Prediction of Cancer Risk in Perspective of Symptoms using Naïve Bayes Classifier

Pallavi Mirajkar [1]
Dr. G. Prasanna Lakshmi [2]

[1] Research Scholar, Faculty of Computer Science

Pacific Academy of Higher Education and Research University, Udaipur.
Andhra University

Abstract— Health Care professionals face the complex task of predicting the type of cancer in patient. This earlier prediction of cancer helps to the practitioner to recommend cancer treatment. Various studies have been carried out that examine patient to improve help practitioner’s prognostic accuracy. Here we worked on major factor that is symptoms. In this thesis paper we have used Naïve Bayes Classification algorithm of data mining to predict the type of cancer. Proposed method of risk prediction aims to predict probability of cancer. Based on the classification algorithm, symptoms of the cancer are classified using Navie Bayes algorithm to recognize the risk of cancer such as lung, breast, ovarian, stomach and oral. Therefore accurate prediction of cancer in patient is important for good clinical decision making in health care strategies.

Index Terms— Data Mining, Naïve Bayes Classifier, Probability, Cancer, Symptoms.

INTRODUCTION

Cancer is a collection of diseases concerning anomalous cell growth with the potential to attack or spread to other parts of the body. Cancer becomes the most life threatening disease now days. Accurate prediction of cancer in patient is important for good clinical decision making and for recommending the treatment, at present, accurate detection remains challenge even for experienced clinicians. The most effective way to decrease cancer deaths is its early recognition. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to diagnose symptoms of cancer.

Data mining plays a crucial role in the discovery of knowledge from huge databases. Data mining has created its important hold in every field including healthcare. Data mining has its most important role in extracting the concealed information. Data mining process is more than the data analysis which includes classification, clustering, association rule mining and prediction. A patient affected with any Cancer may suffer symptoms in the body. Cancer symptoms are used to forecast the risk level of the cancer disease. The main aim of this study is to predict the risk of type of cancer using Naïve Bayes classification algorithm. Naïve Bayes is one of the most effective statistical and probabilistic classification algorithms. Naïve Bayes algorithm is called “Naïve” because it makes the assumption that the occurrence of certain feature is independent of the occurrence of other features. Naïve Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. Naïve Classifiers assumes that the effect of an attribute value on the given class is independent of the values of other attributes. Naïve Bayesian Classifiers depends upon the Baye’s theorem. An advantage of the Naive Baye’s classifier is that it requires a small amount of data to estimate the parameters needed for classification. It performs superior in numerous complex real world situations like weather forecasting, Medical Diagnosis, spam classification. It is suitable when dimensionality of input is high.

RELATED WORK

Kritharth Pendyala et.al (2017) proposed a system that detects and predicts breast cancer. They have applied classification algorithm to data using Naïve Bayes and SVM algorithm. The proposed hybrid algorithm gives more accuracy and reduces the time complexity.

Subrata Kumar Mandal (2017) has applied various techniques such as data cleaning, feature selection, feature extraction, data discretization and classification for predicting breast cancer as accurately as possible. They states that Logistic Regression Classifier has given maximum accuracy.

N.V. Ramana Murty and Prof. M.S. Prasad Babu (2017) analyzed lung cancer data set using WEKA tool with various data mining classification techniques. It is recognized that the Navie Bayes classification algorithms.

Keerti Yeulkar et. al. (2017) classified the SEER dataset in to malignant tumor and benign tumor using
C4.5 and Naïve Bayes algorithm. They have also applied decision tree classification technique.

Younus Ahmad Malla et.al (2017) performed analytical evaluation of machine learning algorithm such as Naïve Bayes, Logistic Regression and Random Forest. WEKA tool is used for running different data mining algorithms for breast cancer data set.

Rashmi M, Usha K Patil (2016) have used Navie Bayes algorithm to diagnose cancer using Gene data set. They stated that data mining can be used to classify cancer patient by analyzing the data.

Animesh Hazra et.al. (2016) have studied various classification techniques and compared them in terms of accuracy and time complexity. They have concluded that Navie Bayes gives better accuracy and it is best classification technique.

Shweta Kharya, Sunita Soni (2016) have developed a predictive model using Navie Bayes algorithm. They have applied weighted concept on Navie Bayes Classification for breast cancer diagnosis. They developed model is more readable, modifiable, efficient and can be handled easily.

Dr. Vani Prumal et.al. (2016) have proposed methodology that uses Naïve Bayes classifier to predict stomach cancer. They stated that Naïve Bayes algorithm is the most appropriate algorithm for the prediction of stomach cancer. It gives more accurate result.

Dr. P. Indra Muthu Meena (2016) implemented C4.5 and Naïve Bayes algorithm for prediction of stomach cancer. They have also analyzed and compared the performance of both algorithms.

Kathija, Shajun Nisha (2016) proposed an algorithm by using support vector machine and Naïve Bayes algorithm. They evaluated performance of the proposed algorithm using confusion matrix. After comparing both algorithms, it is found that Naïve Bayes algorithm gives highest accuracy.

Samuel Giftson Durai et.al (2016) have done survey and shown that decision tree is the best suitable classification algorithm for prediction of breast cancer and experiments have been conducted using MATLAB software.

B. M. Gayathri, C. P. Sumathi (2016) applied Naïve Bayes algorithm and designed GUI to enter details of the patient for predicting whether given data is benign or malignant. Proposed work has shown that Naïve Bayes classifier predicts breast cancer with less attributes.

P. Saranya, B. Satheeskumar (2016) analyzed that data mining techniques helps to recognize data to predict the cancer at an early stage. They have studied various data mining algorithms such as Naïve Bayes, ANN, Support Vector Machine and decision tree.

Monika Hedawoo et. al. (2016) proposed method to classify mammogram using different classification algorithms for predicting tumor cells in breasts. It has been revealed that Naïve Bayes algorithm is more sensitive.

Animesh Hazra et. al. (2016) applied data cleaning, feature selection, feature extraction, data discretization and classification techniques to predict breast cancer correctly. In proposed method Naïve Bayes algorithm gives maximum accuracy.

V. Kirubha, S. Manju Priya (2015) proposed method to analyze risk factors of lung cancer. They have used classification algorithm like Naïve Bayes, Random Forest, REP Tree and Random tree. It is proved that REP Tree provides better result.

S. Kalaivani et.al (2015) studied three classifiers Naïve Bayes, Random Tree and Support Vector Machine. They have calculated the performance of these algorithms and it is observed that Naïve Bayes algorithm gives more accurate results.

Mandeep Rana et.al (2015) has applied different algorithms on dataset such as KNN, logistic regression and Naïve Bayes algorithm. It is observed that Naïve Bayes and logistic regression gives accurate results for diagnosing breast cancer.

PROPOSED WORK

Introduction
For better perceptive of approach, we take the paradigm which is shown below in the following Tables:
Naïve Bayes classifiers are a group of basic probabilistic classifiers based on applying Bayes theorem with independence assumptions between the features. Using Bayesian classifiers, the method can determine the hidden facts related with diseases from symptoms of the patients. According to naïve Bayesian classifier the occurrence (or unoccurrence of a particular feature of a class is considered as independent to the presence (or absence) of any other feature. When the dimension of the inputs is high and more efficient result is expected. Naïve Bayes model identifies the physical characteristics and symptoms of patients suffering from cancer disease. Naive Bayes algorithm is based on Bayes theorem. There are two types of probabilities

- Posterior Probability \[ P(H/X) \]
- Prior Probability \[ P(H) \]

where \( X \) is data tuple and \( H \) is some hypothesis. According to Bayes' Theorem, \[ P(H/X) = \frac{P(X/H)P(H)}{P(X)}. \]

### Table No. 5 Stomach Cancer

<table>
<thead>
<tr>
<th>Symptoms of Stomach Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indigestion or heartburn</td>
</tr>
<tr>
<td>Vomiting and nausea</td>
</tr>
<tr>
<td>Pain in the abdomen</td>
</tr>
<tr>
<td>Constipation or diarrhea</td>
</tr>
<tr>
<td>Bloating Stomach</td>
</tr>
</tbody>
</table>

### System Architecture:

1. **Enter Patient Record**
2. **Classified Data (N Class of Cancer)**
3. **Naïve Bayes Classifier**
4. **Testing Data**
5. **Cancer Risk Prediction**

Implementation of an Algorithm is as given below:
Suppose that there are \( n \) classes of cancer, \( C_1, C_2, \ldots, C_n \) and each data sample is represented by \( n \) dimensional symptoms of cancer, \( S = \{S_1, S_2, \ldots, S_n\} \).

**Step 1:** To calculate the probability of type of cancer. First recognize the symptoms of the specific cancer. \( P(C_1|S_1, S_2, \ldots, S_n) \).

**Step 2:** Calculate the probability of each symptoms. Probability of each Symptom = \( P(S_i|C_1) \) where \( i = \{1, 2, \ldots, n\} \)

**Step 3:** Multiply each probability of symptoms together
\[
P(C_1) = P(S_1|C_1) \times P(S_2|C_1) \times \ldots \times P(S_n|C_1)
\]

**Step 4:** Repeat step 1 to step 3 for each cancer class.

**Step 5:** Predict that \( S \) belongs to the type of cancer which is having the highest posterior probability.
\[
P(C_i|S) > P(C_j|S) \text{ for all } 1 \leq j \leq n \text{ and } j \neq i
\]

Display the risk of one of the type of cancer from predefined set of class.

The Naïve Base algorithm is playing a vital role in mining the necessary information. Based on the above mentioned algorithm and computed the risk of cancer, some tests will be prescribed to confirm the presence of cancer.

**CONCLUSION**

In thesis paper, we have used Naïve Bayes classifier for predicting cancer by taking into consideration the related symptoms. It is essential to work on different symptoms and discover the probability of cancer in patient. Prediction of cancer at an early stage can be helpful for better treatment and survival of the patient. In thesis paper, we compute the probability of risk of the cancer based on symptoms. And then gives a warning of which type of cancer patient may have. The goal of thesis paper is to give the prior warning to the patient. This will eventually spare the time and decrease the cost of treatment and conjointly will increase the possibility of survivability. Future work can be directed on the prediction of cancer by integrating the various dimensions such as cancer causing factors, contaminated food items and symptoms.

**REFERENCES**


**Biography:**

Mrs. Pallavi Mirajkar passed M.Sc. (Computer Science) from North Maharashtra University, Jalgaon. Currently she is pursuing her Ph.D. from Pacific University, Udaipur.

Dr. G.Prasanna Lakshmi, Guide. Associate Professor, Andhra University.