Ensemble Classifier for Cervical Cancer Required Test Prediction

Dipti N. Punjani[1] & Dr. Kishor Atkotiya[2]
[1]Assistant Professor, National Computer College, Jamnagar
[2]Professor, Department of Statistics, Saurashtra University- Rajkot

Received: December 07, 2018 Accepted: January 24, 2019

ABSTRACT: *Cervical Cancer is the most crucial cancer type. The main problem with this cancer is several different tests are required for its detection. There is always a dilemma to select a type of test for initial diagnosis. In cancers like deadly dieases, we can not completely remove role of doctors but we can atleast help them in taking decisions earliest. We can also help patients to be aware about what kind of tests they might need to go through to be prepared accordingly. We have introduced an ensemble classifier to classify a new patient according to the likelihood of a specific test she should go through. According to the real dataset of approx 800 patients of Hospital Universitario de Caracas' in Caracas- Venezuel four important cervical cancer related tests Hinselmann, Schiller, Citology, and Biopsy are identified.*

Key Words: *Cervical Cancer, Hinselmann, Schiller, Citology, Biopsy, Decision Tree, Naïve bayes, Ensemble Classification*

1. Introduction

Cervical cancer is one of those deadly cancers which need immediate medical attention before it becomes severe and sometimes life threatening too. Undoubtedly the medical science has reached to the milestones where doctors can diagnosis with utmost accuracy. Still there is a need of automation which can be used by doctors for faster decision making. At initial level, such automated base system can help doctors to identify the initial screening tests requirements especially when the patients are not in critical stages. Along with the time reduction, such system will help patients to be aware about the tests too. We have seen many patients going through the tests; do not know exactly what tests they are giving and what their significancea are. Nowadays, different tests need patients to be prepared for them differently. A common awareness will help patients to start evaluating the tests and their results in layman way. People can also start relating their life style and past medical history to the possibility of cancers like cervical cancers in detail. According to the real dataset of approx 800 patients of Hospital Universitario de Caracas' in Caracas- Venezuel four important cervical cancer related tests Hinselmann, Schiller, Citology, and Biopsy are identified. We have tried to classify a new patient's detail to suggest which tests out of these 4 tests she should go through. We have tried to improve accuracy with the help of ensemble classifier design where we are using decision tree method and naïve bayes method together. As a part of our research, we had also introduced probabilistic classification to process a large data with a very small subset having classified levels. This paper discusses how our ensemble classifier can predict which tests a patients should go through to diagnosis cervical cancer further [1,2,3,4].

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

2. Ensemble Classifier Design

An ensemble classifier is a classification process where we use multiple classification algorithms to classify same data. In our case, we have used naïve bayes method and decision tree method: two of the most widely used classification methods of data mining. Later on, we had introduced an accuracy and generalization increasing concept to pick one of the results given by both or improve one's result of these algorithms systematically.Figure 1 shows a general model of an ensemble classifier. [3,4,5,6].

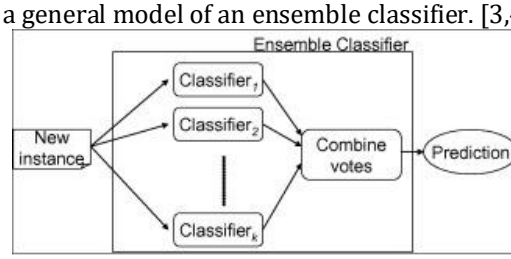
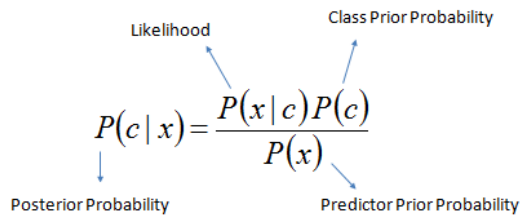


Figure 1 – Ensemble Classifier

2.1. Naïve Bayes Classification

The method uses explicitly calculated probabilities from the training data set. Figure 1 shows how bayes theorem is used by naïve bayes classifier to predict possible classification label. In this case, P(c|x) represents posterior probability to calculate the probability of labelling the new data x with label c. the higher this probability, the more chance of label c as output. P(x|c) is likelihood to represent the probability of data x with label c in dataset. P(c) – class prior probability represents the probability of class c output in dataset where as P(x) – predictor prior probability represents the probability of x in dataset. Here these probabilities are with reference of presence of class label c and availability of x in training data set. x is a record with a set of values (x_1, x_2, x_3,.... x_n) which will be used to calculate final value of P(c|x) as shown below[3,4,5].



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 1 – Naïve Bayes Classification

2.2. Decision Tree Classification

This method though uses the probabilities of available class labels corresponding to the values of various attributes of various records indirectly, in the form of entropy. Figure 2 shows a basic structure of a decision tree classifier. Here we can see that the classifier is in the form of a tree. A tree is with its interior nodes as conditions and exterior nodes as class labels. The interior nodes are conditions which decide the path to reach to one of the exterior nodes. The exterior nodes are the labels which are the outputs in the form of classification labels. The root is the most important condition and it goes downwards, as per the importance. Here importance refers to the relevance, and influencing power of a condition on deciding the classification result in the form of a class label. Decision tree classification is a process of deriving such classification tree from the available dataset by processing a set of entropy values. Here the main goal is to derive important set of conditions which lead us to various class labels [7,8,9,10,11].

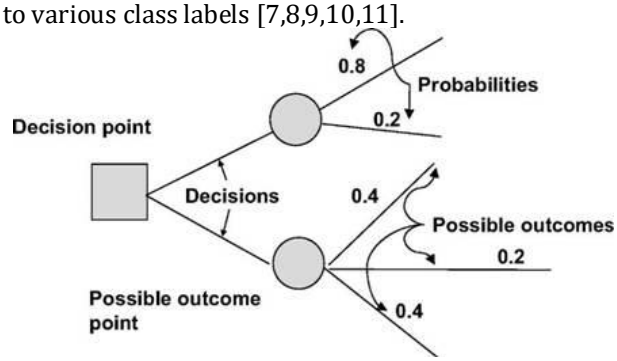


Figure 2 – Decision Tree Classification

3. DATA SET

In any data mining research, data must be real to get accuracy. We have tried to search for the genuine data set. We feel immensely thankful to Hospital Universitario de Caracas' in Caracas- Venezuel for public access of real data of nearly 800 patients. We also appreciate how they have prepared dataset without revealing identity of the patients. This dataset is composed of various fields which are listed in Table – 1.

Age	vulvo_perineal_condylomatosis
Number_of_sexual_partners	syphilis
First_sexual_intercourse	pelvic_inflammatory_disease
Num_of_pregnancies	genital_herpes
Smokes	molluscum_contagiosum

Smokes_years	AIDS
Smokes_packs_year	HIV
Hormonal_Contraceptives	Hepatitis_B
Hormonal_Contraceptives_years	HPV
IUD	Number_of_diagnosis
IUD_years	Time_since_first_diagnosis
STDs	Time_since_last_diagnosis
STDs_number	Cancer
condylomatosis	CIN
cervical_condylomatosis	HPV
vaginal_condylomatosis	Dx

Table – 1 Database of Cervical Cancer Test

4. IMPLEMENTATION

We have done implementation with R. Total 9 classifiers are developed. 4 for naïve bayes method and 5 for decision tree method. Each of these 9 classifiers involves an individual classifier for 4 tests Hinselmann, Schiller, Citology, and Biopsy. In addition 1 classifier to derive all tests together is also developed for decision tree method.

4.1. Naïve Bayes Result

We have implemented Naïve Bayes classification method over our data set. For individual tests and combined group of 4 tests, we have tried to determine various matrices which are shown in below. Table – 2 is for Hinselmann test. Table – 3 is for Schiller test. Table – 4 is for Citology test. Table – 5 is for Biopsy.

Confusion Matrix Hinselmann Test Naïve Bayes		Actual	
		No	Yes
Predictions	No	469	2
	Yes	354	33
Accuracy : 58.51% No Information Rate : 0.9592 Kappa : 0.0882 Sensitivity : 0.56987 Specificity : 0.94286		Pos Pred Value : 0.99575 Neg Pred Value : 0.08527 Prevalence : 0.95921 Detection Rate : 0.54662 Detection Prevalence : 0.54895 Balanced Accuracy : 0.75636	

Table – 2 Hinselmann Test Naïve Bayes

Confusion Matrix Schiller Test Naïve Bayes		Actual	
		No	Yes
Predictions	No	714	38
	Yes	70	36
Accuracy : 87.41% No Information Rate : 0.9138 Kappa : 0.3322 Sensitivity : 0.9107 Specificity : 0.4865		Pos Pred Value : 0.9495 Neg Pred Value : 0.3396 Prevalence : 0.9138 Detection Rate : 0.8322 Detection Prevalence : 0.8765 Balanced Accuracy : 0.6986	

Table – 3 Schiller Test Naïve Bayes

Confusion Matrix Citology Test Naïve Bayes		Actual	
		No	Yes
Predictions	No	449	3
	Yes	365	41

Accuracy : 57.11% No Information Rate : 0.9487 Kappa : 0.0988 Sensitivity : 0.5516 Specificity : 0.9318	Pos Pred Value : 0.9934 Neg Pred Value : 0.1010 Prevalence : 0.9487 Detection Rate : 0.5233 Detection Prevalence : 0.5268 Balanced Accuracy : 0.7417

Table – 4 Citology Test Naïve Bayes

Confusion Matrix Biopsy Test Naïve Bayes		Actual	
		No	Yes
Predictions	No	752	32
	Yes	51	23
Accuracy : 90.33% No Information Rate : 0.9359 Kappa : 0.3055 Sensitivity : 0.9365 Specificity : 0.4182		Pos Pred Value : 0.9592 Neg Pred Value : 0.3108 Prevalence : 0.9359 Detection Rate : 0.8765 Detection Prevalence : 0.9138 Balanced Accuracy : 0.6773	

Table – 5 Biopsy Test Naïve Bayes

4.2. Decision Tree Result

We have implemented decision tree classification method over our data set. For individual tests and combined group of 4 tests, we have tried to determine various matrices. We got following accuries as shown in Table – 6 for individual decision tree approach. One decision tree is shown in Figure – 2 for reference too.

Sr.	Test	Accuracy
1	Hinselmann	95.92%
2	Schiller	91.38%
3	Citology	94.87%
4	Biopsy	93.59%

Table –6 Accuracy of Decision Tree Methods

4.3. Ensemble Classification

We have noticed that naïve bayes classification can classify both cases: patients with need of test and patients with no need of test. While decision tree classification can only classify patients with no need of test. At the same time, accuracy of naïve bayes classifiers is less as compared to of decision tree classifiers. So we decided to ensemble two methods. 1st we select the classifier with utmost accuracy and later on we tried to improve accuracy. Here we prefer a small loss of accuracy if we can achieve a classifier which can classify both the cases: patients who need to go for tests and patients who do not need to go for tests. The detail discussion on how we tried to improve efficiency using experimental thresholds and probabilities class labels is discussed in our paper entitled "Cervical Cancer Test Identification Classifier using Decision Tree Method". The approach is shown in Figure 3.

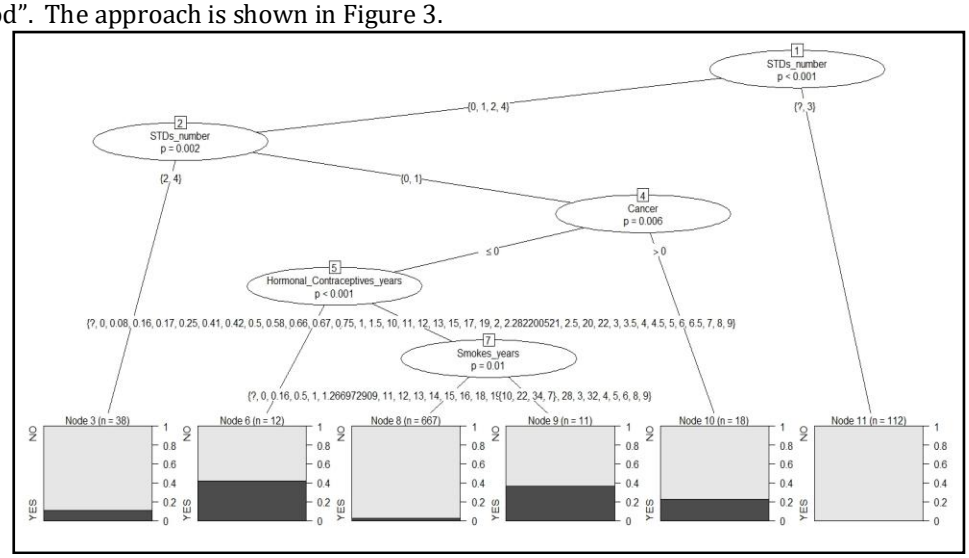


Figure 2 – Decision Tree for Hinselmann Test

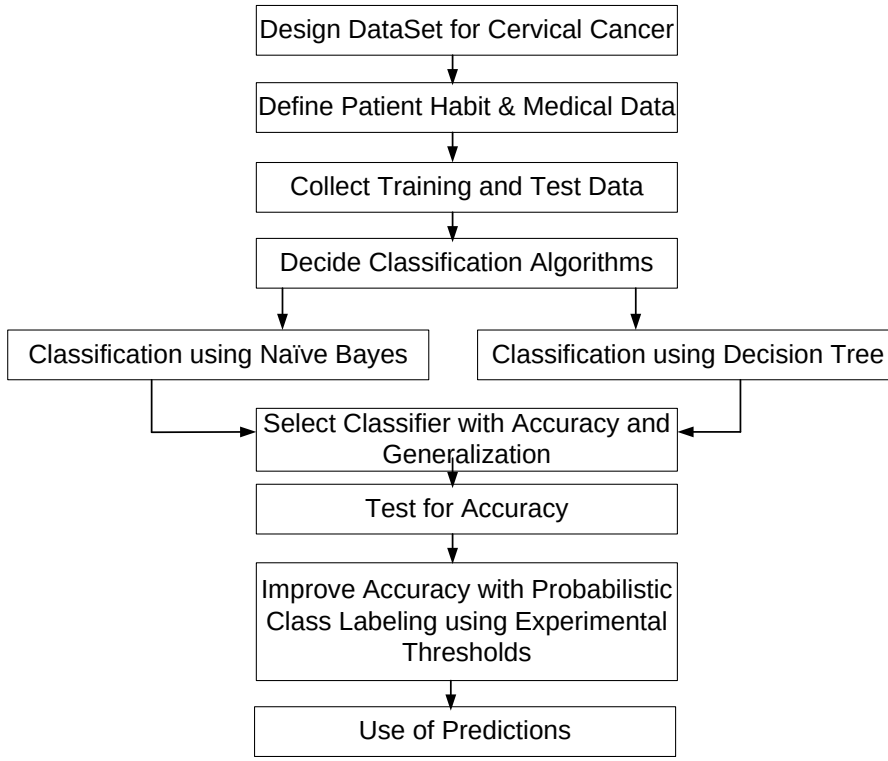


Figure 3 – Cervical Cancer Test Identification Ensemble Classifier System

CONCLUSION

In this research paper we tried to solve problem of cervical cancer test identification requirement prediction problem using various classification methods. We started with finding a real data of patients to process with. The purpose is to get real and accurate model by processing real and accurate data of patients. We have processed data of around 800 patients of Hospital Universitario de Caracas' in Caracas- Venezuel which is made to be publicly accessible for research purpose. As the problem which we target is very sensitive because of life threatens of the diseases, we tried to explore various ways of prediction so that acceptable accuracy can be achieved. Initially we tried to use one classification method, but later on we seen that a single method can not be applicable accurately for all patients for all tests. We shifted our research focus to ensemble classification where multiple methods simultaneously take part into decision making. We used naïve bayes method. This method could able to identify patients who need to go for tests and patients who do not need to go for tests but with some what low accuracies. We used decision tree method. This method could able to identify patients who do not need to go for tests only with high accuracy (obviously because in our database, a large number of patients are those who do not require to go for tests). Later on as discussed we introduced probabilistic classes with experimental thresholds to maintain accuracy as well as to able to classify both the cases.

FUTURE WORK

Above all of the points which we covered, there is much scope for further improvement and further research in this direction. We have used two algorithms naïve bayes and decision tree. In future more classification algorithms could be used to analyze accuracies of more number of classification algorithms in this problem. It is also possible to go through the concept of over fitting and under fitting for to make it more scalable. At present, we could get dataset of one hospital. We can also visit medical colleges and hospitals to collect database and can check if some new fields can be added or not. Further to the appropriateness of this system, we can consult medico officers and doctors to use our system and to find out how useful, accurate, complete, efficient, effective, scalable, latest and easy to use our system is. We can also take feedback from people, who use them. Cervical cancer is a very sensitive cancer which needs utmost attention and care not just by doctors but also by patients and their family members. As it is female centric, further awareness is required in rural area due to lack of knowledge. Our work can be extended to be more focused on life style and less focused on medical tests to make it easy to be used by all people at initial stage. Later on, if it is

required, medico people can use the system for advance analytics with reference of some primary tests results.

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. Abid Sarwar, Mehbob Ali, Jyotsna Suri, Vinod Sharma "Performance Evaluation of Machine Learning Techniques for Screening of Cervical Cancer" Proceedings of the 9th INDIACom; INDIACom-2015; IEEE Conference ID: 35071 2015
9. Sunny Sharma ,"Cervical Cancer stage prediction using Decision Tree approach of Machine Learning" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016
10. R. Vidya and G. M. Nasira, "Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns" Indian Journal of Science and Technology, Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/82085, August 2016
11. 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.