Estimation based on Data Mining Approach for Health Analysis

Priyanka Vijay Pawar
Department Of Computer Engineering
Ramrao Adik Institute of Technology
Nerul,Navi Mumbai
Email: pawarp0712@gmail.com

Megha Sakharam Walunj
Department Of Computer Engineering
RamraoAdik Institute of Technology
Nerul,Navi Mumbai
Email:meghaswalunj@gmail.com

Pallavi Chitte
Department Of Computer Engineering
RamraoAdik Institute of Technology
Nerul,Navi Mumbai
Email: pallavi.chitte@gmail.com

Abstract—In this world of global marathon, everyone is too close to smart technology but are moving miles away in taking care of our health. The traditional approach has been replaced by smart technologies integrated in every discipline of science. In this paper, we present a methodology to predict diseases based on user input symptoms. We have built a prototype to demonstrate the efficiency of these methods which will inform users about the disease they are suffering from. Itpredicts probable diseases by mining data sets and provides suggested doctors and remedial solutions. It will also guide the users by giving tips to live a healthy life, some diet tips and also usefulness of plants and food items. We are identifying probability of diseases using Apriori technique.

Keywords—*Apriori, mining, predict, probable, symptoms*

*****

I. INTRODUCTION

Everyone falls ill sometime or the other and we all want quick analysis. We all are under the wrong impression that doctors are expertise in their domains and have good knowledge of all their solutions. Today, the rat race is so fast and furious that everyone just wants good amount of salary, name and fame but neglect and do not pay any heed to good health. Technology is far ahead and medical science is today doing miracles. But it can't possibly happen that everything cannot be at their fingertips. Even if we have access to the massive amounts of data we still need expert teams to predict analysis in various domains. This kind of detailed research and statistical analysis is more than what a human mind can think. It is the reason everyone is moving towards predictive analytics. Predictive Analytics (PA) uses manipulations to search through huge amounts of information, and analyse it to predict outcomes for users. It can include data from past results as well as latest advancements. It is used for predictions and also gives that we have never thought of. In medicine, predictions can range from responses to smart health care.

II. CONCEPT OF DATA MINING

Data mining is the process of extracting knowledge from a huge amount of data available with us which also helps us gain monetary benefits and save time doing everything manually. It allows us to study every aspect of data, process it with techniques available and finally result out with most appropriate data items. It is the process of pattern matching among various fields in huge databases. Data mining can answer questions that cannot be proven by other techniques
The key properties of data mining are:

- Pattern Discovery
- Probable Outcomes
- Useful Results

1) *Pattern Discovery*

Mining of data can be done by making new models. It uses a set of procedure to handle information. The concept of pattern discovery is the prototype of this model analysis. They can be used to mine the data on which they are built, but most types of models prefer new data.

2) *Probable Outcomes*

Whatever data is mined, will surely give some results. These results may not be precise and they showcase all the possibilities that can occur with the data items.

3) *Useful Results*

From whatever is obtained, we need to segregate the most useful results for the solution. This technique provided the same by applying the algorithm on data items.

III. LITERATURE REVIEW

Divya Jain *et.al* [1] presents a review of the implementation of Apriori Algorithm on datasets using machine learning tool Weka. Ruijuan Hu [2] states the details of the idea on two-step frequent data items using Apriori algorithms and Association Rules. This mentions a new improvised method called Improved Apriori Algorithm to eliminate cons of Apriori algorithm. Gitanjali J, *et.al.*[3] proposed study of huge datasets from various angles and obtaining gist of useful information. These methods are useful in detecting diseases and providing proper remedy for the same. Krishnaiah*et.al.* [4] aims to evaluate various methods of data mining in applications to develop precise decisions and also provides a detailed discussion of medical data mining techniques can improve various angles of clinical predictions. Dan A. Simovici [5] proposed that association rules represent knowledge in data sets as results and are directly related to calculation of frequent item sets. Mohammed Abdul Khaleel [6] states data mining as a concept that studies large amount of data and extracts patterns that can be converted to useful knowledge.

In this paper, we set out to identify efficient algorithm for mining results. By using all these predictive analytics and data mining techniques, we can create versatile applications for medicine sector so as to fulfil:

1. This tells how Apriori algorithm is used to find frequent data items and compares them with the existing algorithms.
2. Data mining techniques can be applied on medical data which has abundant scope to improve health solutions.
3. Electronic health records and other historical medical data can prove miracles if used for a right purpose.
4. Huge amounts of complex data generated by health care sector includes details about diseases, patients, diagnosis methods, electronic patient’s details hospitals resources.

IV. PROPOSED SYSTEM

Sometimes the situation occurs when you need the doctors help immediately, but they are not available due to some reason.This system allows the users to get analysis on the symptoms they give for predicting the disease they are suffering from.User will be asked to enter the symptoms, then system will processes those symptoms for various illness or disease that user could be aliked with. In this system we use some techniques of data mining to guess the most accurate diseases or illness that could be related with patient’s symptoms. If the system is unable to provide solutions, it informs the user about probable disease they have. If the user symptoms does not exactly match with any disease in our database, shows the diseases user could probably have judging the input symptoms.

This system tends to replace the existing system for going to the doctor for getting diagnosis on illness you are suffering from to an smart solution where you get instant diagnosis on entering symptoms in the system. The main features of this system will be giving instant diagnosis on the user entered symptoms and getting tips for remaining fit. In our proposed system, we use data mining method where user entered symptoms are cross checked in the database and frequent item sets are mined out of the existing datasets. The proposed system is an efficient algorithm implemented for the diagnosis of diseases. Fig. 1 illustrates the entire working of the system.

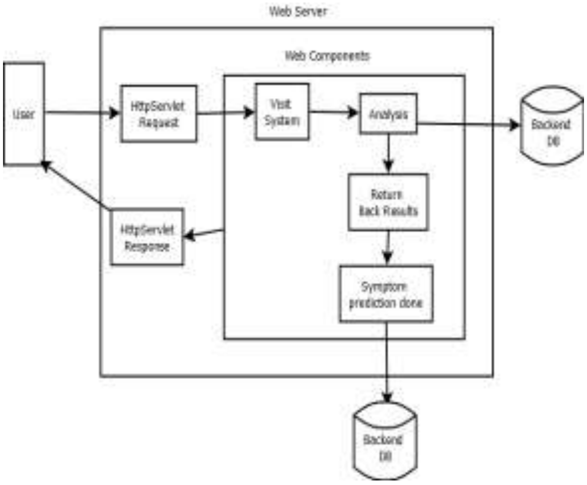


Fig. 1Block diagram of system

1. *APRIORI ALGORITHM*

Apriori is the algorithm for getting associated results from available data items. Apriori is used to work on innumerable data items. In association rule mining, with a set of items, the algorithm tries to find subsets which are common to a minimum set of items. Ituses a process whereby frequent items are used a single at an instant and group by the frequent items against the information in the database. The algorithm stops when no further successful results are found. Thepseudo code for the algorithm used in this paper for mining out results is presented in Algorithm1. We identify the support and confidence for finding frequent item sets where they are defined as

a) Support: The percentage of useful data sets for which the pattern is true.

b) Confidence: The measure of truth or trust associated with each pattern.

For example, the patient having headache as a symptom can have fever at the same time. This is acquired from the association rule below.

Support (Headache -> Fever)

$$= \frac{\text{No. of transactions containing both headache and fever}}{\text{No. of total transactions}}$$

(1)Confidence (Headache -> Fever)

$$= \frac{\text{No. of transactions both Headache and fever}}{\text{No. of transactions containing (Headache)}} \qquad (2)$$

The algorithm wants to find the rules which satisfy both a minimum support and confidence criteria. Rules originating from the same item set have identical support but can have different confidence. With a set of data items, this result finds all rules having minimum confidence level.

Algorithm 1 Algorithm for finding associated diseases

```
// Mn: Item set of size N
// Fn : frequent item set of size N

Step 1: F1 = {frequent items};
Step 2: for (N = 1; N<=Fn; N++) {
Step 3: Mn+1 = Medical symptoms derived from Fn after entering first symptom
Step 4: each t transaction in the database do {
Step 5: Increment count of all the symptoms in Mn+1
Step 6: Fn+1= min_support medical data in Mn+1  }
Step 7: end}
Step 8: return union NFn;
```

2. *WORKING OF APRIORI:*
   1. *Frequent Itemset Search:*
      - Obtain item occurrence:
        - Items that occur more than one times in the entire dataset.
      - Get frequent item sets:
        - Generate items that occur frequently.
   2. *Obtain rules that have greater confidence*
      - Rules which satisfy minimum confidence are listed.

The working can be explained with a flow diagram as shown in Fig. 2

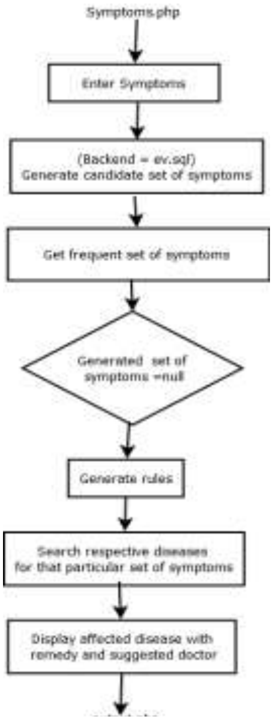

Symptoms.php
Enter Symptoms
(Backend = ev.sql)
Generate candidate set of symptoms
Get frequent set of symptoms
Generated set of symptoms =null
Generate rules
Search respective diseases for that particular set of symptoms
Display affected disease with remedy and suggested doctor

Fig. 2 Workflow of Apriori Algorithm

V. PERFORMANCE ANALYSIS

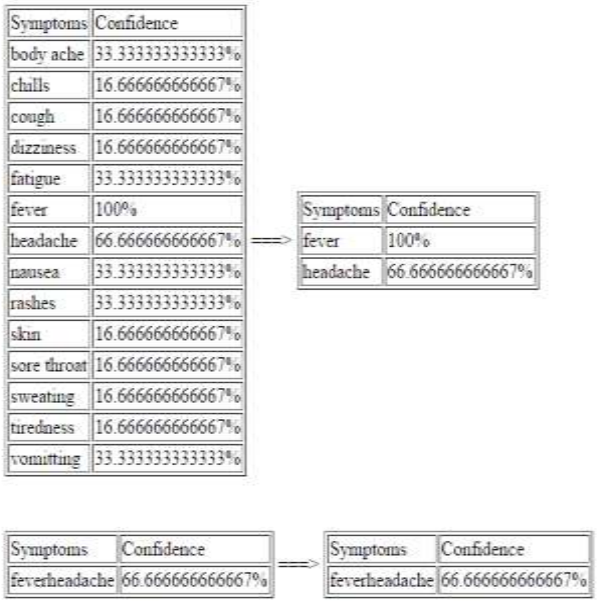
The system uses a mining algorithm which calculates the number of items used and the probability that they are occurring in the data sets mined. It is done using support and confidence of the used data sets. The following parameters are considered:

1. Support: The number of times an item occurs in the input data.
2. Confidence: The probability of a data item occurring in the data set.

Therefore, all the symptoms with that above it are carried forward to find frequent item sets in matching combinations. Systems which check symptoms online to predict diseases ask a variable set of random questions and then give innumerable results of the prediction. This can be a bit confusing for users who are not totally aware of the disease terminologies. Fig 3 illustrates the finding of frequent item sets.

Data items in the sample:

```
$dataset=array(array('headache','fever','fatigue','skin','rashes'
,'vomitting'),array('fever','chills','dizziness','headache','sweati
ng',' tiredness','body ache'),
array('fever','headache','fatigue','nausea','sorethroat'),
array('fever','headache','nausea'),
array('vomitting','fever','body ache'),
array('fever','cough','rashes')
);
```



Symptoms	Confidence
body ache	33.333333333333%
chills	16.666666666667%
cough	16.666666666667%
dizziness	16.666666666667%
fatigue	33.333333333333%
fever	100%
headache	66.666666666667%
nausea	33.333333333333%
rashes	33.333333333333%
skin	16.666666666667%
sore throat	16.666666666667%
sweating	16.666666666667%
tiredness	16.666666666667%
vomitting	33.333333333333%

==>

Symptoms	Confidence
fever	100%
headache	66.666666666667%

Symptoms	Confidence
feverheadache	66.666666666667%

==>

Symptoms	Confidence
feverheadache	66.666666666667%

Fig. 3 Analysis of symptom combinations

Table 1: Parameters in existing and current system

Parameters	Symptoms	Diseases
Existing System	10	15
Current System	25	7

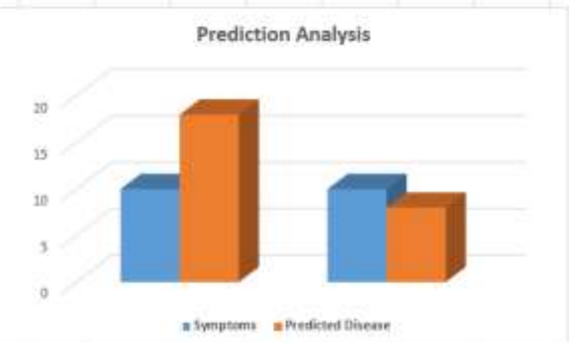


Fig. 4 Analysis of diseases with respect to symptoms

VI. CONCLUSION AND FUTURE WORK

In the proposed system, hidden knowledge will be extracted from the historical data by preparing datasets by applying Apriori algorithm. Predicting smart health can be done only is system responds that way. These datasets will be compared with the incoming queries and the final report will be generated using Association Rule Mining. Since this proposed methodology will work on real historical data, it will provide accurate and efficient results, which will help patients get diagnosis instantly. This system will also guide the users of how to remain healthy and fit using tips provided here.

The further enhancements that can be done would be integrating this web application in an Android app. This will be available to users on mobile basis and its use can be further increased. Also feature like getting the doctor online on chat so that patients can directly talk to the concerned doctors. The modules doing cancer analysis can be integrated to find how close the person associated with cancer is. This will make this web application predictable in true sense.

REFERENCES

[1] D. Jain, S. Gautam, “Implementation of Apriori Algorithm in Health Care Sector: A Survey,” Univ. ITM, Gurgaon, Nov. 4, 2013, Vol. 2.

[2] R. Hu, “Medical Data Mining Based on Association Rules,” University PLA, Luoyang, China, Nov. 2010, Vol. 3.

[3] Gitanjali J., C.Ranichandra, M. Pounambal, “APROIRI algorithm based medical data mining for frequent disease identification,” Univ. VIT, Vellore, 4, April 4, 2014, Vol 2.

[4] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, “A study on Clinical Predictions Using Data Mining Techniques”, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831,Vol. 3, Issue 1, Mar 2013, 239-248

[5] M. A. Khaleel, S. K. Pradhan, G. N. Dash, “Finding Locally Frequent Diseases Using Modified Apriori Algorithm,” International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) Vol. 2, Issue 10, October 2013.

[6] J. Han and M. Kamber, “Data Mining, Concepts And Techniques”.

[7] M. Tiwari, M. B. Jha, O. Yadav, “Performance Analysis of Data Mining Algorithms,” IOSR Journal of Computer Engineering (IOSRJCE), ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 3, PP 32-41, Sep-Oct 2012.

[8] Concepts Of Data Mining and Predictive Analytics [Online]. Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/ datamining.htm http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-6.html

[9] Dharmender Kumar,Suman, “Performance Analysis Of Various DataMining Algorithms: A Review”, International Journal of ComputerApplications (0975 – 8887) Vol.32– No.6, Oct-2011.