International Journal of Research in Advent Technology, Vol.7, No.4, April 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Cervical Cancer Test Identification Classifier Using Decision Tree Method

Dipti N. Punjani[1], Dr. Kishor H. Atkotiya[2]
[1]*Assistant Professor, National Computer College, Jamnagar.*
[2]*Professor, Department of Statistics, Saurashtra Unviersity, Rajkot.*

Abstract: Cervical cancer is one of the most crucial and important type of cancers which needs to be treated as early as possible. The most difficult task is to detect a cancer at early stage so that necessary cure can be done on time. Several medical research works have been done towards the various medical tests which cervical cancer patients should go for. The four most frequent tests are: hinselmann, schiller, citology, and biopsy. These four tests might be expensive to afford all at a time. This research work focuses on prediction of which test a patient should go first for cervical cancer identification. a real dataset of approx. 800 patients of hospital universitario de caracas' in caracas, venezuel is used for the training purpose. Decision tree method is used to build a classifier.

Keywords: Cervical Cancer, Hinselmann, Schiller, Citology, Biopsy, Decision Tree, Classification

1. INTRODUCTION

Prediction of any disease becomes very difficult when the disease itself has unpredictable symptoms at early stage. Undoubtedly, cancers are one of the most difficult diseases to detect and cure. Medical science is continuously searching for the detection process at early stage as well as cure process at last stage. A significant amount of success has been found in recent years which have reduced fear of cancers from the people. Still cancers are considered as life threatening by most of the non medical person. Not every cancer is always life threatening and even a life threatening cancer can be cured if detected and cured properly. This research work focuses on a special type of cancer called – cervical cancer. A cancer is growth of cells in an out of control way where an area of body could be affected in an unnatural way. Cervical cancer affects the cells at cervix - lower area of uterus of female. The cervix connects uterus to the vagina which is a canal through which a birth takes place. Cervix is composed of two parts having different types of cells.

Endocervix is near to uterus with glandular cells. Exocervix (Ectocervix) is a part near to the vagina with squamous cells. These two types of cells meet at transformation zone [1][2]. Figure 1 shows the structure.

Cervical cancer is one of the most crucial cancers to handle. Four most important tests related with cervical cancer are Hinselmann, Schiller, Citology, and Biopsy. The indeed question which test a patient should go for first? Which test results in most accurate cancer state detection first? Are all tests are required? The answers of these tests are tried to find out with this piece of research. The main motive is to let patients and medical person know what test is required to be done at a particular stage of diagnosis without any manual interpretation of habits and other tests reports. Doing a realistic research, enough care has been kept that the predictors must be real, easy to arrange and enough to decide which test a patient should go for accurately. A real data of 850 patients taken at Hospital Universitario de Caracas' in Caracas, Venezuel is used for training as well as testing purpose [8]. The implementation is with R.

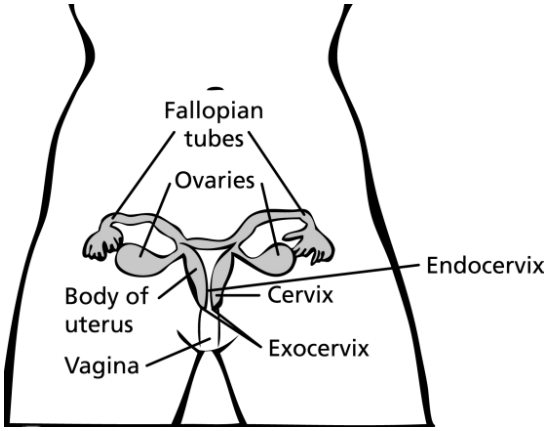

Figure 1 Cervical cancer related body parts

Section 2 discusses basis of decision tree method. Section 3 discusses the dataset. Section 4 discusses the results. The paper ends with conclusion and future directions.

2. DECISION TREE METHOD

A decision solely depends on what analytics fields we have. The accuracy depends on the logic behind the decision. What if the decision could be taken by a system instead of human? That's what the data science is all about. What if the data science is used with the medical science? The purpose becomes the collaboration of data science and medical science to take decisions more accurately and speedily. Undoubtedly decision tree method is one of the most accurate, easy and simple methods to build a classifier [3][4][5].

2.1. Decision Tree Structure

The method uses the training data to partition based on the tree construction concept. Every interior node is a condition while every exterior node represents the prediction. As discussed in section 1, this research work depends on prediction of test required for cervical cancers, the interior nodes are the patient's history related conditions and exterior nodes are the various tests. The input attributes are partitioned into two sets: Predictors – whose values are known and Class – whose value needs to be predicted. Given a set of predictors (P1, P2, P3, P4…Pn) output will be value of a class C.

The set of rules can be easily implemented and interpreted as a series of nested If….Else…. structure. Once a classifier model is build there is no need to store the training data anymore as the learning process is kind of an eager learner where the model itself is sufficient to carry the analytics [6][7].

2.2. Implementation with R

R language is used for statistical analysis, data analytics, graphical representation etc. it is freely available under the GNU General Public License. R is very easy to use with efficient data management capabilities. It has a large number of tools for data analysis. Data analysis can be in text form as well as in graphical form which makes users to understand outputs easily. R has a package called party for the usage of decision tree. The package party's ctree method is used to create a decision tree. The basic steps of defining the decision tree with R are: Set training data set in an object and then calling ctree method. ctree(f,d) where f represents the formula with information of all predictors and class variables. d represents the object with training data. ctree method returns an object which is in the form of a decision tree. This decision three can be printed or plotted for text or graphical representation point of view.

3. DATA SET

Prediction needs to be done with utmost accuracy. The first step in achieving accurate predictions is to have an accurate prediction model. An accurate prediction model can be developed only if the training data is accurate. As we are targeting predictions related with cervical cancers, accuracy is indeed requirement. The target task has a lot of unpredictable and inaccurate factors and so an accurate model needs to balance the predictions at acceptable level.

3.1. Important Properties

We have identified a list of quality in test data which is indeed required before we can proceed for model development.

- Genuine:-It must represent data of actual patients.
- Accurate:-The results of samples must be accurate and precise.
- Complete:-All types of possibilities must be covered.
- Unbiased:-Not specific to a small category.
- Large:-It should be of at least 500 patients.
- Random:-No repetition of data is allowed.
- Crisp:-All values must be crisp in nature. No vague fields.
- Real:-Must be realistic to get it filled by experts.
- Latest: - The data must be recent.

3.2. Cervical Cancer Dataset

The best is to process real data collected / experienced at real places. There is no better option than processing data given by hospitals of their patients. We have used A real data of 850 patients taken at Hospital Universitario de Caracas' in Caracas, Venezuel. The data is composed of 36 attributes which are based on patient's personal life style with habits, medical history etc. The data base tries to make it as simple as possible by asking Boolean values rather than test results in numerical forms. So that a patient himself or a medical person can use it without changing the user interface. The attributes are categorized into two groups [8].

(1) Patient's personal life style with habits: This category includes various attributes like age, number of sex partners a patient involved with, number of pregnancies, age at which 1st sexual intercourse was experienced, whether a patient smokes or not? If Yes then since how many years, whether a patient takes Contraceptives? If Yes then since ho w many years? Etc. These are the fields which are less likely to be help in predictions as compared to the fields of medical history group but yet important and so included[8].

(2) Patient's medical history: This category includes some important information about patient's medical past and present. The set of attributes like whether a patient has HIV, AIDS, Hepatitis B, any other Cancer, HPV etc. The more to the detail how many times a patient is diagnosed is also included. condylomatosis and its aligned attributes like cervical condylomatosis ,vaginal condylomatosis and vulvo-perineal condylomatosis are also included [8].

The rest is the most important field, a field which describes the most important and urgent test a patient should go for. The test is the next step towards the diagnosis of cervical cancer. Test attribute refers to the four tests: Hinselmann, Schiller, Citology, and Biopsy[8].

4. IMPLEMENTATION

4.1. Test Decision Tree

We have derived decision trees for all the four tests. The decision trees for Hinselmann, Schiller, Citology, and Biopsy are shown in Figure 2, Figure 3, Figure 4 and Figure 5 respectively. Figure 6 shows a combined classifier for all the four tests.

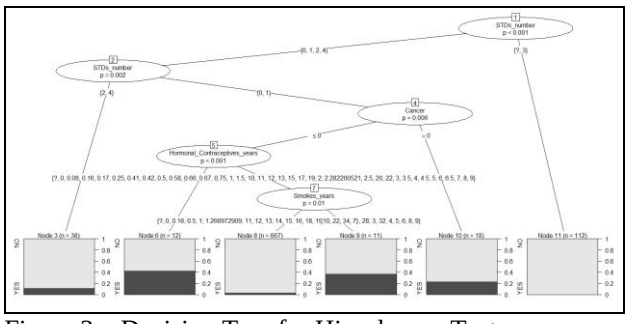

Figure 2 – Decision Tree for Hinselmann Test

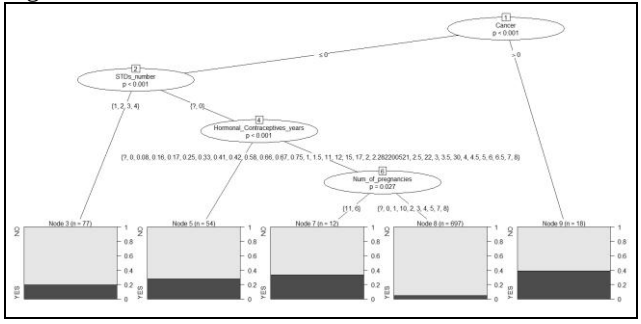

Figure 3 – Decision Tree for Schiller Test

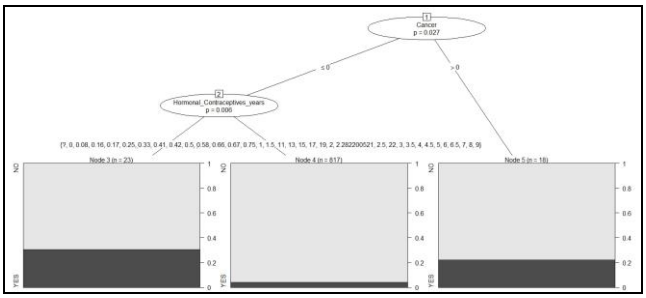

Figure 4 – Decision Tree for Citology Test

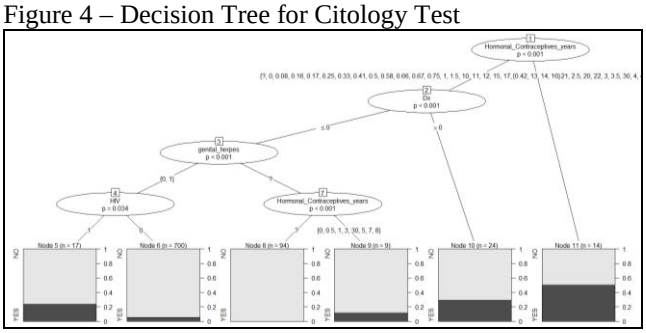

Figure 5 – Decision Tree for Biopsy Test

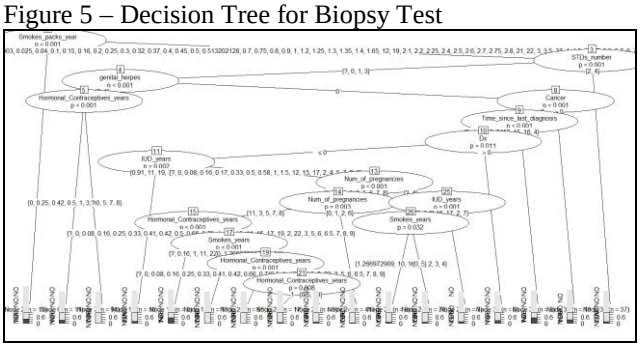

Figure 6 – All Tests Classifier

4.2. Evaluation

We can see from the above decision trees that the cases where the prediction Yes is answer have less probabilities and so of the answer is No in majority of cases. Such predictions are of no use as they are not accurate. The main reason of this problem is very less number of records having positive answers for tests as compared to negative answers for tests. To overcome this issue, we have classified categorical attributes corresponding to Hinselmann, Schiller, Citology, and Biopsy tests as numerical attributes with probabilities of belief in respective test. This enhances the accuracy as we are not concluding with Yes or No but the probability between 0 and 1. The higher the probability, the more likely the answer is to be Yes and vice versa. Experimental thresholds are derived for each of the four tests which are listed as below.

(1) Hinselmann Test is No if Probability is < 0.20 otherwise Yes.
(2) Schiller Test is No if Probability is < 0.23 otherewise Yes.
(3) Citology Test is No if Probability is < 0.22 otherwise Yes.
(4) Biopsy Test is No if Probability is < 0.30 otherwise Yes.

4.3. Confusion Matrix

Accuracy can be calculated based on the confusion matrix interpretation. The purpose here is to find out how many Yes cases are predicted as Yes and how many No cases are predicted as No. The four confusion matrices corresponding to four tests are shown below.

Hinselmann Test		Actual	
		No	Yes
Predictions	No	797	22
	Yes	26	13

Table 1 - Confusion Matrix of Hinselmann

Schiller Test		Actual	
		No	Yes
Predictions	No	745	43
	Yes	39	31

Table 2 - Confusion Matrix of Schiller

Citology Test		Actual	
		No	Yes
Predictions	No	789	36
	Yes	25	8

Table 3 - Confusion Matrix of Citology

Biopsy Test		Actual	
		No	Yes
Predictions	No	775	36
	Yes	28	19

Table 4 - Confusion Matrix of Biopsy

4.4. Accuracy

Accuracy is calculated based on the confusion matrix interpretation. The purpose here is to find out how many times our classifiers succeed in providing correct prediction by comparing the actual values with the predicted values. The accuracy can be calculated by taking ratio of correct predictions by total predictions. Converting it into the percentage we get the success rate of classification. The formula to calculate accuracy is,

P_{Yes} = No. of Correct Cases for Class "YES"
P_{No} = No. of Correct Cases for Class "NO"
N_{Yes} = No. of Incorrect Cases for Class "YES"
N_{No} = No. of Incorrect Cases for Class "NO"

$$Accuracy = (P_{Yes} + P_{No}) / (P_{Yes} + P_{No} + N_{Yes} + N_{No})$$

For example, Accuracy of Hinselmann test is = (13+797) / (13+797+22+26) = 0.9441

Accuracies of all four tests are shown below.

Sr.	Test	Accuracy
1	Hinselmann	94.41 %
2	Schiller	90.44 %
3	Citology	92.89 %
4	Biopsy	92.54 %

Table 5 - Accuracy of Predictions

5. CONCLUSION

This paper discussed how classification model can be developed for cervical cancer related test requirement prediction. The basis of cervical cancer and related concepts are discussed. Decision tree algorithm is discussed along with an example for better understanding of the concept. Most importantly, a dataset of cervical cancer is found and studied so that

more realistic model can be developed. A decision tree method is used to build a model with R. the accuracy is calculated based on the testing with the training data. Two novel concepts are introduced. 1st is the introduction of probabilistic intermediate state where the categorical class labels are found based on the probabilities of belief. 2nd is the usage of experimental thresholds - different for different tests. Both of these concepts have helped us in improving the accuracy as well as removing the problems due to extreme small number of positive classes as compared to large number of negative classes with natural randomness. It is obvious that the real data affects accuracy as compared to settled data but it is indeed a requirement to come out with a model which is near to the reality. Acceptable accuracies for each of the four tests are found.

6. FUTURE WORK

Further research work can be carried out towards finding more detail about cervical cancer related parameters. In this data set we have included only life style and medical history further to it, symptoms can be added. A generalize model can be designed to accurate the model by using multiple classification algorithms instead of only decision tree method. Data pre-processing steps could be used to improve training phase too. As discussed earlier, every cancer has unpredictable symptoms and it is difficult to detect test requirement easily. Still our decision tree classifier with experimental thresholds provides acceptable accurate results. The more to the improvement, experimental thresholds can be replaced with dynamic thresholds. The comparison could be made of independent test prediction and combined tests predictions. More to the accuracy improvement, multiple classification methods can be used and results can be compared for more accuracy. The same results can be accepted easily while the different results can be analyzed further to conclude with a single result. The analysis can be done in more detail with other indirect parameters like Kappa, Sensitivity, Specificity etc.

REFERENCES

[1] Cervical Cancer Prevention. [Online - 2015]. Available: https://www.cancer.gov/types/cervical/hp/ cervical-prevention-pdq

[2] World Cancer Report, World Health Org., Geneva, Switzerland, 2014.

[3] Aggarwal, Charu C. Data mining: the textbook. Springer, 2016

[4] Jiawei Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001

[5] Kaur H, Wasan SK. Empirical Study on Applications of Data Mining Techniques in Healthcare. J Comput Sci. 2006;2(2):194-200. doi:10.3844/jcssp.2006.194.200.

[6] Venkatadri.M and Lokanatha C. Reddy ,"A comparative study on decision tree classification algorithm in data mining" , International Journal Of Computer Applications In Engineering ,Technology And Sciences (IJCAETS), Vol.- 2 ,no.- 2 , pp. 24- 29 , Sept 2010.

[7] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics21.3 (1991): 660-674.

[8] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.