

Complex Engineering Activity

Probability and Random Variables

Abdullah
Dept. of Computer Engineering
Air University
Islamabad
180374@students.au.edu.pk

Abdur Rehman
Dept. of Computer Engineering
Air University
Islamabad
180380@students.au.edu.pk

Muhammad Usama Ameer
Dept. of Computer Engineering
Air University
Islamabad
180369@students.au.edu.pk

Abstract—This document is a model and instructions of an expert system we built that will predict the occurring and non occurring of a disease on the basis of previous data such as recognized symptoms, demographic information, spread rate, geographical information, previously diagnosed conditions, current diseases, etc. we built this expert system using probabilistic decision trees and Bayesian Inference on a real-world problem.

I. INTRODUCTION

“Disease Prediction” system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output.

At present, when one suffers from particular disease, then the person has to visit to doctor which is time consuming and costly too. Also if the user is out of reach of doctor and hospitals it may be difficult for the user as the disease can not be identified. So, if the above process can be completed using a automated program which can save time as well as money, it could be easier to the patient which can make the process easier.

Disease Predictor is a system that predicts the disease of the user with respect to the symptoms given by the user. With the help of data set given to the system, Disease Predictor will be able to know the probability of the disease with the given symptoms.

As the use of internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to internet than hospitals and doctors. People do not have immediate option when they suffer with particular disease. So, this system can be helpful to the people as they have access to internet 24 hours.

II. METHODOLOGY

Disease Predictor we have implemented is using techniques like , decision tree and Bayesian Inference algorithm. From the analysis it was found that it is one of the most accurate technique than the others.

A. Decision tree

Decision tree (DT) is one of the earliest and prominent probabilistic algorithms. A decision tree models the decision logics i.e., tests and corresponds outcomes for classifying data items into a tree-like structure.

The nodes of a DT tree normally have multiple levels where the first or top-most node is called the root node. All internal nodes (i.e., nodes having at least one child) represent tests on input variables or attributes. Depending on the test outcome, the classification algorithm branches towards the appropriate child node where the process of test and branching repeats until it reaches the leaf node . The leaf or terminal nodes correspond to the decision outcomes. DTs have been found easy to interpret and quick to learn, and are a common component to many medical diagnostic protocols . When traversing the tree for the classification of a sample, the outcomes of all tests at each node along the path will provide sufficient information to conjecture about its class. An illustration of an DT with its elements and rules is depicted in Fig.1

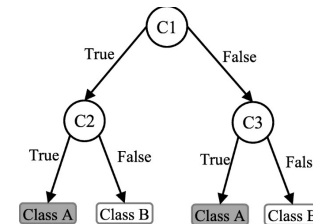


Fig. 1. A Decision Tree

B. Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics.

Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law. In the philosophy of decision theory, Bayesian inference is closely related to subjective probability, often called "Bayesian probability".

BAYESIAN ANALYSIS

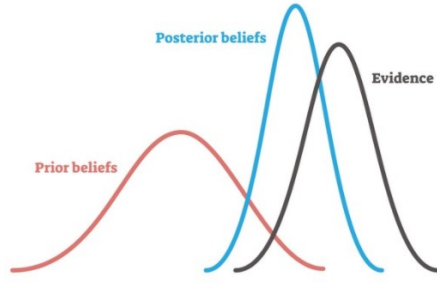


Fig. 2. Bayesian Inference

III. BACKGROUND

Machine learning algorithms employ a variety of statistical, probabilistic and optimisation methods to learn from past experience and detect useful patterns from large, unstructured and complex datasets. These algorithms have a wide range of applications.

The scope of this system is primarily on the performance analysis of disease prediction approaches using different variants of supervised algorithms. Disease prediction and in a broader context, medical informative, have recently gained significant attention from the data science research community in recent years. This is primarily due to the wide adaptation of computer-based technology into the health sector in different forms (e.g., electronic health records and administrative data) and subsequent availability of large health databases for researchers. These electronic data are being utilised in a wide range of healthcare research areas such as the analysis of healthcare utilisation.

IV. SYSTEM DESIGN

A. Pseudo Code

- We preferred bayesian Inference to build our system.
- Bayesian inference mainly works on Bayes rule.
- Bayes' theorem is stated mathematically as the following equation:

$$P(y | X) = \frac{p(X | y)p(y)}{p(X)} \quad (1)$$

Here X is random Variable whose value can be from x1 to xn.

- When we are working with random variables. Bayes' rule can be rewritten as:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(x_1|y)P(x_2|y) \dots P(x_n|y)(P(y))}{P(x_1)P(x_2) \dots P(x_n)} \quad (2)$$

- We will count the symptoms and then calculate their conditional probabilities.
- Calculated probabilities will be then used in Bayes's rule.
- Then Bayes' rule will give us final probability of disease.

- Now we will normalize probability by adding calculated probability of all diseases which will give us total probability.
- The probability of particular disease will be then divided by this total probability. which will normalize our probability.

B. Code Snippets

- Python Module of Bayesian inference
Make module of Bayesian Inference, which you can import in your code.

```
class BayesianClassifier:
    """This class contain all required methods for reading, cleaning
    calculating initial probabilities and predicting new incoming symptoms
    This class is the coded implementation of Bayesian Rule
    ATTRIBUTES:

    Functions:
```

Fig. 3. Class BayesianClassifier

- You can import module and use it Function. It has following Functions.(getData, CleanGivenData, train, Predict)

```
from BayesianClassifier import BayesianClassifier

myClassifier = BayesianClassifier()
myClassifier.getData("../DataSet/Training.csv", ["n/a", "na", "--"], ',')
myClassifier.cleanGivenData()
myClassifier.train([132], list(range(0, 132)))
results = myClassifier.runTests([ "red_sore_around_nose"])
print(myClassifier.mostProbably(results))
```

Fig. 4. Import BayesianClassifier

- Finding NAN rows from the Data and then Drop(Delete) these rows from Dataset.

```
24 # Data cleaning of training data
25 missing_values = ["n/a", "na", "--"]
26 myData = pd.read_csv("Dataset.csv", na_values = missing_values)
27
28 FindingNaN = myData.isnull()
29
30 droppingrow = []
31
32 NaN_index = FindingNaN.any(axis=1) # finding index of NAN row
33
34 for i in range(0, len(myData)):
35     if(NaN_index[i] == True):
36         droppingrow.append(i)
37
38 myData = myData.drop(droppingrow) # Drop rows which contain NAN
39 #####
```

Fig. 5. Cleaning Data

- Find Probability of all Disease given symptoms and after that normalize the probabilities of all Diseases.

```
# Calculating Percentage of Symptoms
for a in range(0,F):
    sympercentage0 = []
    for i in range(1, len(sym)+1):
        sympercentage0.append(D[a][i]/D[a][0])
    sympercentageD1.append(sympercentage0) # Calculating percentage of all sym
```

Fig. 6. Training Data

Fig. 7. Visualization Symptoms

```
# making decision what is the most probability the disease
disease = str()
lastProbability = int()

for eachDisease, diseaseProbability in probabilityResultsList.items():
    if lastProbability < diseaseProbability:
        lastProbability = diseaseProbability
        disease = eachDisease
return disease
```

Fig. 8. Predicting Data

C. State Diagram

It explains different state of the system. First the user opens Disease Predictor. The user selects the symptoms. When finished selecting symptoms the user submits the symptoms. Disease Predictor analyzes the symptoms and displays the results.

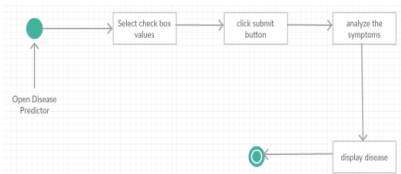


Fig. 9. State Diagram

D. Sequence Diagram

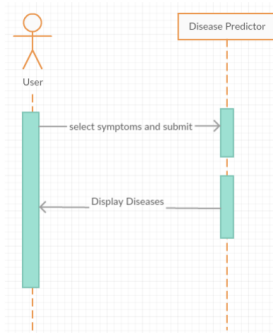


Fig. 10. Sequence Diagram

It explains the sequence of the Disease Predictor. Initially system shows the symptoms to be selected. The user selects the

symptoms and submits to the system .The Disease Predictor predicts and display the results.

V. RESULTS

- We gave some of the symptoms to the system to predict disease according to our algorithm.

```

[15]: nosb(brother[\"refstates\", \"itching\", \"skin_rash\", \"nodal_skin_signs\", \"dischroic_patches\"])

For disease Fungal infection calculated probability is: 0.9099999999919731
For disease Allergic reaction calculated probability is: 0.7368208265978161
For disease DMO calculated probability is: 0.7368208265978161
For disease Chronic osteoarthritis calculated probability is: 0.8739480120307846
For disease Drug Reaction calculated probability is: 0.7368208265978161
For disease Psoric urine disease calculated probability is: 0.348208365078161
For disease AIDS calculated probability is: 0.7368208265978161
For disease Diabetes calculated probability is: 0.7368208265978161
For disease Gastroenteritis calculated probability is: 0.7368208265978161
For disease Bronchial Asthma calculated probability is: 0.7368208265978161
For disease Hypertension calculated probability is: 0.7368208265978161
For disease Migraine calculated probability is: 0.7368208265978161
For disease Cervical spondylitis calculated probability is: 0.7368208265978161
For disease Paralysis calculated probability is: 0.7368208265978161
For disease Jaundice calculated probability is: 0.8739480120307846
For disease Multiple sclerosis calculated probability is: 0.7368208265978161
For disease Chicken pox calculated probability is: 0.5524476651604611
For disease Dengue fever calculated probability is: 0.8739480120307846
For disease Typhoid calculated probability is: 0.7368208265978161
For disease Hepatitis A calculated probability is: 0.348208365078161
For disease Hepatitis B calculated probability is: 0.348208365078161
For disease Hepatitis C calculated probability is: 0.348208365078161
For disease Hepatitis D calculated probability is: 0.348208365078161
For disease Hepatitis E calculated probability is: 0.348208365078161
For disease Alcoholic hepatitis calculated probability is: 0.348208365078161
For disease Tuberculosis calculated probability is: 0.7368208265978161
For disease Common Cold calculated probability is: 0.348208365078161
For disease Measles calculated probability is: 0.7368208265978161
For disease Dimorphic hemophilus(plis) calculated probability is: 0.7368208265978161
For disease Heart attack calculated probability is: 0.7368208265978161
For disease Varicose veins calculated probability is: 0.7368208265978161
For disease Hypothyroidism calculated probability is: 0.7368208265978161
For disease Rheumatoid arthritis calculated probability is: 0.7368208265978161

```

Fig. 11. Input of Symptoms

- After processing the system gave us the probability of all the diseases and the disease with most probability.

```

for disease alcoholic hepatitis calculated probability is: 7.35029852659718e-21
for disease tuberculosis calculated probability is: 7.35029852659718e-21
for disease common cold calculated probability is: 7.35029852659718e-21
for disease pneumonia calculated probability is: 7.35029852659718e-21
for disease idiopathic hemorrhoids(piles) calculated probability is: 7.35029852659718e-21
for disease heart attack calculated probability is: 7.35029852659718e-21
for disease varicose veins calculated probability is: 7.35029852659718e-21
for disease hypothyroidism calculated probability is: 7.35029852659718e-21
for disease hyperthyroidism calculated probability is: 7.35029852659718e-21
for disease hypoglycemia calculated probability is: 7.35029852659718e-21
for disease osteoarthritis calculated probability is: 7.35029852659718e-21
for disease arthritis calculated probability is: 7.35029852659718e-21
for disease vertigo Paronychia Positional vertigo calculated probability is: 7.35029852659718e-21
for disease acne calculated probability is: 8.37934032028785e-16
for disease urinary tract infection calculated probability is: 7.35029852659718e-21
for disease Psoriasis calculated probability is: 8.37934032028785e-16
for disease Impetigo calculated probability is: 8.37934032028785e-16

Out[15]: 'Fungal infection'

In [16]: 1 - 2.321852641469978e-07

Out[16]: 8.999997678347585

In [ ]:
```

Fig. 12. Disease Probability

VI. COMMAND LINE OUTPUT

[illegible]

Fig. 13. Command Line Output

We Will run CEP _ dict_ implementation.py by typing
python [filename] [testing_ data.csv] [Row number]
or
python CEP_dict_implementation.py ../DataSet/Testing.csv

program will read test.csv and read specific row then find all the symptoms of that row and train all model then predict disease.

VII. CONCLUSIONS

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease. Total diseases included were 41 with 133 different symptoms. Average prediction accuracy probability is very satisfactory.

Our proposed system helps to improve the accuracy of diagnosis and greatly helpful for further treatment. In future enhancements, the accuracy has to be tested with different dataset and to apply other AI algorithms to check the accuracy estimation.

In this work, we show that Automated Disease Prediction System can help people who are facing difficulties, better understand their physical condition by predicting potential diseases. We also show that our framework enables the system to perform significantly better than existing ones. Having said that, our system accuracy can be increased further as there is space left for improvement. Like the decision tree and parent tree generation is a cumbersome task but it is a continuous process, same goes with the enrichment of the database. It will get better and better over time and accuracy of disease prediction will also be on the rise.

VIII. RECOMMENDATIONS

This project has not implemented recommendation of medications to the user. So, medication recommendation can be implemented in the project. History about the disease for a user can be kept as a log and recommendation can be implemented for medications.

IX. FUTURE WORK

Every one of us would like to have a good medical care system and physicians are expected to be medical experts and take good decisions all the time. But it's highly unlikely to memorize all the knowledge, patient history, records needed for every situation. Although they have all the massive amount of data and information; it's difficult to compare and analyse the symptoms of all the diseases and predict the outcome.

So, integrating information into patient's personalized profile and performing an in-depth research is beyond the scope a physician. So the solution is ever heard of a personalized healthcare plan – exclusively crafted for an individual. Predictive analytics is the process to make predictions about the future by analyzing historical data. For health care, it would be convenient to make best decisions in case of every individual. Predictive modeling uses artificial intelligence to create a prediction from past records, trends, individuals, diseases and the model is deployed so that a new individual can get a prediction instantly. Health and Medicare units can use these predictive models to accurately assess when a patient can safely be released.

X. ACKNOWLEDGEMENT

we express our sincere gratitude towards our guide of Prof. Dr. Mudassar Farooq for his constant help, encouragement and inspiration throughout the project work. Also we would like to thank him for valuable guidance , ability to motive us and even willingness to solve difficulty made it possible to make our project unique and made task easier.

XI. REFERENCES

- <https://github.com/Abdullah296/Disease-Prediction>
- <https://www.kaggle.com/neelima98/disease-prediction-using-machine-learning>