ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/82085, August 2016* ISSN (Online) : 0974-5645

Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns

R. Vidya[1*] and G. M. Nasira[2]

[1]Department of Computer Science, MS University, Tirunelveli - 627012, Tamil Nadu, India; vidya.sjc@gmail.com
[2]Department of Computer Science, Chikkanna Government College, Tirupur – 641602, Tamil Nadu, India; nasiragm99@yahoo.com

Abstract

Background/Objective: Cervical Cancer is one among the most vulnerable and highly affected diseases among women around the World. Normally, cells grow and divide to produce more cells only when the body needs them. This orderly process helps to keep dividing when new cells are not needed. These cells may form a mass of extra tissue called a growth of tumor. Tumors can be classified as Benign or Malignant. First Benign Tumors are not cancer. They can usually be removed, and in most cases, they do not show up. Most important, the cells in benign tumors do not spread to other parts of the body. Second, malignant tumors are cancer cells. These tumors can damage nearby tissues and organs. Malignant tumors are threat to life. In this research work, prediction of normal cervix or Cancer cervix is determined with the aid of Powerful Data Mining algorithms. **Methods/Statistical Analysis:** In this research work, prediction of normal cervix or Cancer cervix is determined with the aid of Powerful Data Mining algorithms. Data mining plays an indispensable role in prediction especially in medical field. Using this concept, Classification and Regression Tree algorithm, Random Forest Tree algorithm and RFT with K-means learning for prediction of normal cervix or cancer cervix is introduced. Collection of data from NCBI (National center for Bio-technology Information) in our work, we used the data set that contains 500 records and 61 variables (i.e. Biopsy numerical value with gene identifier). The output has been presented in the form of prediction tree format. As stated, we selected a sample of 100 records with 61 biopsy features. Based on this biopsy data, an awareness program is conducted and survey is followed up to identify the changes of women during this transition period. To collect data efficiently, a Personal Interview program was conducted among rural women in various places. Collaboration with JIPMER hospital people were checked up for the test of cervical cancer. The results obtained through biopsy test were put through statistical analysis and was given through MATLAB for algorithm testing. To ascertain the results obtained are segregated and delivered in various heads with 100 test data and 60 training data. **Findings:** Comparison of the performance of various algorithms was used under the techniques in terms of sensitivity, specificity and accuracy to determine the best predictor for the cervical cancer. At first, Regression tree algorithm methodology was used for prediction. The CART binary tree yields two results, either normal cervix or cancer cervix. A Splitting Criterion called GINI index is used to identify the diversity that exists in cervical data. RFT validated optimal accuracy, a new logic was applied i.e. "combinations of two algorithms" is used. It is also an ensemble supervised machine learning algorithm. The process of whitening is used as a pre-process in k-means clustering, to get the best prediction result. The result showed the 83.87% accuracy with CART TREE output. Random Forest Tree (RFT) is used to improve the prediction accuracy. With MATLAB Coding we achieved 93.54% of prediction accuracy. The K-Means algorithm is considered efficient for processing huge datasets and hence a high accuracy of 96.77% is achieved with RFT - K-MEAN LEARNING TREE output. The Randomization of Algorithm is presented in two ways: 1. Bagging for random bootstrap sampling and 2. Input attributes are selected at random for decision tree generation. This creates an unbiased estimate of generalization error as growing of tree into forest progressed and the derivation of time complexity of K-Means is achieved. **Applications/Improvements:** Cervical cancer diagnosis and prognosis are two medical applications which pose a great challenge to the researchers.

** Author for correspondence*

The algorithms optimize a cost function defined on the Euclidean distance measure between the data points and means of cluster. Combination of RFT with K-means algorithm is the novelty of our research work, where we have achieved high accuracy result. Accurate prediction of occurrence of cervical cancer has been the most challenging and toughest task in medical data mining because of the non-availability of proper dataset. Many researchers have been done to develop different techniques that can solve problems and improve the prediction accuracy of cervical cancer through images. But in our research work, the prediction of cervical cancer is with Numerical Data. NCBI (National Center for Biotechnology Information) data set has been used. This research paper is a boon to create expert medical decision making systems and a solution for medical practitioners to construct an optimal prediction model for Cervical Cancer Prediction.

Keywords: Cervical Cancer, CART, Data Mining, Hereditary Pattern, K-Mean, RFT

1. Introduction

This research work focuses on a particular type of cancer called 'Cervical Cancer'[1]. Cervical cancer is the prominent form of cancer worldwide and ranks as the first common cancer among women in India. The body is made up of many types of cells. Normally, cells grow and divide to produce more cells only when the body needs them. This orderly process helps to keep dividing when new cells are not needed. These cells may form a mass of extra tissue called a growth of tumor[2]. Tumors can be classified as Benign or Malignant.

There are four methods to diagnose cervical cancer. First is PAP-SMEAR TEST-A basic test that describes whether a person is having cancer or not[3]. Second, HPV TEST identifies the presence of HPV (Human Papilloma Virus) that can create genital Warts at an early stage. Third type is COLPOSCOPE TEST-A to confirm the presence of normal or abnormal cells in a women cervix or vagina. Fourth type is BIOPSY - It is the removal of a small amount of tissue for examination under the microscope.

There are two types of tumors; first Benign Tumors are not cancer. They can usually be removed, and in most cases, they do not show up. Most important, the cells in benign tumors do not spread to other parts of the body. Second, MALIGNANT TUMORS are cancer cells. These tumors can damage nearby tissues and organs. Malignant tumors are threat to life. In this research work, prediction of normal cervix or Cancer cervix is determined with the aid of Powerful Data Mining algorithms that is an umbrella term that helps the researcher to identify patterns in data and obtain optimal predictions in medical field[4].

2. Research Empirical Background

The main objective of the work is prediction of "Benign" that is non- Cancerous or a "Malignant" [5], that is cancerous using various data mining algorithms. Data mining [6] plays an indispensable role in prediction especially in medical field. Using this concept, Classification and Regression Tree algorithm, Random Forest Tree algorithm and RFT with K-means learning for prediction of normal cervix or cancer cervix is introduced. The major objectives of this work are enumerated as follows:

Developing various machine learning techniques to predict cervical cancer and non-cancerous from the data set[7]. Comparison of the performance of various algorithms used under the techniques in terms of sensitivity, specificity and accuracy to determine the best predictor for the cervical cancer[8]. Collection of data from NCBI (National Center for Bio-technology Information).

The diagnosis and prognosis of cervical cancer are considered as two challenging medical applications for the researchers. The optimal solution is expected to be obtained through data mining techniques failing by which machine learning algorithms can be hybridized to it. Hence a new dimension of research is yet to be proclaimed to give optimal solution for cervical cancer prediction.

In order to achieve and prove my research work, collection of data work plays a vital role. We have collected the data set in NCBI[9]. In our work, we used the data set that contains 500 records and 61 variables (i.e. Biopsy numerical value with gene identifier). At first, Regression tree algorithm methodology was used for prediction[10]. The result showed the 83.87% accuracy with CART TREE output. The binary tree yields two results, either normal cervix or cancer cervix. There were some limitations in CART Algorithm.

Hence, the next algorithm namely Random Forest Tree (RFT) is used to improve the prediction accuracy. With MATLAB Coding we achieved 93.54% prediction accuracy with RFT TREE output. To substantiate high accuracy rate, a new logic was applied i.e. "combinations

of two algorithms" is used. The k-means based methods are efficient for processing the large data sets, thus very attractive for data mining. With k-means learning and RFT, high accuracy of 96.77% is achieved with RFT-K-Mean Learning Tree output. The process of whitening is used as a pre-process in k-means clustering, to get the best prediction result[10].

3. Proposed System

The Major objective of this proposed research work is to construct decision tree using Classification and Regression Tree (CART) and Random Forest Tree (RFT) with numerical datasets. If the result is not identified properly, a hybridized algorithm[11] will be created with machine learning algorithms to obtain optimal prediction for cervical cancer. NCBI Data set for numerical data origin (original) is used. The output has been presented in the form of prediction tree format.

3.1 Prediction with Cart Algorithm

Classification and Regression Tree (CART) helps in predicting continuous dependent and predictor variables by predicting the most likely value of the independent variable. The CART algorithm splits the entire dataset into subsets recursively based on similar values of the dataset. The decision tree gets growing based on rules and decisions in search of the optimal objective. It always considers the optimal split from the classification to predict normal from cancer cervix.

3.2 Algorithm for Prediction using Tree Generation

Classification and Prediction can be done with utmost efficiency using decision trees. The decision trees represent rules to achieve refined classification of datasets[12]. Rules and Rule sets can be expressed easily such that the human understanding is much better with all database applications. In few applications, accuracy of prediction is given much priority due to the sensitiveness of the result to be achieved. One such area is the cervical cancer prediction as even slight misplaced result might be costly for the life of a Woman. The implementation of CART algorithm is given below:

Input : NCBI dataset – numerical biopsy values
Output : prediction with Regression Tree output
Begin
Start at the root node (t=1)
For each X, find the set S
Training cases be *N*
Number of variables in the classifier be *M*
m of input variables to determine decision
IF
Min the node impurities (Sum)
Select 2 child nodes
choosing *n* times with replacement from all *N*
Choose *m* variables on which to base the decision at that node.
Calculate the best split based on these *m* variables in the training set
Choose the split $\{X^* \in S^*\}$ min overall X and S
Search for a split s^* among the set if all possible candidates s that gives the purest decrease in impurity.
Split node 1 (t = 1) into two nodes (t = 2, t = 3) using the split s^*
Repeat the split search process (t = 2, t = 3) as indicated until the tree growing the tree growing rules are met.
Else
Stopping criterion, exit
Estimate the error of the tree, by predicting their classes.
Each tree is fully grown and not pruned
End

The Splitting of Tree is carried out by a single variable which is responsible to generate each step to maximize class purity within two resulting subsets. The subset is further split based on the initial variable based on independent variables[13]. The steps for creating tree has various paths like creating node, selecting leaf without error, creating subordinate node, replacing leaf with created node and repeating the former steps again and again.

3.3 Steps involved in CART Tree Generation

CART being a recursive method for classification and regression of trees by predicting dependent variables and categorical predictor variables. The tree classifiers are introduced to the algorithm and presented with the classification Trees Analysis facilities discussing the same information. The Steps for working of CART are furnished below:

3.3.1 Splitting Criteria

CART algorithm uses GINI Index to decide the attributes to be selected. The result should be to choose the attribute that has minimum GINI Index after splitting. CART uses GINI diversity index ahead of information criteria as a splitting criterion for better accuracy. The CART algorithm supports the GINI Index diversity[14,15] to gain more information and give the optimal result of prediction. The GINI index searches the best features at each internal node and then creates the decision. CART can also provide towing splitting criterion for multiclass problems. The classes are separated at each node into two super classes that contains mutually exhaustive and disjoint classes. The splitting criterion can find right attribute to optimize two super classes criterion.

In this Research Work, methodology is carried out with CART tree containing 500 records where 100 records with 60 training set and 40 testing set are used at first place. The CART algorithm is applied to NCBI dataset to find the root and leaf of the tree using GINI Index splitting criteria. Same operation is performed for each attribute until last split and decision tree is created. After every leaf storing the continuous-valued prediction, the average value for the training tuples is created until it reaches the destination target of prediction. Pruning method is used on the decision tree for better accuracy. The CART is implemented and achieved a good prediction rate of 83.37% by writing code in the MATLAB[16].

4. RFT (Random Forest Tree)

Random Forest Tree achieves the predicted output by integrating independently distributed vectors of random type contained as a collection of tree-structured classifiers. Data Mining will increase the outcome of prediction of diseases where huge dataset is involved. The special feature of the RFT algorithm is that it can be associated with Machine Learning Algorithms to increase the prediction result. Random Forest Tree (RFT) Technique

Step 1: Decision rules - partition sample of data
Step 2: Terminal node (leaf) indicates the class assignment
Step 3: Tree partitions samples into mutually exclusive groups
Step 4: One group for each terminal node
Step 5: All paths
Step 6: Start at the root node
Step 7: End at a leaf
Step 8: Each path represents a decision rule
Step 9: Joining (AND) of all the tests along that path
Step 10: Separate paths that result in the same class are disjunctions (ORs)
Step 11: All paths - mutually exclusive
Step 12: For any one case - only one path will be followed
Step 13: False decisions on the left branch
Step 14: True decisions on the right branch

generates multiple decision trees where randomization is used in two ways:

- Random Sampling of Data for bootstrap samples like bagging.
- Random selection of input attributes for generating individual base decision trees.

RFT can handle huge input dataset and can select right variable[17] to get the desired output. This also generates unbiased internal estimation of generalization error as growing of forest progressed. It is considered as an effective method for prediction with high accuracy. RFT principle is based on bagging where bootstrap samples are generated for induction of each decision tree. Attribute selection is considered as yet another source of randomization.

4.1 Prediction

To make a prediction for a query point x, each tree independently predicts

$$f_n^i(x) - \frac{1}{N^e(A_n(x))} \sum_{y_i \in A_n(x) I_i = e} Y_i \qquad (1)$$

Where, F_n is the sequence of the estimator based on (x) auxiliary variable. $A_n(x)$ denotes the leaf containing x and $N^e(An^{(x)})$ denotes the number of estimation points it contains. N represents the sequence of classifiers. e embodies the value of risk functional, i the number of sequence. Y signifies the randomness in the tree construction. The predictions made by each tree depend only on the estimation points in that tree.

4.2 RFT Algorithm

Decision trees are a popular method for various machine learning tasks. Tree learning comes closest to meeting the requirements for serving as an off-the-shelf procedure for data mining, because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate. Random forest (or random forests) is an ensemble classifier[17] that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The steps involved in RFT is shown below

Input : NCBI Data Set
Output: Prediction result with RFT tree output
Begin
For K=1 to c do
Draw a bootstrap data of size N from the training data
N-records are sampled at random but with replacement, from the original data
Grow a tree T using bootstrapped data
M-input variables
m<<M, a number m-is selected such that at each node
m-variables are selected at random out of M
Repeat at each node
Randomly select x out of p predictor variables
Pick out from the x predictor variables
The best split on these m attributes is used to split the node
Spilt the node into two daughter nodes
Value of m is held constant during forest growing
Until some stopping condition for splitting the tree further is satisfied
There are two tuning parameters for these classifiers
a. C- The number of trees used to build the classifier
b. x- The number of particular variables selected for splitting each node in the tree (x=p)
Each tree is grown to the largest extent possible
There is no pruning
End

4.3 Methodology using in RFT Algorithm

Random Forest Tree (RFT) method is generally used to handle two problems, one to construct a prediction rule in supervised learning problem and two to assess and rank the variables along with their ability to predict the response and is done by considering Variable Importance Measures (VIMs) which are computed automatically for each predictor within RFT Algorithm. The Steps involved are:

- Random forest is a combined tree predictor mechanism where each tree depends on the values of a random vector sample in an independent manner with equal distribution.
- RFT Algorithm is applied on NCBI dataset to find the root and leaf of the tree.
- Decide the best split among each node in the Tree.
- Use Estimation Points to make a prediction for RFT.

$$f_n^i(x) - \frac{1}{N^e(A_n(x))} \sum_{y_i \in A_n(x) I_i = e} Y_i \qquad (2)$$

- Using 2013 MATLAB code, for sample 100 records, we have achieved 93.54% prediction accuracy with RFT tree output contain either normal cervix or cancer cervix.

Dataset is a collection of data. Dataset of the column description of GDS3233 that has been used is given in Figure 1. The Column Description GDS3233 along with column name, column ID and identifier.

The dataset was retrieved from NCBI (National center for biotechnology information), GEO (Gene Expression Omnibus) database, record #GDS3233 it represents the pre invasive and invasive squamous cells carcinoma. This dataset is used to identify gene expression profiles in cervical cancer. Columbia university medical center published this dataset from pathology department at cancer cytogenetic lab.

- **GSM (Geo Sample) DNA – One Gene Chip**
  A DNA Microarray or biochip[18] is identified as an organization of Micro-DNA spots attached to a solid surface. Microarray DNAs are used by scientists to measure the levels of expression of genes in large number at the same time.
- **Column Name**
  Example**:** mRNA values…..GSM 246087, GSM 246088, GSM 246089, and GSM 246090 One GSM is one GENE CHIP (Data from the use of single chip). For each gene there will be multiple score including the main one, held in the column "VALUE".
- **Identifier-gene symbol**
- For example**:** DDR1, RFC2, HSPA6. A gene's unique observation consists of italicized uppercase Latin Letters and Arabic Numbers formally assigned by HUGO Gene Nomenclature committee after gene identification.
  Colm ID = Prob ID
  Eg: 1007-s-at, 1053-at, 117-at.
- **Probe**
  Explore or Examine especially with hands or an instrument and mapping Probe do Affymetrix Feature to Gene.

S.No	COLUMN DESCRIPTION GDS3233	COLUMN-NAME	COLM-ID	IDENTIFIER
1	'GSM246087 = Value for GSM246087: Cervical cancer cell line, C4-I; src: Cervical cancer cell line, C4-I'	'GSM246087'	'1007_s_at'	'DDR1'
2	'GSM246088 = Value for GSM246088: Cervical cancer cell line, CaSki; src: Cervical cancer cell line, CaSki'	'GSM246088'	'1053_at'	'RFC2'
3	'GSM246089 = Value for GSM246089: Cervical cancer cell line, C-33A; src: Cervical cancer cell line, C-33A'	'GSM246089'	'117_at'	'HSPA6'
4	'GSM246090 = Value for GSM246090: Cervical cancer cell line, HT-3; src: Cervical cancer cell line, HT-3'	'GSM246090'	'121_at'	'PAX8'
5	'GSM246119 = Value for GSM246119: Cervical cancer cell line, SiHa; src: Cervical cancer cell line, SiHa'	'GSM246119'	'1255_g_at'	'GUCA1A'
6	'GSM246120 = Value for GSM246120: Cervical cancer cell line, SW756; src: Cervical cancer cell line, SW756'	'GSM246120'	'1294_at'	'UBA7'
7	'GSM246121 = Value for GSM246121: Cervical cancer cell line, MS751; src: Cervical cancer cell line, MS751'	'GSM246121'	'1316_at'	'THRA'
8	'GSM246122 = Value for GSM246122: Cervical cancer cell line, ME-180; src: Cervical cancer cell line, ME-180'	'GSM246122'	'1320_at'	'PTPN21'
9	'GSM246123 = Value for GSM246123: Cervical cancer cell line, HeLa; src: Cervical cancer cell line, HeLa'	'GSM246123'	'1405_i_at'	'CCL5'
10	'GSM246422 = Value for GSM246422: Normal cervix, commercial_Ambion; src: Normal cervix, commercial_Ambion'	'GSM246422'	'1431_at'	'CYP2E1'
11	'GSM246423 = Value for GSM246423: Normal cervix, commercial_Stratagene; src: Normal cervix, commercial_Stratagene'	'GSM246423'	'1438_at'	'EPHB3'
12	'GSM246484 = Value for GSM246484: Normal cervix, commercial_BioChain; src: Normal cervix, commercial_BioChain'	'GSM246484'	'1487_at'	'ESRRA'
13	'GSM246485 = Value for GSM246485: Normal cervix epithelium_CaCX3; src: Normal cervix epithelium_CaCX3'	'GSM246485'	'1494_f_at'	'CYP2A6'

Figure 1. Column Description GDS323.

Table 1. Comparison between different GSM ID

GSM ID	Probe ID	Gene Symbol	Description
GSM246087	1007_s_at	DDR1	Discoid in domain receptor tyrosine kinase 1
GSM246422	1431_at	PTPN21	Protein Tyrosine Phosphatase, Non- Receptor Type 21
GSM246485	1494_f_at	CYP2A6	Cytochrome P450 2A6
GSM247164	20001_at	CAPNS1	Calpain, Small Subunit 1

The Column name along with its Probe ID and Gene Symbol is given in Table 1. The GDS3233 is used for analysis of Cervical Cancer (CC) primary tumors and cell lines Chromosomal amplifications are a common cellular mechanism of gene activation in tumor genesis. Chromosome 20 is a commonly gained chromosome in cervical cancer. Results provide insight in the potential role of chromosome 20 gain in cervical cancer progression. The RNA GSM246087 COLUMN numerical value is its TPM (Transcription Per Minute)[19] value for its feature values. Transcription is a process of making RNA from a DNA template. It normalizes to transcript copies instead of reads and corrects for cases where the average transcript length differs between samples. For example, the RNA GSM246087 is given in Table 2 and Table 3 describes the line description long with different GSM ID characteristics.

Table 2. The RNA GSM246087

S.No	Feature	TPM-Value
1	1007_s_at	1353.7
2	1053_s_at	135.8
3	117_at	3.3
4	121_at	393.2

5. RFT with K-MEANS Learning

Cervical carcinoma still continues to be the most common cancer among women and accounts for the maximum deaths each year. Persistent infections with High- Risk (HR) Human Papilloma Virus (HPV), Such as HPV 16, 18, 31, 33 and 45 have been identified as a major development of the disease.

5.1 K-MEANs Algorithm

K-Means clustering method of Vector quantization is a form of signal processing that can be used to classify and analyses the clusters in data mining. K-Means[20] partitions 'N' observations into 'K' clusters where every observation is paired with the nearest means that serves as a prototype of the cluster to be identified. The ultimate aim is to identify 'K' centers in C1, C2, ….Ck.

5.2 Algorithm: Construction of K-Mean with RFT Tree

When K-means is constructed, the RF predictors might lead to a dissimilarity measure between observations between unlabeled data. The main idea is to construct RF predictor to distinguish observed clinical data from generated datasets. RF can handle homogeneous and heterogeneous variables[21] much easily than other variables. The steps involved in the hybridization of RFT and K-Means algorithm is given below:

Table 3. Describes the line description long with different GSM ID characteristics

GSM ID	Title	Source name	Description	Characteristics
GSM246087	Cervical cancer cell line, C4-I	Cervical cancer cell line, C4-I	Cervical cancer cell line C4-I', 'squamous cell cancer'	cervical cancer cell line
GSM246422	Normal cervix, commercial_Ambion	Normal cervix, commercial_Ambion	Commercial RNA from normal cervix	Commercial RNA from cervix
GSM246485	Normal cervix epithelium_CaCX3	Normal cervix epithelium_CaCX3	Normal cervix, age 27 years, micro-dissected squamous epithelium	Microdissected_Normal Cervical epithelium
GSM247164	Normal cervix_03-5611	Normal cervix_03-5611	Normal cervix epithelium, microdissected, Age 44	Normal Cervix Epithelium, microdissected

Input : NCBI Data Set
Output: Prediction result K-Mean with RFT tree output
Begin
For K=1 to c do
Draw a bootstrap data of size N from the training data
N-records are sampled at random but with replacement, from the original data
Grow a tree T using bootstrapped data
M-input variables
The K-mean algorithm is formed with the following operations
Choose initial cluster centers Z1, Z2, ...Zk randomly from the n points w1, w2, wN, wI ∈ Rm
Assign point Wi, I = 1, 2, ..., N to cluster Cj = 1, 2, ..., K if and only if ||Wq - Zq, p = 1, 2, ..., K, and J ≠ P ties are resolved arbitrarily.
Compute the new cluster centers Z1∗, Z2∗,...,ZK∗, as follows

$$Zi* = (I/n) \Sigma Wj \qquad I = 1, 2, ...,K$$

Wj ∈ CJ
If Zi∗ Zi I = 1, 2, ...,K then terminate otherwise Zi - Zi∗ and go to step 10
RFT process:
m<<M, a number m-is selected such that at each node
m-variables are selected at random out of M
Repeat at each node
Randomly select x out of p predictor variables
Pick out from the x predictor variables
The best split on these m attributes is used to split the node
Spilt the node into two daughter nodes
Value of m is held constant during forest growing
Until some stopping condition for splitting the tree further is satisfied
There are two tuning parameters for these classifiers
C- The number of trees used to build the classifier
x- The number of particular variables selected for splitting each node in the tree (x=p)
Each tree is grown to the largest extent possible
There is no pruning
End

6. Result and Discussion

Time Complexity for K-Means

$O(Knt_{dist})$

K – No.Of Centroids

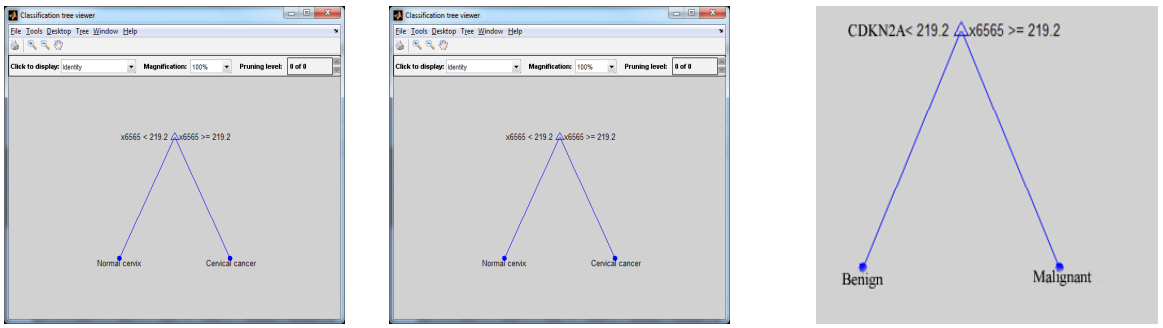
n – Number of Objects

t_{dist}- time to calculate the distance between two objects.

Data mining provide the methodology and technology to analysis the useful information of data. Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in dataset. K-Mean is a greedy algorithm can only converge to local minimum. K-Means is numerical unsupervised non- deterministic interactive method.

In the given work, data were collected from NCBI, including biopsy features. K-Means learning as a preprocessing and the result had been given to the input for RFT (hybrid). We had achieved 97.77% accuracy in prediction i.e. the novelty in our research work. This research work will promote best prediction models and strong classifiers to make effective decision making systems in the medical world.

The present study was an attempt to find out the solution for cervical cancer and give awareness to the women regarding the health issues. As stated, we selected a sample of 100 records with 61 biopsy features. On this biopsy data, a survey was carried out to find out the extent of awareness and problems faced by women during this



(a) (b) (c)

Figure 2. Tree output.

period of transition. A personal interview schedule was used to collect data for the study. Collaboration with JIPMER hospital people were checked up for the test of cervical cancer. The results obtained through biopsy test were put through statistical analysis and was given through MATLAB for algorithm testing. For the better understanding the results were divided and presented under following three heads with 100 test data and 60 training data as shown in Table 4 and tree output has been explained in Figure 2.

Table 4. Different methods and prediction

No.	Methods	Prediction in percentage
1	CART	83.87%
2	RFT	93.54%
3	RFT with K-mean	96.77%

7. Conclusion

Accurate prediction of occurrence of cervical cancer has been the most challenging and toughest task in medical data mining because of the non-availability of proper dataset. Many researchers have been done to develop different techniques that can solve problems and improve the prediction accuracy of cervical cancer through images. But in our research work, the prediction of cervical cancer is with Numerical Data. NCBI (National Center for Biotechnology Information) data set has been used. However, no such system is designed exclusively till now for cervical cancer prediction with datasets.

The research has described the prediction of cervical cancer in two stages i.e. Benign or Malignant of women with data mining algorithms, with reasonable accuracy. This dissertation describes a finite, well defined numerical dataset which is well suited for cancer prediction. This work presents various machine learning techniques for cervical cancer prediction.

The different data mining techniques i.e. CART, RFT and K-means with RFT have been investigated. Several experiments were conducted using these algorithms with MATLAB coding. The achieved prediction performances are comparable to the existing techniques. The experiments are performed with NCBI Dataset.

CART (Classification and Regression Tree) can be effectively used for the prediction of cervical cancer with NCBI dataset. Integration of algorithms can give better performance and reduce prediction errors. During this study work, the accuracy percentage of CART is 83.87% with Binary tree output.

To increase the correctness of the prediction level, RFT (Random Forest Tree) algorithm is used to predict cancer and it is classified as Benign or the accuracy level reached to the extent of 93.54%. Still the results are efficiently obtained using K-means learning algorithm which is best suited for data mining because of its efficiency in processing large data sets. The algorithms optimize a cost function defined on the Euclidean distance measure between the data points and means of cluster. Combination of RFT with K-means algorithm is the novelty of our research work, where we have achieved high accuracy result. By this process, the accuracy of RFT with K-means for cervical cancer prediction is enhanced to 96.77% while comparing these three.

8. Acknowledgement

This research was supported by Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER).

9. References

1. Wright TC, Stoler MH, Behrens CM, Sharma A, Zhang G, Wright TL. Primary cervical cancer screening with human papillomavirus: End of study results from the ATHENA study using HPV as the first-line screening test. Gynecologic Oncology. 2015; 136(2):189-97.
2. Stevanovic S, Draper LM, Langhan MM, Campbell TE, Kwong ML, Wunderlich JR, Restifo NP. Complete regression of metastatic cervical cancer after treatment with human papilloma virus–targeted tumor-infiltrating T cells. Journal of Clinical Oncology. 2015.
3. Athinarayanan S, Srinath MV. Classification of cervical cancer cells in PAP smear screening test. ICTACT Journal on Image and Video Processing. 2016; 6(4).
4. Rose PG, Java J, Whitney CW, Stehman FB, Lanciano R, Thomas GM, Di Silvestro PA. Nomograms predicting progression-free survival, overall survival and pelvic recurrence in locally advanced cervical cancer developed from an analysis of identifiable prognostic factors in patients from NRG Oncology/Gynecologic Oncology Group randomized trials of chemo-radiotherapy. Journal of Clinical Oncology. 2015; 33(19):2136-42.
5. Kuang F, Yan Z, Li H, Feng H. Diagnostic accuracy of diffusion-weighted MRI for differentiation of cervical cancer and benign cervical lesions at 3.0 T: Comparison with routine MRI and dynamic contrast-enhanced MRI. Journal of Magnetic Resonance Imaging. 2015; 42(4):1094-9.
6. Chau R, Jenkins MA, Buchanan DD, Ouakrim DA, Giles GG, Casey G, Lindor NM. Determining the familial risk distribution of colorectal cancer: A data mining approach. Familial Cancer. 2016; 15(2):241-51.
7. Athinarayanan S, Srinath MV. Classification of cervical cancer cells in PAP smear screening test. ICTACT Journal on Image and Video Processing. 2016; 6(4).
8. Narayan G, Xie D, Ishdorj G, Scotto L, Mansukhani M, Pothuri B, Murty VV. Epigenetic inactivation of TRAIL decoy receptors at 8p12-21.3 commonly deleted region confers sensitivity to Apo2L/trail-Cisplatin combination therapy in cervical cancer. Genes, Chromosomes and Cancer. 2016; 55(2):177-89.
9. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Astashyn A. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion and functional annotation. Nucleic Acids Research. 2015; 1189.
10. Belfatto A, Riboldi M, Ciardo D, Cattani F, Cecconi A, Lazzari R, Cerveri P. Kinetic models for predicting cervical cancer response to radiation therapy on individual basis using tumor regression measured in vivo with volumetric imaging. Technology in Cancer Research and Treatment. 2015; 15(1):146-58.
11. Sukumar P, Gnanamurthy RK. Computer aided detection of cervical cancer using PAP smear images based on hybrid classifier. International Journal of Applied Engineering Research. 2015; 10(8):21021-32.
12. Gertz EM, Chowdhury SA, Lee WJ, Wangsa D, Heselmeyer-Haddad K, Ried T, Schaffer AA. FISHtrees 3.0: Tumor phylogenetics using a ploidy probe. PloS One, 2016; 11(6):e0158569.
13. Yamal JM, Guillaud M, Atkinson EN, Follen M, MacAulay C, Cantor SB, Cox DD. Prediction using hierarchical data: Applications for automated detection of cervical cancer. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2015; 8(2):65-74.
14. Yamal JM, Guillaud M, Atkinson EN, Follen M, MacAulay C, Cantor SB, Cox DD. Prediction using hierarchical data: Applications for automated detection of cervical cancer. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2015; 8(2):65-74.
15. Hamad R, Rehkopf DH, Kuan KY, Cullen MR. Predicting later life health status and mortality using state-level socioeconomic characteristics in early life. SSM-Population Health. 2016; 2:269-76.
16. Pinker K, Andrzejewski P, Baltzer P, Polanec SH, Sturdza A, Georg D, Poetter R. Multiparametric [18 F] Fluorodeoxyglucose/[18 F] Fluoromisonidazole positron emission tomography/magnetic resonance imaging of locally advanced cervical cancer for the non-invasive detection of tumor heterogeneity: A pilot study. PloS One. 2016; 11(5):e0155333.
17. Bountris P, Haritou M, Pouliakis A, Karakitsos P, Koutsouris D. A decision support system based on an ensemble of random forests for improving the management of women with abnormal findings at cervical cancer screening. IEEE 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015 Aug. p. 8151-6.
18. Bhat S, Kabekkodu SP, Noronha A, Satyamoorthy K. Biological implications and therapeutic significance of DNA methylation regulated genes in cervical cancer. Biochimie. 2016; 121:298-311.
19. Nagasaka K, Taguchi A, Kawana K, Hashimoto K, Plessy C, Nakamura H, Banks L. A new approach for screening cervical cancer by characterization of transcripts using CAGE technology. In ASCO Annual Meeting Proceedings. 2015 May; 33(15):e16514.
20. Nguyen HT, Jia G, Shah ZK, Pohar K, Mortazavi A, Zynger DL, Knopp MV. Prediction of chemotherapeutic response in bladder cancer using K-means clustering of Dynamic Contrast-Enhanced (DCE)-MRI pharmacokinetic parameters. Journal of Magnetic Resonance Imaging. 2015; 41(5):1374-82.
21. Peterson L, O'Sullivan J, O'Sullivan F, Koh WJ, Swensen R, Krohn K, Rajendran J. FMISO PET uptake and spatial heterogeneity to predict response in patients with cervical cancer. Journal of Nuclear Medicine. 2015; 56(3):1175.