needle aspiration test. Here we are using R programming language for the statistical analysis purpose. The proposed system consists of these major steps: the data set is loaded into the main memory, then it is pre-processed by using PCA (i.e. Principle Component Analysis) algorithm, then data is analysed and finally machine learning algorithms (LDA, RF model) are applied. There are total 30 number of parameters in the fine needle aspiration test, and we successfully reduce it to 20.finally the result is displayed using visual analytics tool.

***Keyword:* -** Principal component analysis *(PCA), Linear Discriminant Analysis (LDA), Random*

I INTRODUCTION

Diseases such as cancer, are the leading cause of death worldwide [1]. Breast cancer is the most common cancer among women in the world. Medical doctors and researchers in bio-medicine are increasingly confronted with complex patient data, posing new and difficult analysis challenges. One of the grand future challenges of biomedical informatics research is to gain knowledge from complex high-dimensional datasets[2]. Within such data, relevant and interesting structural and/or temporal patterns (‘‘knowledge’’) are often hidden and not accessible to domain experts.

In this paper we are working on a data set which contains the records of the cancer patients. This records are obtained by fine needle aspiration test. Firstly the data is cleaned. The raw data is not linear in nature. We make it linear by making use of PCA (Principal component analysis) and LDA(Linear Discriminant Analysis) Algorithm. Then the data is partitioned into two halves , training and testing data set, after that some machine learning algorithms are applied. And finally we reduce the number of parameters of the test.

II BACKGROUND AND RELATED WORK

This section provides relevant related work and background information. Firstly we will discuss the importance of data analytics

2.1 Data analytics

Data Analytics refers to qualitative and quantitative techniques and processes that are used to increase productivity and business profit. "To identify and analyze behavioral data, techniques and patterns, data is removed, accepted, and is divided, can be dynamic according to the requirement or requirement of a particular business [3].

Global organization collects and analyzes data related to customers, business processes, market economics or practical experience. Data is classified, stored and analyzed to study purchasing trends and patterns. Developed data provides complete decision-making facility.

2.2 Machine Leaning

Machine learning is a type of artificial intelligence that allow systems to learn automatically and improve from experience without being explicitly programmed. The basic assertion of machine learning is to use statistical analysis to predict an output value in an acceptable range by making use of build algorithms that can get input data.

Machine learning is closely related to computational data, which is also focused on making predictions through the use of algorithms. It has strong links with mathematical optimization, which provides methods, principles and application domains in the field.

Machine learning algorithms are often classified as supervised or unsupervised. In addition to submitting feedback to the supervised algorithm about the accuracy of prediction during training, humans need to provide both input and output. Once the training is completed, the algorithm will be applied which has learned new data. Unsupervised algorithms need not be trained with the desired result data. Instead, they use a recapitulation approach called deep education to review the data and to reach conclusions.

2.3 Visual analytics

Visual Analytics is a type of query that displays insight data to solve an issue in a mutually, graphical way. This approach uses data visualization technology to help data scientists and other professionals identify the trends, patterns and relationships in the data they are working with. It is easy for non-technical users to use packaged visual analytics software tools to set and change analytical parameters and to include drag and drop options.

Visual Analytics can be seen as an integral approach combined with visual, human factors, and data analysis. Automated and visual analysis methods are combined with human and interactive tight clips to gain knowledge of information in the visual analytics process.

III SYSTEM ARCHITECTURE

This section will describe the whole working of the project. The proposed system is working on the data set of the cancer patients. This data set is the result of the fine needle aspiration test, which is performed on the breast cancer patients. The system architecture is as follows.

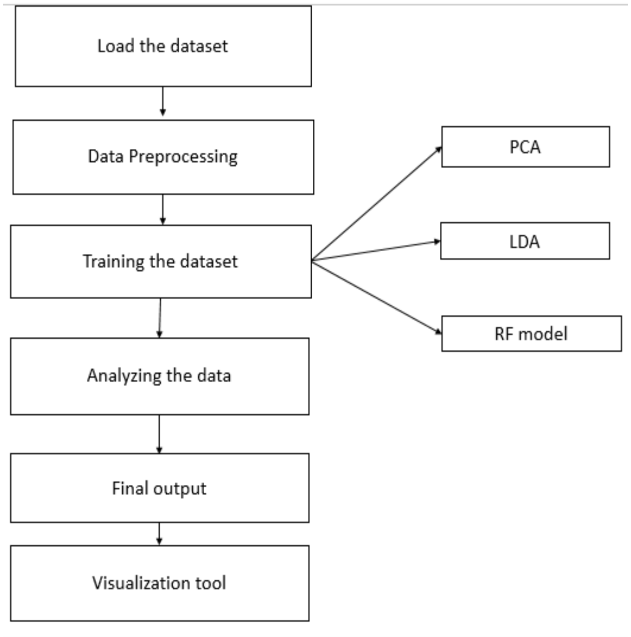


Figure -1: System Architecture

IV ALGORITHMS

In this section we will discuss the algorithms used in the project

4.1 PCA Algorithm

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

The main idea of the Principal Component Analysis (PCA) is to reduce the dimension of the data set, which includes many variables related to each other, while maintaining the diversity in the dataset, to the maximum extent, either heavy or lightly. By changing the variable into a new set of variables, which are known as principal components (or simply, PCs) and are orthogonal, it is ordered that the variation in the variables in the original variable decreases as we go down in order.[4] Therefore, the first main component maintains the maximum components that were present in the original components. The main components are the eigenvectors of a convergence matrix, and therefore they are orthogonal.

The important thing is that the dataset on which PCA technology should be used should be scaled. Results are also sensitive to relative scaling. As a common man, this is a way to summarize the data. Imagine some bottles of wine on a dining table. Each alcohol is described by virtues such as color, strength, age, etc. but unnecessary will arise as many of them will measure related properties. So in this case the PCA will do a summary of each liquor in the stock with less stock. Generally, the principal component analysis allows the user to supply the user with the least-dimensional picture, projection or "shadow" when viewed from the least informed perspective.

Dimensionality: This is the number of random variables or the number of features in a dataset, or more easily, the number of columns in your dataset.

Correlation: It shows how strongly two variable are related to each other. The value of the same ranges for -1 to +1. Positive value indicates that when one variable increases, the other increases as well, on other side negative indicate the other decreases on increasing the former. And the modulus value of indicates the strength of relation.

Orthogonal: Correlation between any pair of unrelated, ie, any pair of variables is 0.

Covariance Matrix: In this matrix, convergence is included between pairs of variables. I (i, j) th element is converse between i-th and j-th variables

Applying the PCA to 2-D Datasets

Step 1: Generalize the data

Step 2: Calculate the Converse Matrix

Step 3: Calculate eigenvalues and eigenvectors

Step 4: Selecting components and making a feature vector:

Step 5: Creating Principal Elements:

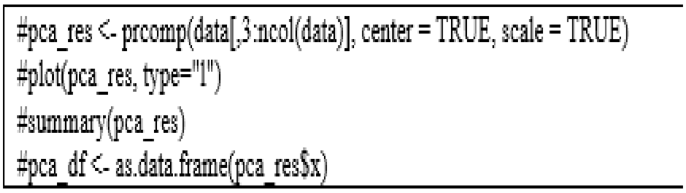
PCA in R:

In R, there are several functions from different packages that allow us to perform PCA. Here we are using prcomp() method.

PCA Using Prcomp:

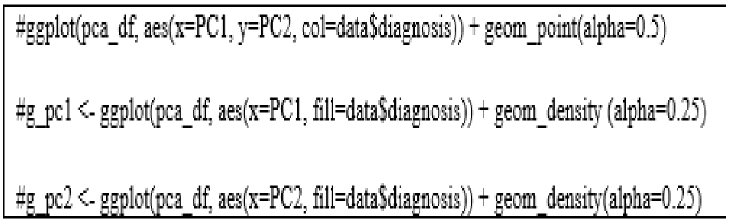
prcomp():

It creates linear combination of original data and removes corelation between the columns . this technique is called PCA(principle component analysis)



```
#pca_res <- prcomp(data[,3:ncol(data)], center = TRUE, scale = TRUE)
#plot(pca_res, type="l")
#summary(pca_res)
#pca_df <- as.data.frame(pca_res$x)
```

Figure -2: Implementation of PCA in R



```
#ggplot(pca_df, aes(x=PC1, y=PC2, col=data$diagnosis)) + geom_point(alpha=0.5)

#g_pc1 <- ggplot(pca_df, aes(x=PC1, fill=data$diagnosis)) + geom_density (alpha=0.25)

#g_pc2 <- ggplot(pca_df, aes(x=PC2, fill=data$diagnosis)) + geom_density(alpha=0.25)
```

Figure -3: PCA plots

4.2 LDA: Linear Discriminant Analysis Algorithm

Linear Differentiation Analysis (LDA) is used in the pre-processing phase for pattern-classification and machine learning applications, usually as a dimensional reduction technique. The goal is to project a square set with good class separation on a low-dimensional space so that the ordering ("curse dimension") can be avoided and also to reduce computational cost.[5]

There are two types of LDA technique to deal with classes: class-dependent and class-independent. In the class-dependent LDA, one separate lower dimensional space is calculated for each class to project its data on it whereas, in the class independent LDA, each class will be considered as a separate class against the other classes [6]

In this there are statistical properties of your data, it is calculated for each class. For an input variable (x) it is the mean and variation of the variable for each class. For multiple variables, this is the same property which is based on multivariate Gaussian, i.e. instrument and convergence matrix LDA makes some simplifying assumptions about your data:

1. Your data is Gaussian that each variable is shaped like a bell curve when plotted.
2. Each attribute has the same variance that values of each variable vary around the mean by the same amount on average.

With these assumptions, the LDA model estimates the mean and variance from your data for each class

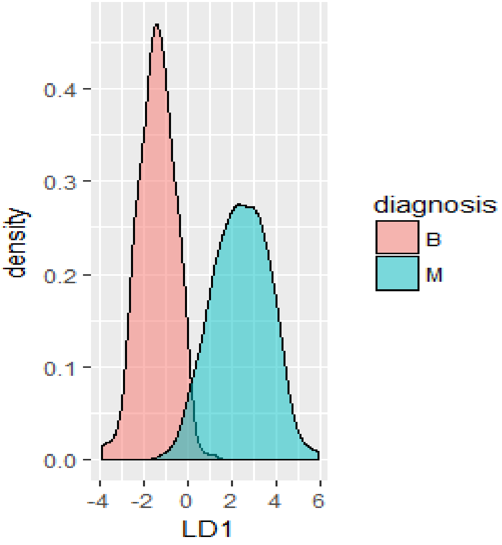


Figure -4: Result of LDA

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two techniques used for data classification and dimensional reduction The main difference between LDA and PCA is that PCA does feature classification and LDA data classification. In PCA, the size and location of the original data set are changed when converted to a different place, while the LDA space does not change, but only tries to separate the class and make the decision area between the classes.

4.3 RF model: Random Forest Model

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm generates the forest with a number of trees.

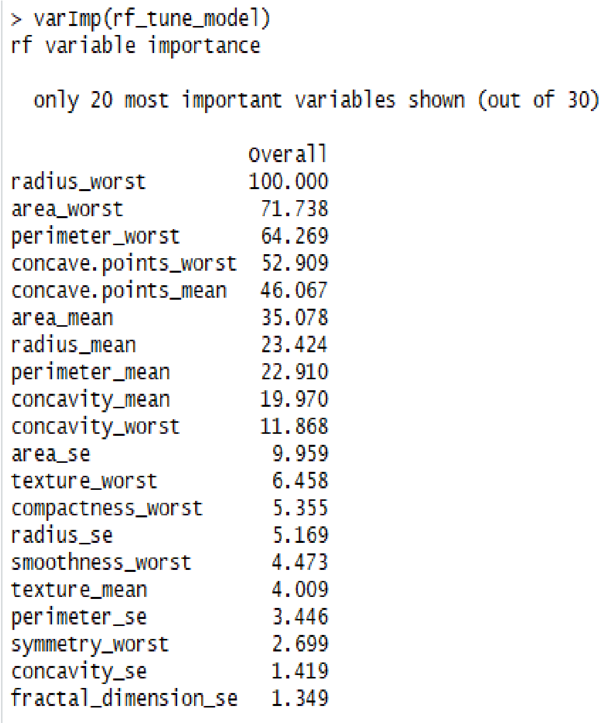
Generally, more trees in the forest look like a jungle. Similarly, in the random forest classification, the higher number of trees in the forest gives results of high accuracy.

Each tree is grown as follows:

1. **Random Record Selection :** Each tree is trained on roughly 2/3rd of the total training data. Cases are drawn at **random with replacement** from the original data. This sample will be the training set for growing the tree.
2. **Random Variable Selection :** Some predictor variables (say, x) are selected at **random** out of all the predictor variables and the best split on these m is used to split the node.(By default, x is square root of the total number of all predictors for classification)
3. For each tree, using the leftover data, calculate the misclassification rate **- out of bag (OOB) error rate.** Aggregate error from all trees to determine **overall OOB error rate** for the classification.
4. Each tree gives a classification on leftover data (OOB), and we say the tree "votes" for that class. The forest chooses the classification having the most votes over all the trees in the forest

Out-of-Bag is equivalent to validation or test data. In random forests, there is no need for a separate test set to validate result. It is estimated internally, during the run.

V RESULTS



```
> varImp(rf_tune_model)
rf variable importance

  only 20 most important variables shown (out of 30)

                     Overall
radius_worst         100.000
area_worst            71.738
perimeter_worst       64.269
concave.points_worst  52.909
concave.points_mean   46.067
area_mean             35.078
radius_mean           23.424
perimeter_mean        22.910
concavity_mean        19.970
concavity_worst       11.868
area_se                9.959
texture_worst          6.458
compactness_worst      5.355
radius_se              5.169
smoothness_worst       4.473
texture_mean           4.009
perimeter_se           3.446
symmetry_worst         2.699
concavity_se           1.419
fractal_dimension_se   1.349
```

Figure -5: 20 important parameters out of 30

VI CONCLUSIONS AND FUTURE SCOPE

Breast cancer is becoming a big issue to women health. By this project we have reduces parameters of breast cancer which increases treatment efficiency and can be easily find out whether the patient is anti-estrogenic or not.

In this paper three machine learning methods were compared to suggest one best that can have better performance than others when considering breast cancer data analysis. It is observed that among the three techniques compared namely ‘PCA’, ‘LDA’ and ‘randomforest’, ‘Randomforest’ is the best based on the performance measure precision which distinctly differs from the rest of measures used. The method of ‘randomforest’ gives us 95% accuracy, hence we conclude saying that usage of ‘Randomforest’ method best suits cancer analytics.

As a future scope we shall be exploring the other types of cancer detection mechanism.

This model can be used for different kind of cancers to reduce the number of parameters in a single test

REFERENCES

[1]. B. Novakovic, J. Jovicic, N. Milic, F. Jusupovic, M. Grujicic, and D. Djuric, “Nutrition care process in cancer,” HealthMED Journal, vol. 4, no. 2, pp. 427-433, 2010.

[2]. Holzinger A (2014) Biomedical informatics: discovering knowledge in big data. Springer, Heidelberg

[3]. IJARIIE-ISSN (2017): Understanding tools and applications of Data Analytics

[4]. E. Barshan, A. Ghodsi, Z. Azimifar and M.Z. Jahromi, Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, Pattern Recognition 44(7) (2011), 1357–1371.

[5]. A. Sharma and K.K. Paliwal, Cancer classification by gradient lda technique using microarray gene expression data, Data & Knowledge Engineering 66(2) (2008), 338–347. doi:10.1016/ j.datak.2008.04.004.

[6]. R. Haeb-Umbach and H. Ney, Linear discriminant analysis for improved large vocabulary continuous speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (1992), Vol. 1, IEEE, 1992, pp. 13–16.

[7]. P. Viszlay, M. Lojka and J. Juhár, Class-dependent twodimensional linear discriminant analysis using two-pass recognition strategy, in: Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), IEEE, 2014, pp. 1796– 1800.

[8]. Pervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, Pattern Recognition 44(7) (2011), 1357–1371.

[9]. E. Barshan, A. Ghodsi, Z. Azimifar and M.Z. Jahromi, Supervised principal component analysis: Visualization, classification and regression on subspaces and sub manifolds, Pattern Recognition 44(7) (2011), 1357–1371.

[10]. E. Barshan, A. Ghodsi, Z. Azimifar and M.Z. Jahromi, Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, Pattern Recognition 44(7) (2011), 1357–1371.

[11]. E. Barshan, A. Ghodsi, Z. Azimifar and M.Z. Jahromi, Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, Pattern Recognition 44(7) (2011), 1357–1371.

[12]. D.R. Umesh, and B. Ramachandra, “Predicting Breast Cancer Recurrence Using Machine Learning Techniques”, International Journal of Latest Trends in Engineering and Technology, vol. 5(1), pp. 255-259, 2015.

[13]. V. Chaurasia, and S. Pal, “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability”, International Journal of Computer Science and Mobile Computing, vol. 3(1), pp. 0-22, 2014.