International Journal of
Scientific Engineering and Technology Research

ISSN 2319-8885
Vol.05,Issue.45
November-2016,
Pages:9230-9234

www.ijsetr.com

Application of Training Dataset using Naïve Bayes Classifier for Prediction of Stomach Cancer in Female Population

DR. VANI PERUMAL[1], SHIBU SAMUEL[2], DR. P. INDRA MUTHU MEENA[3]

[1]Assistant Professor, RCAS, Sultanate of Oman.

[2]Lecturer, RCAS, Sultanate of Oman.

[3]Assistant Commsr & Oncology Researcher, Chennai, TN, India.

Abstract: Research says that Stomach cancer is the fifth most common cancer in the world. Prediction of this is a vital task. Data mining plays a significant role for predicting such dangerous diseases. It is a process of dredge up information from the enormous datasets or other repositories. This paper describes a prototype which uses a set of training data developed by Naïve Bayes Classifier for the prediction of stomach cancer. The maximum likelihood training was done for all the independent attributes then based on the results, prediction will be done. From the experimental result it is observed that Naïve Bayes Algorithm is a better classification algorithm for the prediction of Stomach cancer.

Keywords: Data Mining; Female Stomach Cancer; Naïve Bayesian Classifier; Machine Learning; Probabilistic Model.

I. INTRODUCTION

Cancer becomes the most life threatening disease now a days. Stomach cancer, also called gastric cancer, is a cancer that starts in the stomach. Stomach cancers tend to develop slowly over many years. Before a true cancer develops, pre-cancerous changes often occur in the mucosa of the stomach. These early changes rarely cause symptoms and therefore often go undetected [1]. The early prediction of cancer plays a vital role in saving human life. Nowadays, since the world become too small with the help of technology, the researchers and the health care enterprises can collect huge amount of healthcare data from a large population. This data should be "mined" to discover hidden information. Data mining is acting as a key factor for predicting more type of diseases. The entire database report of medical patients can be utilized in a more efficient manner for the prediction of Stomach cancer. This paper also utilizes the medical reports of some other people, who are not predicted with the same disease. These two set of data are utilized for the machine learning process for producing training data set. Then these data set are further used in the mining algorithm for the prediction of the disease. There are number of factors which increases the risk of stomach cancer in female population. Some of the important factors are extracted and utilized from the collected medical data set for the prediction algorithm. Those important attributes are listed below:

- Age
- Food habit
- Menopause
- Profession
- Smoking habit - Passive smokers or Non smokers
- Soft drink consumption
- Tobacco Use
- Hematology - Blood Group
- Hematology - Oleic Acid measurement 32.93 and above
- Hematology - Palmitic Acid measurement 13.22 and above
- Hematology - Lino Leic Acid measurement 25.20 and above
- Hematology - Amino Acid Profile: Proportion of Lucine and Iso Lucine measurement 2.3 and above

The paper focuses only female population. The data set used for this machine learning process were collected from the female population in the various districts from Tamil Nadu, India. The size of the training data set is fifty with different values for all the above mentioned twelve attributes.

II. LITERATURE REVIEW

Dr. S. Vijayarani et.al used Naïve Bayes Algorithm and Support Vector Machine for the prediction of Liver disease. They predicted normal liver diseases, CBCL, Acute Hepatitis and Outliers using six attributes [6]. Ankita et.al utilized Naïve Bayes Classifier for the prediction of Swine Flu disease. They have used the values of eight different attributes for the prediction [8].Sukhmeet Kaur et.al used Naïve Bayes algorithm for the prediction of future manufacturing of number of cars which is useful for the car manufacturing industry. They produced the prediction results. Then they compared the prediction results with the actual and real world values in order to validate the results. The utilized Naïve Bayes algorithm for predicting the result

[14]. Dhamodharan et.al predicted three major liver diseases such as Liver cancer, Cirrhosis and Hepatitis with the help of distinctive symptoms. The authors used Naïve Bayes algorithm and FT Tree algorithm for the prediction of those diseases. Comparison of the performance of these two algorithms has been done based on the classification of accuracy measure. Based on the experimental results they concluded that the Naïve Bayes algorithm as a better algorithm for the prediction of the diseases with maximum classification accuracy than the other algorithm [3].

Dhanashree et.al implemented a classifier approach for the detection of heart disease. Also they have shown how Naïve Bayes algorithm can be used for classification. They have used thirteen parameters as a classifiers for prediction of heart disease [4]. Rosalina et.al predicted a hepatitis prognosis disease using SVM and Wrapper Method. First they have used wrapper methods to remove the noise features then they have used the classification process. Features selection were implemented to minimize the noisy or irrelevant data. Using the experimental results they observed the increased accuracy rate in the clinical laboratory test cost with minimum execution time. They achieved the target by combining Wrappers Method and Support Vector Machine techniques [12]. Omar S. Soliman et.al has projected a hybrid classification system for HCV diagnosis, using Modified Particle Swarm Optimization algorithm and Least Squares Support Vector Machine. In their paper, they used Feature vectors which are extracted using Principle Component Analysis algorithm. LS-SVM algorithm is sensitive to the changes of values of its parameters, so they used Modified-PSO Algorithm to search for the optimal values of LS-SVM parameters. They obtained in a less number of iterations. Their proposed system was implemented and evaluated in the benchmark HCV data set from UCI repository of machine learning databases. Then the database was compared with another classification system. That particular system utilized PCA and LS-SVM. From their experimental results, they proposed a new system, which obtained maximum classification accuracy than the other systems [11].

Sneha et.al used Apriori algorithm to classify the web pages based on the results extracted after submitting the query to the search engine. Then they applied Naïve Bayes algorithm to calculate the probability of each feature to classify the page into the respective class based on its probability [13]. Karthik et.al applied a soft computing technique for intelligent diagnosis of liver disease. They implemented classification detection and its type detection in three phases. In the first phase, they classified the disease using Artificial Neural Network classification algorithm. In the second phase, they generated the classification rules using Learn by Example algorithm. In the third phase fuzzy rules were applied to identify the types of the liver disease. Thus they used ANN for classification [7]. Chaitrali S. Dangare et.al has structured prediction systems for Heart disease using more number of input attributes. They used the data mining classification techniques like Decision Trees, Naive Bayes and Neural Networks. The performances of these techniques are compared, based on accuracy. Their analysis shows that out of these three classification models Neural Networks has predicted the heart disease with highest accuracy [2]. Dipali Bhosale et.al used Naïve Bayes algorithm for feature selection. First they noted classification results without doing any kind of feature selection techniques. They also used Co-relation based Feature Selection, Wrapper, and Information Gain on the data sets. Then, by using these three feature selection techniques they separate feature subsets which are chosen for each technique then they passed the selected features to the classifiers and they derived the final results [5].

III. DATA MINIG AND NAÏVE BAYES CLASSIFIER

A. Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step it can also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, post-processing of discovered structures and visualization. Data mining is the analysis step of the knowledge process in the databases. Data mining software is used number of different tools for analyzing data. It also allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among different of fields in large relational databases. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Actually Data mining is the process of extracting knowledge hidden in the large set of databases.

B. Naïve Bayes Classifier

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Naive Bayes algorithm is based on Bayesian Theorem. Naïve Bayes classifier an effective Bayesian classifier built upon the strong assumption that different features are independent with each other. Classification is done by taking the highest posterior of classification variable or attribute given a set of feature. It assumes that the effect of a variable value on a given class is independent of the values of other variable. This assumption is known as class conditional independence. An advantage of the Naïve Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification. Abstractly, Naïve

Bayes is a conditional probability model The following is the probabilistic model of Naïve Bayes classifier.

In this model, each data texture is represented by an n dimensional feature vector, depicting n measurements made on the sample from n attributes. Consider that there are m classes labeled from C1 to Cm. Given an unknown data sample, Xi with no class label, the classifier will state that Xi belongs to the class having the highest posterior probability, conditioned on Xi. The Naïve Bayes probability assigns an unknown sample Xi to the class Ci. Classification problem is given to the classifier which has the combination of different values of selected variables with a related value of class variable; the classifier then returns a posterior probability distribution over the class variable. Therefore the value of P(Ci|Xi) is maximized. The class Ci for which P(Ci|Xi) is maximized is called the maximum posteriori hypothesis. By Bayes theorem P(Ci/Xi)= (P(Xi/Ci)P(Ci))/P(Xi), since P(Xi) is constant for all classes, only P(Xi|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and therefore therefore P(Xi|Ci) alone should be maximized. Otherwise P(Xi|Ci)P(Ci) will be maximized. Note that the class prior probabilities may be estimated by P(Ci) = Si/S , where Si is the number of training samples rate of P(C|xi) = P(xi1|C) * P(xi|C) * P(xi3|C) * P(xi4|C) * …P(xin|C) * P(|C) class Ci, and S is the total number of training samples.

IV. METHODOLOGY and RESULTS

The proposed methodology uses Naïve Bayes algorithm for the prediction of stomach cancer in female population. The entire algorithm consists of the following three steps.

Step. 1: Each data sample is represented as an n dimensional feature vector, X = (x_1, x_2….. x_n). This depicts n measurements made on the sample from n different attributes like A_1, A_2…..A_n respectively.

Step. 2: Assume that there are m classes, C_1, C_2……C_m in a data sample X, the classifier will predict that X belongs to the class having the highest posterior probability, which is conditioned as: if and only if: P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= i Thus P(Ci|X) is maximized. As mentioned above, thus the class Ci for which P(Ci|X) is maximized and referred as maximum posteriori hypothesis.

Step. 3: As described above, P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. But for the class, prior probabilities are unknown. Hence it is assumed that the classes are equally likely, so P(X|Ci)P(Ci) is maximized. If they are not equally likely, P(X|Ci) is maximized. Note that the class prior probabilities can be estimated by P(Ci) = Si/S On X, the Naïve probability assigns an unknown sample X to the class Ci.

The stomach cancer related Data set was collected from various private and government hospitals and also from the public population. It is a real time data gathered from the female cancer patients and public female population. This dataset has fifty instances and twelve attributes. The attributes are age, food habit, menopause, profession, smoking habit, soft drink consumption and tobacco use. This dataset also contains the details Hematology test details. These Data set was trained in a program such that the probabilities of all the classes with all the conditions were calculated. Result was gathered in database and when the test data was given, the probabilities for the different classes for the given symptom values were obtained on the basis of which it is implied that the patient may fell into the class with the highest probability. This is a simple and powerful technique that is instrumental in helping the society to predict the category a patient falls into. From this results, stomach cancer can be detected or if there is a possibility then precautionary measures can be taken. The following tables gives the values derived from the training data set using Naïve Bayesian classifier. In the following Tables, P denotes prone to Stomach Cancer and N denotes not prone to Stomach Cancer.

TABLE I: Probability of the Attribute "Age"

Age	P	N
<30	4/35	4/15
30-40	8/35	6/15
40-50	12/35	2/15
>50	11/35	3/15

TABLE II: Probability of the Attribute "Food Habbit"

Food	P	N
Veg	3/35	6/35
NV	8/35	3/15
NV-M	10/35	4/15
NV-M&B	14/	2/15

In Table.2, NV stands for Non vegetarian, NV-M stands for Non vegetarian frequent meat consumers and NV-M&B stands for Non vegetarian frequent meat and beef consumers.

TABLE III: Probability of Attribute "Menopause"

Mp.	P	N
Yes	28/35	5/15
No	7/35	10/15

TABLE IV: Probability of the Attribute "Smoking"

Passive Smokers	P	N
Yes	20/35	10/35
No	15/35	5/15

In Table.4, Smoking only considers Passive smokers since the data set used for this prediction only consider Female population in Tamil Nadu-India.

TABLE V: Probability of the Attribute "Profession"

Prfsn.	P	N
F	16/35	5/15
S	9/35	4/15
T	10/35	6/15

TABLE VI: Probability of the Attribute "Carbonated Drink Consumption"

C.D.Cons.	P	N
MI	9/35	8/15
MO	11/35	5/15
EX	15/35	2/15

Table.5 depicts three broad classifications of profession based on the maximum occurrences of stomach cancer. They are Farmer-F, Sedentary Jobs-S and Textile Professionals-T. Table.6 denotes the consumption of Carbonated drinks. Here it is divided into three categories named Mild-MI, Moderate-MO and Excessive-EX.

TABLE VII: Probability of the Attribute "Blood Group"

B.G.	P	N
A	13/35	3/15
B	14/35	7/15
Other	8/35	5/15

TABLE VIII: Probability of the Attribute "Oleic Acid" in Hematographic Studies

O.A. ≥ 32.93	P	N
Yes	23/35	9/15
No	12/35	6/15

Table.7 classifies the Blood Group as A, B and Other groups. A stands for both A+ve, A-ve and B stands for both B+ve, B-ve.

TABLE IX: Probability of the Attribute "Linoleic Acid" in Hematographic Studies

L.A. ≥ 25.20	P	N
Yes	24/35	7/15
No	11/35	8/15

TABLE X: Probability of the Attribute "Palmitic Acid" in Hematographic Studies

P.A. ≥ 13.22	P	N
Yes	22/35	10/35
No	13/35	5/15

TABLE XI: Probability of the Attribute "Amino Acid Profile" in Hematographic Studies

P.L& IL ≥ 25.20	P	N
Yes	28/35	5/15
No	7/35	10/15

TABLE XII: Probability of the Attribute "Tobacco Users"

Tobacco	P	N
Yes	21/35	3/35
No	14/35	12/15

From the mentioned tables, first all possible probabilities conditioned on the target attribute of Stomach Cancer is computed by using Bayesian Classifier method.

P(St. Cancer = P) = 35/50 = 0.7

P(Age<30 | St. Cancer = P) = 0.11

P(Age=30-40 | St. Cancer = P) = 0.23

P(Age=40-50 | St. Cancer = P) = 0.34

P(Age>50 | St. Cancer = P) = 0.32

P(St. Cancer = N)= 15/50 = 0.3

P(Age<30 | St. Cancer = N) = 0.27

P(Age=30-40 | St. Cancer = N) = 0.4

P(Age=40-50 | St. Cancer = N) = 0.13

P(Age>50 | St. Cancer = N) = 0.2

Like the above mention method, all the attribute's possible probabilities were calculated using the before mentioned twelve tables. Now by implementing the Naïve Bayes classifier and the information derived by using the Classification algorithm for the above data set, we can obviously predict the possibility Stomach Cancer for any observed information.

TABLE XIII: The Observed Values Of All The Twelve Attributes in a Female Patient

S.No	Attributes	Values
1.	Age	30-40
2.	Food Habit	NV
3.	Menopause	No
4.	Smoking	No
5.	Profession	Farmer (F)
6.	Carbonated drink consumption	Mild (MI)
7.	Blood Group	B +ve (B)
8.	"Oleic Acid" in Hematographic Studies	O.A. < 32.93 (No)
9.	"Linoleic Acid" in Hematographic Studies	L.A. < 25.20 (No)
10.	"Palmitic Acid" in Hematographic Studies	P.A. < 13.22 (No)
11.	"Amino Acid Profile" in Hematographic Studies	Proposition of Lucine & Iso Lucine < 25.20 (No)
12.	Tobacco User	No

To mine the data regarding the prediction of Stomach Cancer, all the Bayesian classifiers listed in the above twelve tables are utilized. The prediction process starts with splitting the two cases, one for P and another for N. P1 is considered as the prediction case of P. Therefore the value of P1 is computed as follows.

P1 = P(St. Cancer = P) * P(Age=30-40 | St. Cancer = P) * P(Food Habbit=NV | St. Cancer = P) * P(Menopause = N | St. Cancer = P) * P(Smoking = N | St. Cancer = P) * P(Profession = F | St. Cancer = P) * P(Carbonated drink

consumption = MI | St. Cancer = P) * P(Blood Group = B | St. Cancer = P) * P(Oleic Acid = N | St. Cancer = P) * P(Linoleic Acid = N | St. Cancer = P) * P(Palmitic Acid = N | St. Cancer = P) * P(Amino Acid Profile = N | St. Cancer = P) * P(Tobacco User = N | St. Cancer = P)

P1 = 0.7 * 0.23 * 0.23 * 0.2 * 0.43 * 0.46 * 0.26 * 0.4 * 0.34 * 0.31 * 0.37 * 0.2 * 0.4

So P1 = 0.0000004753

Now P2 is considered as the prediction case of N. Therefore the value of P2 is computed as follows.

P2 = P(St. Cancer = N) * P(Age=30-40 | St. Cancer = N) * P(Food Habbit=NV | St. Cancer = N) * P(Menopause = N | St. Cancer = N) * P(Smoking = N | St. Cancer = N) * P(Profession = F | St. Cancer = N) * P(Carbonated drink consumption = MI | St. Cancer = N) * P(Blood Group = B | St. Cancer = N) * P(Oleic Acid = N | St. Cancer = N) * P(Linoleic Acid = N | St. Cancer = N) * P(Palmitic Acid = N | St. Cancer = N) * P(Amino Acid Profile = N | St. Cancer = N) * P(Tobacco User = N | St. Cancer = N)

P2 = 0.3 * 0.4 * 0.2 * 0.67 * 0.33 * 0.33 * 0.53 * 0.47 * 0.4 * 0.53 * 0.33 * 0.67 * 0.8

So P2 = 0.0000817847

Therefore the argument Probability P1 is greater than P2 and hence the patient is not predicted with Stomach Cancer as described in the Naïve Bayes classifier algorithm [8, 9, 10]. Thus the Naïve Base algorithm is playing a vital role in mining the necessary information from a large set of Data. It supports probabilistic learning and probabilistic prediction. The algorithm was developed and implemented for machine learning and data mining. It also can be applied for any size of data set. Specifically in health care sectors, huge amount of data is available and data mining becomes an ease of use of this data set for disease prediction.

V. CONCLUSION AND FUTURE WORK

Various data mining techniques can be collaborated with Naïve Bayes classifier algorithm and can be used for the prediction of Stomach Cancer in female population. This proposed approach attained promising results, which may lead to further attempts to utilize Information Technology for the prediction of Stomach Cancer. Classifiers are playing a vital role in Data mining techniques, which are used in the prediction of diseases. From the experimental results and the real time data set, this work concludes that Naïve Bayes algorithm is the most suitable algorithm for the prediction of Stomach Cancer. The execution time of the same algorithm is also minimum.

This approach defines a technique based only on female data set in Naïve Bayes classification algorithm. In future it can be combined with male data set and implemented using various data mining techniques such as clustering and other classification algorithms.

VI. REFERENCES

[1] American Cancer Society. Cancer Facts & Figures 2016. Atlanta, GA.

[2] Chaitrali S. Dangare and Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications. JUN-2012 (Vol.47,Issue.10) 44-48.

[3] Dhamodharan. S. Liver Disease Prediction Using Bayesian Classification, Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies. MAY-2014, 1-3.

[4] Dhanashree S. Medhekar, Mayur P. Bote and Shruti D. Deshmukh. Heart Disease Prediction System using Naïve Bayes, International Journal of Enhanced Research in Science Technology & Engineering. MAR-2013 (Vol.2, Issue.3) 1-5.

[5] Dipali Bhosale and Roshani Ade., Feature Selection based Classification using Naive Bayes, J48 and Support Vector Machine, International Journal of Computer Applications. AUG-2014 (Vol.99,Issue.16) 14-18.

[6] Dr. S. Vijayarani and Mr.S.Dhayanand. Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research. APR-2015 (Vol.4,Issue.4) 816-820.

[7] Karthik. S, Priyadarishini. A, Anuradha. J and Tripathi, B. K. Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Advances in Applied Science Research. JUN-2011 (Vol.2,Issue.3) 334-345.

[8] Ms. Ankita R. Borkar and Dr. Prashant R. Deshmukh. Naïve Bayes Classifier for Prediction of Swine Flu Disease, International Journal of Advanced Research in Computer Science and Software Engineering. APR-2015 (Vol.5,Issue.4) 120-123.

[9] Ms. Ankita R. Borkar and Dr. Prashant R. Deshmukh, Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier, Artcom international conference, (978-1-4244-8093-7).

[10] Mrs.G.Subbalakshmi and Mr.M.Chinna Rao, Decision Support in heart disease prediction system using Naïve Bayes, International Journal of Computer Science and Engineering. (Vol.2,Issue.2) 0976-5166.

[11] Omar S.Soliman and Eman Abo Elhamd. Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine, International Journal of Scientific & Engineering Research. MAR-2014 (Vol.5,Issue.3) 122-129.

[12] Rosalina. A.H, Noraziah. A. Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method, IEEE 2010, 2209-2222.

[13] Sneha K. Dehankar, K.P. Wagh and Dr. P. N. Chatur, Web Page Classification using Apriori Algorithm and Naïve Base Classifier, International Journal of Advanced Research in Computer Science and Management Studies. APR-2015 (Vol.3,Issue.4) 527-533.

[14] Sukhmeet Kaur and Kiran Jyoti. Predicting the future of car manufacturing industry using Naïve Bayse Classifier, International Journal for Science and Emerging Technologies with Latest Trends. 2012 (Vol.4, Issue.1) 25-34.