DISEASE DIAGNOSIS SYSTEM BY EXPLORING MACHINE LEARNING ALGORITHMS

ALLEN DANIEL SUNNY
Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore, India
allensunny1242@gmail.com

SAJAL KULSHRESHTHA
Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore, India
ksajalnm@gmail.com

SATYAM SINGH
Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore, India
warframesam@gmail.com

SRINABH
Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore, India
isrinabhjha@gmail.com

Mr. MOHAN BA
Associate Professor, CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
ba.mohan@gmail.com

Dr. Sarojadevi H.
Professor, CSE
Nitte Meenakshi Institute of Technology
Bangalore, India
sarojadevi.n@nmit.ac.in

***Abstract*—In this paper, aspects of the design of a system for diagnosis of Common disease that can be detected by the Doctor and patient on entering the symptoms into the system. Before developing the system, an analysis is done by comparing the machine learning algorithms on Disease-Symptom Knowledge Data-set prepared by New York Presbyterian Hospital of patients admitted during 2004. Based on the analysis the most accurate algorithm is used in the system to achieve the reliability. Also, in this report we will be discussing the hurdles we faced to achieve the required results, as in our project the machine learning algorithms are implemented on the pure text based data set. This project can possibly help doctors and patients as well, as early detection is beneficial for right treatment and early recovery.**

***Index Terms*—Disease Diagnosis System, Machine learning Algorithms, Naive Bayes, Apriori**

I. Introduction

Machine learning is one of the key methods used in modern day analysis. With the explosion of the tech industry, and the rise of big data; it became necessary to analyze and predict trends. Slowly over the years machine learning has branched out into almost every major industry, and performs functions that were almost unheard, compared to a mere few years ago.

In this project we have concentrated exclusively upon the supervised algorithms. Supervised algorithms can be broadly broken down into two sub divisions-

- Regression algorithms- Regression algorithms are used to predict continuous values. For example , Nave Bayes and KNN algorithms.
- Classification algorithms- Classification algorithms are used in order to predict discrete values. For example, linear regression and K Means algorithms.

For this project we have extensively used scikit Learn which is a module that is built on top of sci py library in python version 2 onwards. Scikit Learn is a python library that exclusively focuses on data science and the various classification, regression and cluster-ing algorithms including support vector machines, k-NN, random forests, Logistic Regression, gradient boosting, Nave Bayes, k-means, and Decision Tree, and is designed to operate within the python script for the Python numerical and scientific libraries NumPy and SciPy.

Libraries used

- NumPy which is Based on n-dimensional array package
- SciPy which is Fundamental library for scientific computing
- IPython is used for the Enhanced interactive console
- Sympy gives us the advantage of using Symbolic mathematics

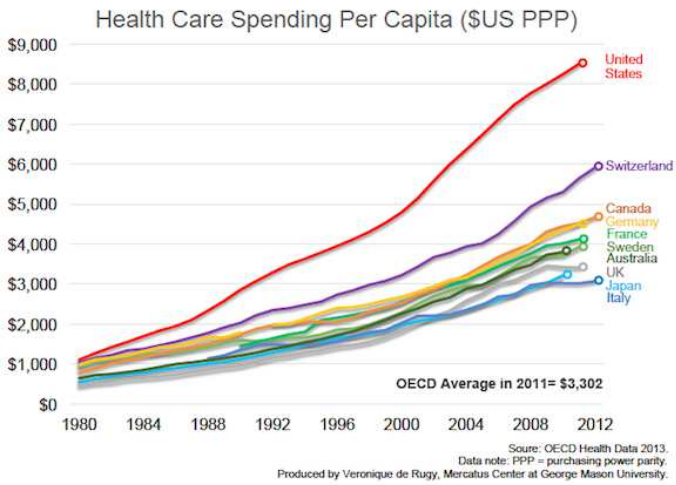
A. Brief history

In recent years we have observed a growing and worrying trend across the developed world. Medical costs are being driven sharply upwards in countries, who are recognized across the world as leading pioneers in science and industry. Countries include the United States and western European nations. The rising cost of medical treatment can be attributed to a number of factors including very high pricing for pharmaceutical drugs and the use of more expensive medical techniques , to name a few. Increasing medical costs seem to be thus intertwined with the overall development of a country.

As India slowly joins the ranks of world powers, it is not inconceivable to think that we might eventually face the

same problem that the other developed nations grapple with currently. In order to preempt such a situation, it is vital to introduce more economic and viable alternatives within the medical industry.

Fig. 1. Health Care Spending per Capita [6]



With the growth of machine learning, we can introduce disease diagnostic tools to the general public, lowering the need for them to depend on expensive medical treatment.

B. Problem Statement

This project will also attempt to answer questions based on the use of various algorithms including,

- Which is the best suited algorithm for the text based data ?
- What is the most optimum method of coding a given algorithm ?
- What is the most optimum way of converting text to tokens and then running an algorithm with it ?
- Various ways in order to pre-process the data ?

II. RELATED WORK

In [1], the authors propose an ontology, called Generic Human Disease Ontology (GHDO), which deals with the knowledge about human diseases with their types, causes, symptoms and treatments.

In [2],the database deals with number of patient cases as proto-type and stored in a separate database. The Disease of the Patients can be predicted with the help of patient datbase. The system has made use of production rules and a neural network.

In [3], the authors have tried to develop the system using artificial intelligence to identify the cause of the headache based on the data input by the patient.

In [4], the approach to develop the disease diagnosis system based on symptom is carried out by making the use of Bayesian network. To make machine understand, it creates the knowledge graph by the extracted information.

III. SYSTEM REQUIREMENTS SPECIFICATIONS

The requirements depends upon the various factors which are taken into consideration, for this project the requirements are based on algorithm used, programming language used and the data-set on which the project is based on.

A. Product Perspective

The final application is coded on Python 3.0 and Python 2.7, so until the newer version of this application is developed which will be independent of the required environment, the product needs Python3.0 and Python 2.7 installed on the system. The user can use any IDE such as Spider and IDLE or can use the freemium open source distribution such as Anaconda, the application will run smoothly.

B. User characteristics

The Users of this system will be mostly doctors and the people who may encounter with some symptom of the disease which they are unaware off. It will be helpful for the doctors as the doctor can be reminded and cross-check with the possible diseases (to overcome human errors like diligence, versatile, tiredness). This can also be helpful for the patients to find out the diseases, when no other help is possible and also can go to specialist of that particular disease instead of going to the general doctor and getting referred which results in time and money wastage and causes human effort. The User need to input the list of Symptoms which they are experiencing. The output of the input given will be the probable disease with the more-likely probability.

C. Functional Requirements

For the proper functioning of the machine learning algorithms, the foremost requirement is conversion of the text data into the numerical data, which is nothing but the data manipulation and the data pre-processing. This is done because the computer understands only the language of 0s and 1s; so, it requires numbers to work on, as the system do not understand words.

D. Performance Requirements

The Disease Diagnosis system tells the probable diseases based on giving Symptom as input. This is related to the lives of human being, so the system is expected to be reliable. If the system is not reliable the consequences can be enormously dangerous as by wrong treatment and medication, the life of the being may come to critical condition.

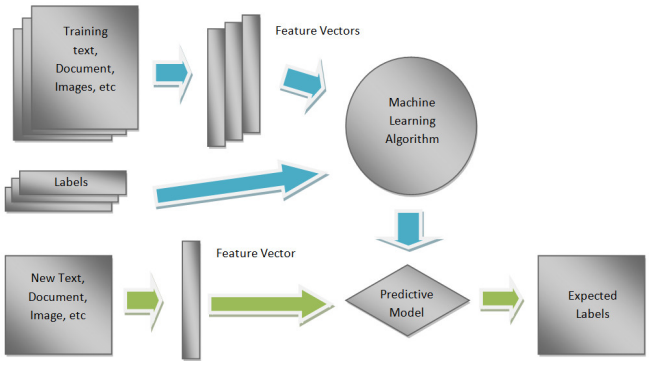
IV. Architectural Design

A. Architecture and characteristics

From the diagram we can see the general representation of the project through its various phases.

1) We have the given data set and the test data. Test data is used to train the given data set, by taking a small portion of the given data, and finding the underlying trends in said data. The logic is then applied to the entire data set, and a result is derived.
2) Data pre-processing must be carried out on both the test data and the data set. Pre-processing removes the inherent errors of the dataset, as well as converting the data set to a form that can be read by the computer.

Fig. 2. Supervised Learning Algorithm

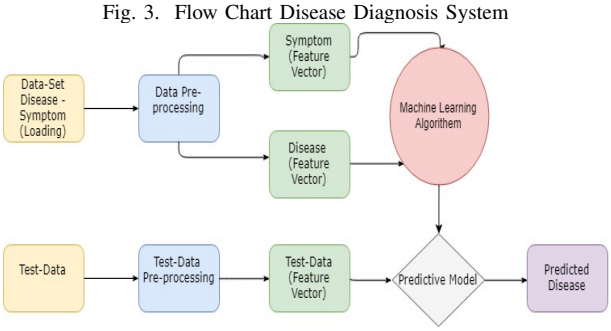


3) The given data set uses vectors in order to map values and deliver a coherent picture. For the test data, a simple text data vector is taken and results are mapped from that. Due to the size of the original data set, there we use 2 vectors, mainly the symptoms, and the disease. Thereby matching each symptom to the required disease. The probability that the disease is the one specified is then generated, based upon the algorithm used.
4) The vectors are then fed into a machine leaning algorithm. For this project, we have used algorithms such as Nave Bayes and Apriori. The machine learning algorithm then outputs a predictive model, based upon the algorithms parameters.
5) Predictive model, the predictive model is used to generate an accurate prediction og the given data set. The predicted model is then compared with the actual model, and the accuracy of the said system is measured. Measured accuracy is then determined to be an accurate value of the said system.
6) The final output is of course the predicted disease. It is then needed to see if the outputted disease is the same one that was should be generated from the inputted symptoms.

B. Architectural Design

System architecture refers to the underlying conceptual model of the system. This can refer to ta systems views,

Fig. 3. Flow Chart Disease Diagnosis System



the structure of the system, as well as how the system behaves under certain conditions. In other words, a strong grasp of the system architecture can help the user understand the scope as well as the limitations of the said system.

Some of the various parts of our system are-

- Data Set: One of the key principles of data mining is the ability to extrapolate the underlying trends inherent in the data, by sampling a relatively small size of the data. The conclusions drawn from the small sample are then mapped onto the entire data set.

- Data preprocessing: Real world data requires cleaning and preprocessing before it can be handled by the algorithm. This is due to the obvious reason that real world data often contains errors that the algorithms that are used, cannot process such errors. Data Preprocessing thus takes this raw data and further processes it, removing errors , and saves it for further analysis.

  Data goes through a sequence of steps during preprocessing:
  - Data Cleaning: Data is cleansed through various methods. These methods include the addition of data, for example filling in data slots that are left blank. Or data reduction, for example the removal of commas or other unknown characters.
  - Data Integration: In this step, the data is integrated from various sources. The data is then checked for any integration errors and they are summarily handled.
  - Data Transformation: Data in this step is normalized based upon the given algorithm. Data normalization can be done through a variety of methods. This step is required in most data mining algorithms, as the data inputted needs to be as clean as it can. Data is then combined and built up.
  - Data Reduction: This step in the process aims to

reduce the data to more manageable levels.

- Data set and test data: The data set is separated into parts which is training and testing data sets. The training data is used to calculate the underlying patterns of the same. Testing, then sees if the patterns hold. Similar data is needed for training and testing, usually derived from the same data set.

  After the model has been preprocessed, the second step is to test the accuracy of the system. We can do this by plugging in various examples that are run against the testing data set.

- Predictive model: Predictive modeling techniques are used for the calculation of the probability of a given occurrence happening, based on the set of input factors. This makes it highly useful in disease prediction, as it can be used to predict what disease the individual is suffering from, based upon a persons input symptoms and previous diagnoses.

- Predicted Disease based upon the various algorithms in the end we have a final output of a disease that has been predicted based upon certain input symptoms.

V. ALGORITHMS USED

A. K-Nearest Neighbor

The KNN algorithm is one of the backbone of data mining and predictive analysis. The KNN algorithm stores all previous cases of such data and based on such, generates new cases and patterns. The KNN algorithm was one of the earliest cases of a predictive modelling algorithm. It uses statistical formula in order to measure the relative distance between each of the data nodes. The formulas used can be of various types including,

- Euclidean Distance

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

- Manhattan Distance

$$\sum_{i=1}^{k}|x_i - y_i| \tag{2}$$

- Minkowski Distance

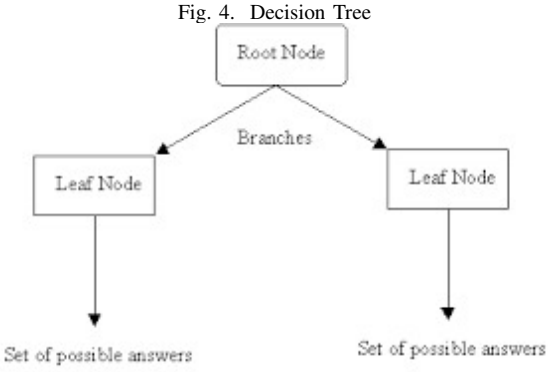
$$(\sum_{i=1}^{k}|x_i - y_i|^p)^{\frac{1}{p}} \tag{3}$$

Data can only be analyzed if they are continuous variables. For non- continuous variables, we need to first standardize the data, Hamming distance and other standardizations are to be used in this case.

B. Decision Tree

Decision trees are used to approximate a discrete valued function, in which the final learned function is the decision tree itself. Decision trees are more useful than regular algorithms, as they can have improves readability. The tree does this by reducing all values to individual nodes, and then generating a simple if else case for each node pair. The Decision tree has been very useful in the various fields, from medical diagnosis (How we first approached the algorithm) and other fields such as stock market analysis.

Fig. 4. Decision Tree



Unfortunately some problems with decdion trees include , the problem that datafitted needsot be continous ,choosing the required attribute selecetion , seeing how deep the tree needs to grow and as with any algorithm , the computational speed at which it processes data.

C. Naive Bayes

Nave Bayes is broadly based on the Bayes algorithm with one key difference. The Nave Bayes algorithm assumes that relationship between different parameters of the system do not exist. It is explicitly useful for large data sets, and despite its simplicity, it exceptionally performs other more sophisticated algorithms regularly. To calculate the conditional probability the formula is given below.

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \tag{4}$$

where

- P(H) is the probability of hypothesis H being true. This is known as the prior probability.
- P(E) is the probability of the evidence.
- P(E—H) is the probability of the evidence given that hypothesis is true.
- P(H—E) is the probability of the hypothesis given that the evidence is present.
  As with other algorithms, it is required for the data to be preprocessed before it can be inputted into the algorithm. Here, the various numerical predictors is

required to be converted into their categorical counterparts, and then inputted into various frequency tables. The main method by which we can achieve this, is by normal distribution.

- Mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

- Standard Deviation

$$\sigma = [\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)]^{\frac{1}{2}} \tag{6}$$

- Normal Deviation

$$(\sum_{i=1}^{k} |x_i - y_i|^p)^{\frac{1}{p}} \tag{7}$$

Observe the above equations, as the root of the Nave Bayes equation. These equations are the final solution of the nave bayes equation. The root values of the data set are plugged into the equation and the data is then normalized. Like all data mining equations, the data must first be preprocessed and then set into its function.

D. Apriori

The Apriori Algorithm is an significant algorithm for mining frequent itemsets for boolean association rules. Apriori uses a "bottom up" approach, where frequent subsets compute one item at a time, this step is known as candidate generation, and groups of candidates are tested beside the data.

The support of an itemset is the share of transaction in the database in which the item X appears. It signifies the popularity of an itemset.

First it creates a frequency table of all the items that take place in all the transactions. Then only those elements are significant for which the support is greater than or equal to the threshold support. The next step is to make all the possible pairs of the significant items keeping in mind that the order doesnt matter. After which we count the occurrences of each pair in all the transactions. only those itemsets are considered which cross the support threshold.

VI. ANALYSIS

The K-nearest neighbor and decision tree was executed by using the function provided by the sci-kit learn library, which gave very poor results. we came to the conclusion that this happened because the dataset is Many-to-many structured and also is text-based where as the in-built functions best works on one-to-one and one-to-many numerical dataset.

After which Naive Bayes was implemented manually (not by using the built-in function of sci-kit learn) and also the data was preprocessed according the need of the algorithm. This gave good results but with some limitations. The limitations were if the symptoms belonging to different diseases are entered into a system as an input then it was difficult for the system to find the probable disease which was because of its way of implementing that is, it uses conditional probability.

Then the Apriori algorithm was implemented which gave the best results as the algorithm is based on the concept of frequent item dataset. It not only gave the possibilities of multiple diseases but also gave the probability of its occurrence.

VII. RESULTS

We used supervised machine learning algorithms to develop a system that allows the doctor to predict probable diseases based on patient symptoms as observed which results in better diagnosis and further treatment. This system was designed and implemented by analyzing the dataset obtained from New York Presbyterian Hospital through the data collected by patient diagnosis in the year 2004 by exploring various supervised learning algorithms.

Based on our exploration of various supervised machine learning algorithms on the dataset, Apriori and Naive Bayes algorithms showed better results compared to k-nearest neighbors and Decision Tree algorithms which showed relatively poor performance. Hence final system was designed through the implementation of Naive Bayes and Apriori algorithms that were finally used for disease diagnosis.

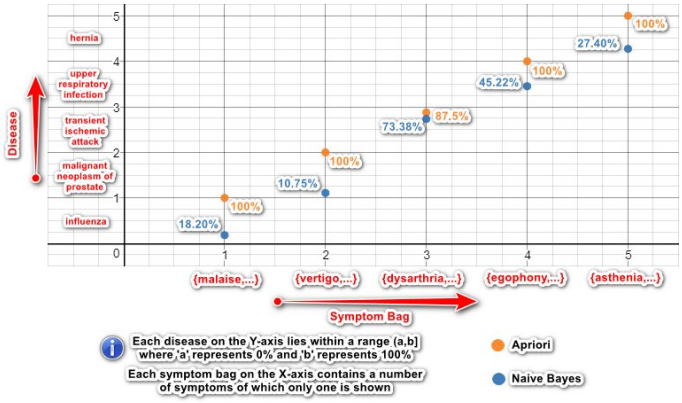
To compare the two working algorithms of our project, The Naive Bayes and The Apriori Algorithm we have plotted the scattered graph between the symptoms and the predicted disease and the probability of the occurrence of the predicted disease.
Below is the table of the set of symptoms, Diseases and the the probability of the occurrence of the predicted disease.

Based on the below table the Scattered Graph is plotted having symptoms on the x-axis and diseases on the y axis.

Based on the analysis we can say that both algorithm are working fine but apriori algorithm is giving better results as the this algorithm is based on frequent item dataset. Moreover it gives us probability of symptom corresponding to all the disease. In that way a person will be able to know if he/she is having more than one disease.

DISEASE	SYMPTOMPS	PROB. NAIVE BAYES	PROB. APRI-ORI
INFLUENZA	uncoordination, pleuritic pain, snuffle, throat sore, malaise, debilitation, dysuria	18.20	100.0
MALIGNANT NEOPLASM OF PROSTATE	passed stones,qt interval prolonged, dysuria, vertigo, paresis, hemianopsia homonymous, tumor cell invasion, hemodynamically stable, orthostasis	10.75	100
TRANSIENT ISCHEMIC ATTACK	syncope, room spinning, headache, Stahliś line, extrapyramidal sign, dysarthria, speech slurred, vertigo	73.38	87.5
UPPER RESPI-RATORY INFECTION	rapid shallow breathing, egophony, indifferent mood, labored breathing, cystic lesion	45.22	100
HERNIA	gag, hyperventilation, excruciating pain, nausea, posturing, pain, pain abdominal, asthenia	27.40	100

Fig. 5. Naive Bayes and Apriori Analysis



VIII. CONCLUSION AND FUTURE WORK

Based on the various algorithm implementations, we have concluded that algorithms such as Nave Bayes and Apriori are highly useful in the implementation of the given data set. Based on the objective, we can conclude that all goals have been met, with the disease being predicted based on the input symptoms, using multiple algorithms.

Since machine learning algorithms are dependent on datasets based upon previous knowledge obtained from the physical diagnosis, the volume of dataset must be comprehensive with minimum number of outliers. As the data set grows more and more, diseases are added, the scope of diagnosis increases with better predictions in the upcoming future.

REFERENCES

[1] Maja Hadzic and Elizabeth Chang. *Ontology based Multi agent Systemn support Human Disease Study and Control*. Proceedings of the conference on Self Organization and Autonomic Informatics 2005, pp.129-141.

[2] Dipanwita Biswas, Sagar Bairagi , Neelam Panse and Nirmala Shinde. *Disease Diagnosis System* International Journal of Computer Science Informatics, Volume-I, Issue-II, 2011

[3] Anthony Farrugia, Dhiya Al-Jumeily, Mohammed Al-Jumaily and David Lamb *Medical Diagnosis: are Artificial Intelligence systems able to diagnose the underlying causes of specific headaches?* Sixth International Conference on Developments in eSystems Engineering 2013.

[4] Prerna Agarwal, Richa Verma, Anupama Mallik. *Ontology Based Disease Diagnosis System with Probabilistic Inference* 2016 1st India International Conference on Information Processing (IICIP).

[5] New York Presbyterian Hospital 2004. *Disease-Symptom Knowledge Database* AMIA Annu Symp Proc. 2008. p. 783-7. PMCID: PMC2656103. http://people.dbmi.columbia.edu / friedma /Projects /DiseaseSymptomKB /index.html

[6] Veronique de Rugy, Mercalus Center at George Mason University. *Health Care spending per capita ($US PPP)* OECD Health Data 2003.