Web Based Disease Detection System

Guided by Hasnuhana Mazumder
Sayantan Saha, Argha Roy Chowdhuri, Anindita Dey and Sourav Haldar
Department of Computer Science and Engineering
Dream Institute Of Technology, Kolkata, India

***Abstract*-This web based Disease Detection System aims at building up a website where user can know about the disease he is infected with by submitting the symptoms that he is experiencing. The system has been developed using clean technology of ID3 (Iterative Dichotomiser 3) algorithm.The concept and the essential activities required for its successful implementation have been elaborated.**

Keywords-Disease Detection, ID3, Medical Website, Clean Technology.

I.INTRODUCTION

Given the current lifestyle that people follow nowadays, health problems are ever increasing and along with it a busy work schedule forces them to take random medicine without consulting a medical practitioner. This calls for a new approach in the field of medical diagnosis.

Who could have predicted that a communication system designed to serve scientists and the military would one day help a common man to cure his medical problems?

With the power of Internet available to almost every individual (through computers, tablets and mobile phones) a website with Disease Detection System will open up a whole new form of diagnosis.In this Internet savvy world, websites are the only common mode of attachment between different people and communities in our society.People of all ages and mostly all classes follow the Internet vastly. So building a website for curing health-related problems is indeed a good idea in today's problem prone world. So we thought of designing a website where people who couldn't make up time to visit a medical professional can avail the facility of getting a correct diagnosis of them and will be spared from taking random medicines which may have some side effects, degrading their health to a greater extent.

This research demonstrates this new Disease Detection System through a prototype model. The detection system has been developed using ID3 (Iterative Dichotomiser 3) algorithm. The algorithm takes into consideration training set and based on it teaches the machine to identify disease based on symptoms. Once the disease is detected we search for corresponding remedy for the disease. The user is advised to try the remedy given for a certain period of time and even if the symptoms persist consult a doctor.

II. RELATED WORKS

There are quite a few number of medical websites already on the internet. Some examples of these websites are references [4], [5], [6].

There are various other websites but they are more or less having the same functionality as the websites mentioned above. The major drawbacks of these websites are:

a. These websites are just informatics websites. Here you can search for different diseases their symptoms and remedy.
b. The Dr. Batra's website has a query option but here it is in the form of enquiry form
c. None of these websites have any dynamic Disease Detection System.

To the best of our knowledge there is no existing website which has a dynamic Disease Detection System. Where user can input their symptoms and get to know about the disease they are infected with.

III. DISEASE DETECTION MODEL

While designing the model we assumed that the user has a clear idea about the symptoms that he is experiencing. The system is first trained to detect disease based on symptoms and afterwards when the user enters his symptoms these symptoms are processed by the machine according to the training that he has received.

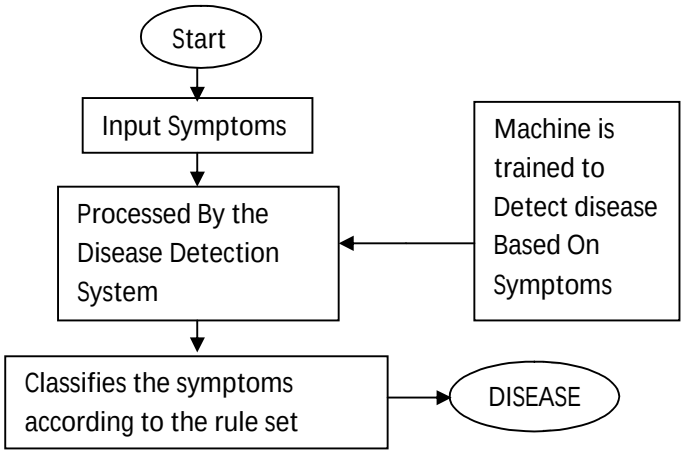

Fig 1: Flow of data in Disease Detection System

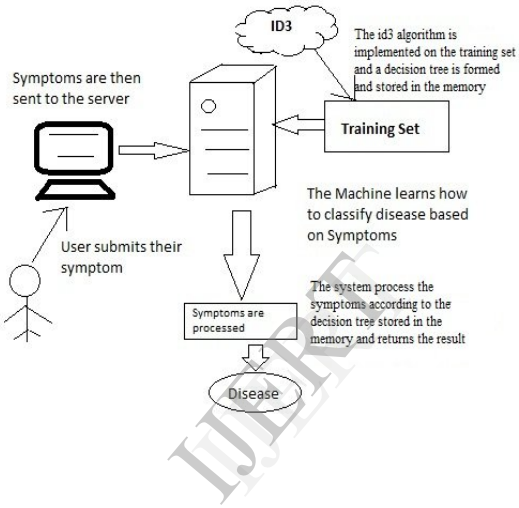

Fig 2: The Disease Detection Model.

After the system is trained with the training set using ID3 algorithm a rule set is formed and when the user enters his/her symptoms those symptoms are processed according to the rule set developed.

When a new disease along with its symptoms are added to the training set a new rule set is developed using the same training mechanism.

IV. METHODOLOGY AND IMPLEMENTATION

The Web Based Diseased Detection System is implemented using ID3 (Iterative Dichotomiser 3) algorithm.

A. Training Set and Test Data

Training set is a document on which the algorithm will be implemented. The training set is prepared in such a manner so that the machine can efficiently learn to classify instances based on attributes. The first rows of the Training Set Document are called the attributes. The attributes are classified to build a decision tree. So these attributes forms the decision node or the non-terminal node of the decision tree. The last column is called the class attribute. This attributes are the terminal node or the result of the decision tree. The training set is prepared in .arff (Attribute Relation File Format) or.csv (comma separated value) format. Every row of data in the training set are called instance.

After training the machine is tested with a set of data having all the attributes of the training set only the class attribute is not specified. If the machine can correctly identify the class attributes based on the training then we can conclude that the machine has learned the system.

B. Iterative Dichotomiser 3(ID3) Algorithm

In Decision Tree Learning, ID3 is an algorithm Invented by Ross Quinlan. It is a non-incremental classification algorithm.

The ID3 algorithm begins with the original set *S* as the root node. On every iteration of the algorithm, it iterates through every unused attribute of the set *S* and calculates the entropy *H(S)*(or information gain *IG(A)*) of that attribute. Then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute to produce subsets of the data. The algorithm continues to recurse on each subset. When every element in a subset belongs to the same class, this subset will no longer be recursed on, and this node in the decision tree becomes a terminal node with a class label same as the class all its elements belong to. The ID3 algorithm terminates when every subset is classified. Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

The ID3 algorithm is used by training on a dataset S to produce a decision tree which is stored in memory. At runtime, this decision tree is used to classify new unseen test cases by working down the decision tree using the values of this test case to arrive at a terminal node that tells you what class this test case belongs to.

ALGORITHM

1. Calculate the entropy of every attribute using the data set *S*
2. Split the set *S* into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes

C. Entropy Calculation

Entropy *H(s)*is the measure of the amount of uncertainty in the (data) set *S*.

The formula for calculating Entropy is:

….. (1)

Where:

- *S*=The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)
- *X*= Set of classes in *S*
- *P(x)*- The proportion of the number of elements in class *X*to the number of elements in set *S*.

EXAMPLE

If *S*is an example of 10 instances and we have with 4 YES and 6 NO instance then the Entropy will be:

Using Equation (1) we get,

H(S) = - (4/10) - (6/10)

= - (0.4*-1.3219) – (0.6*-0.736)

= 0.97036

D. Calculation for Information Gain

Information gain *IG(A)* is the measure of the difference in entropy from before to after the set *S* is split on an attribute *A*. In other words, how much uncertainty in S was reduced after splitting set *S*on attribute *A*

The formula for calculating Information Gain is:

….. (2)

$$Gain(S, A) = Entropy(S) - \Sigma ((|S_v| / |S|)*Entropy(S_v))$$

…… (3)

Where,

- *H(S)*- Entropy of set *S*
- *T*- The subsets created from splitting set *S*by attribute *A* such that
- *P(t)*- The proportion of the number of elements in *t*to the number of elements in set *S*
- *H(t)*- Entropy of subset *t*

In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the **largest** information gain is used to split the set *S*on this iteration.

EXAMPLE

Suppose S is a set of 14 examples in which one of the attributes is *Sore Throat*. The values of *Sore Throat* can be *Yes or No*. The classification of these 14 examples are 9 YES and 5 NO. For attribute *Sore Throat*, suppose there are 8 occurrences of Sore Throat = No and 6 occurrences of Sore Throat = Yes. For Sore Throat = No, 6 of the examples are YES and 2 are NO. For Sore Throat = Yes, 3 are YES and 3 are NO.

$$Gain(S, Sore\ Throat) = Entropy(S) - (8/14)*Entropy(S_{No}) - (6/14)*Entropy(S_{yes})$$

= 0.940 - (8/14)*0.811 - (6/14)*1.00

= 0.048

Entropy(S_{No}) = - (6/8)*log2(6/8) - (2/8)*log2(2/8) = 0.811

Entropy(S_{Yes}) = - (3/6)*log2(3/6) - (3/6)*log2(3/6) = 1.00

For each attribute, the gain is calculated and the highest gain is used in the decision node.

IMPLEMENTATION

The reason for choosing ID3 algorithm is it can easily build prediction rules based on the training data, builds the fastest and the shortest tree. Only need to test enough attributes until all data is classification finding leaf nodes enables text to be pruned, reducing number of tests, whole dataset is searched to create tree. Mathematical algorithm for building the decision tree.

The decision tree has been constructed using WEKA(Waikato Environment for Knowledge Analysis) GUI TOOL; it is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand is free software available under the GNU (General Public License).

A. Disease Detection Using ID3 and Weka.

1. Select the dataset for which the test is to be retrieved.
2. By using Weka and ID3 algorithm sort the specific pattern and classify dataset based on symptom.
3. Processed the symptoms entered by the user based on this classification and retrieve the class attribute.
4. The retrieved class attribute is the disease. Now search the database for a remedy of this disease and return it to the user.

B. Result Obtained From Id3 algorithm through Weka

Disease Detection System using Id3 algorithm through Weka is implemented using Java platform. It is web-based, user friendly and flexible. The Dataset that we have collected consists of 24 different diseases along with their symptoms. These data are shown in Table- 1. The ID3 classification algorithm was fed with this training set and it generated the following decision tree (Fig-3)

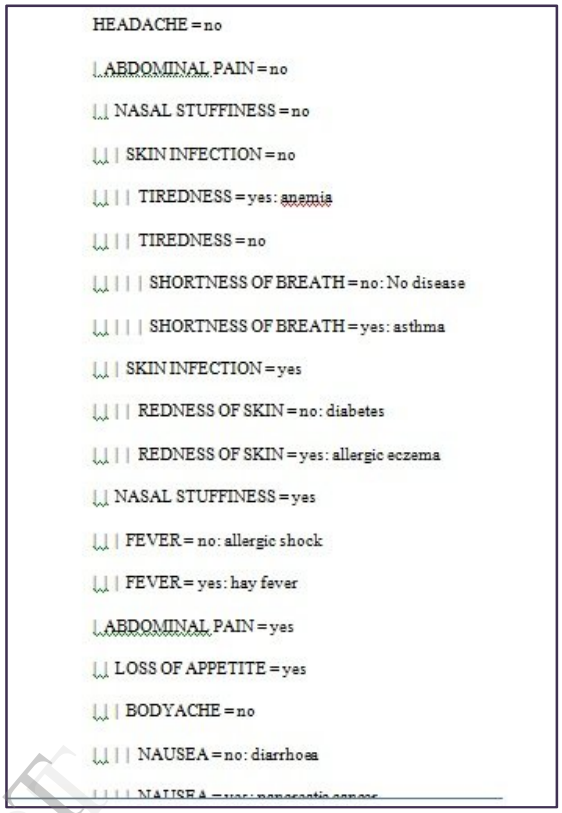

Fig-3: A part of the decision tree generated by id3

The Figure (Fig- 4) given below shows the accuracy of the algorithm and the time taken to build the tree. The Disease Detection is performed based on the tree that has been generated. The leaf node of this tree are the diseases and the non-terminal nodes are the symptoms that were considered to reach a decision.

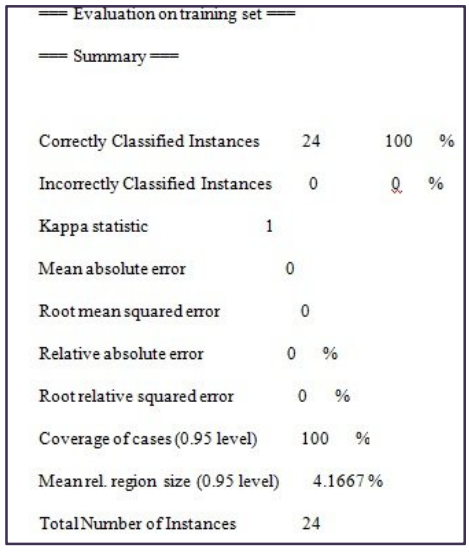

Fig-4: The classification accuracy of ID3

The above statistics shows that the Id3 algorithm has correctly classified all the 24 instances with a Kappa Statistics of 1. While generating this decision tree Id3 algorithm considered only 14 attributes out of the 51 attributes that were present in the dataset.
The above generated decision tree was converted into a rule set using If-Else structure in java platform and was implemented in Web Based Disease Detection System. A corresponding User Interface was developed where user has to input all the symptoms based on which the diagnosis will be done and the result along with remedy was also shown on the user interface.

Table1: The dataset that was used to train the Disease Detection System using Id3Algorithm (double click to view the entire table)

HEADACH	BODYACH	REDNESS	ITCHY SKI	RASH	SWELLING	STOMACH	NAUSEA	VOMITIN	BLOOD PR	DEHYDRA	WEIGHT L	BLADDER	SKIN INFE	DIZINESS	BLURRED	SHAKING
no	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	yes
no	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	no
yes	yes	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	yes
no	no	yes	yes	yes	no	no	no	no	normal	no	no	no	yes	no	no	no
no	no	yes	no	no	yes	yes	yes	yes	low	no	no	no	no	no	no	no
no	no	no	no	no	no	no	yes	yes	normal	yes	yes	yes	yes	no	no	no
yes	no	no	no	no	no	no	no	no	normal	no	no	no	no	yes	yes	no
yes	yes	no	no	no	no	no	no	no	normal	no	no	no	no	yes	no	no
yes	yes	no	no	no	yes	no	yes	yes	normal	no	no	no	no	yes	no	yes
yes	yes	no	no	no	no	yes	yes	yes	normal	yes	yes	no	no	no	no	no
yes	yes	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	yes
yes	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	no
yes	no	no	no	yes	no	yes	yes	yes	normal	no	no	no	no	no	no	no
yes	yes	yes	yes	yes	yes	no	no	yes	normal	no	no	no	no	no	no	no
no	no	no	no	no	no	no	yes	no	normal	no	yes	no	no	no	no	no
yes	yes	no	no	no	no	no	no	no	normal	no	yes	no	no	no	yes	no
no	yes	no	no	no	no	no	no	no	normal	no	yes	no	no	no	no	no
no	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	yes
no	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	yes	normal	no	no	no	no	no	no	no
no	yes	no	no	no	no	no	no	no	low	no	no	no	no	no	no	no
yes	yes	no	no	no	no	yes	no	yes	normal	no	no	no	no	no	no	yes
no	no	no	no	no	no	no	no	no	normal	no	no	no	no	no	no	no

V. RESULT AND ANALYSIS

This section presents the experimental results with the various classification of the disease based on attribute selection. The result focuses on the classification of disease on the basis of symptoms and the rule set that was defined using the decision tree.

In Fig 5 the user selects the symptoms that he or she is experiencing.

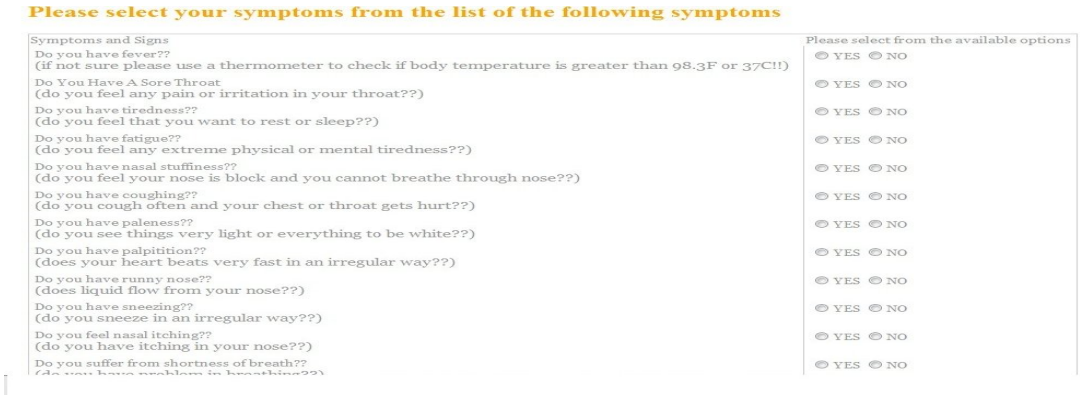

Fig 5: Step 1- The user selects the symptoms that he/she is experiencing

After the selections has been made and submitted the symptoms are processed and the disease is searched based on the rule set that has been defined. After the disease is detected the remedy database is searched for the corresponding remedy of the detected disease and then the disease along with its remedy is returned back to the user. This step is depicted in Fig 6.

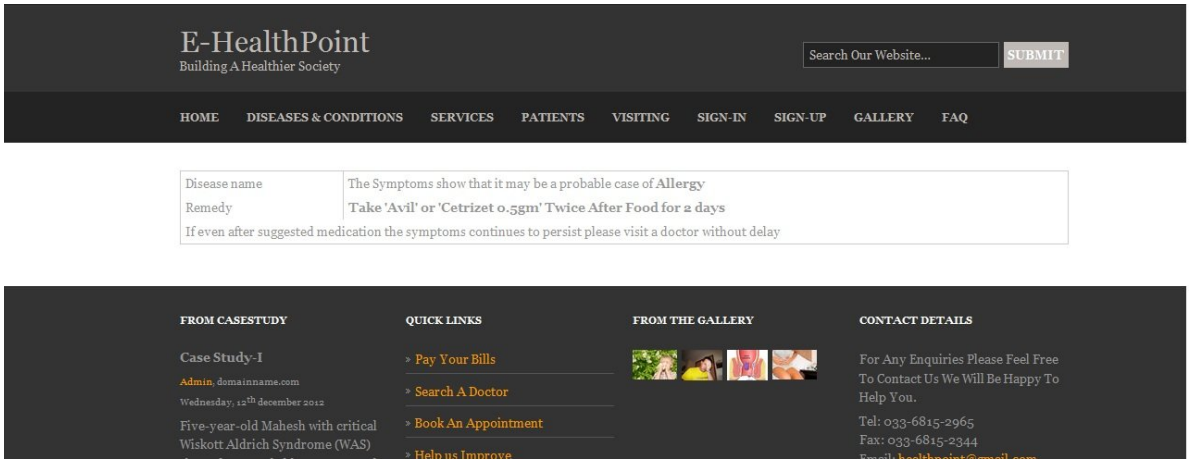

Fig 6: The web page displaying the disease and the remedy for the corresponding disease

The result shows that the classification done by the Id3 algorithm and the decision tree generated was correct since after training the algorithm could correctly classify disease based on symptoms.

VI. CONCLUSION

The historical development of Machine Learning and its application in medical diagnosis shows that from simple and straightforward to use algorithms, systems and methodology have emerged that enable advanced and sophisticated data analysis. In the future, intelligent data analysis will play even a moreimportant role, due to the huge amount of information produced and stored by modern technology.Current machine learning algorithms provide tools that can significantly help medical practitioners toreveal interesting relationships in their data. solving newproblems. The physicians found that the combination of classifiers was the appropriate way of improving the reliability and comprehensibility of diagnostic systems. Machine learning technology has not been accepted in the practice of medical diagnosis to an extent that the clearly demonstrated technical possibilities indicate. However, it is hard to expect that this disproportion between the technical possibilities and practical exploitation will remain for very much longer.

VII. REFERENCES

[1] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106

[2] Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997. pp. 55–58.

[3] Grzymala-Busse, Jerzy W. "Selected Algorithms of Machine Learning from Examples." *FundamentaInformaticae* 1 8, (1993): 193–207.

[4] Dr. Batras (URL: www.drbatras.com)

[5] Medicinet(URL: www.msdicinenet.net)

[6] SymptomChecker(URL: symptoms.webmd.com)

VIII. ACKNOWLEDGEMENT

We would like to thank our Director Sir, Mr. DipankarSarkar who had not only encouraged us but also guided us in our research. A special thanks to our Project guide Ms. HasnuhanaMajumdar for her expert guidance and motivation.