Prediction of Favourable Rules to Identify Suspected Patients of HIV Using Integration of Expert System and Data Mining

Authors
Amol Joglekar¹, Dr. G. Prasanna Lakshmi², Maunash Jani³
¹Research Scholar, Pacific University, Udaipur
²Guide and Woman Scientist.
³Member

ABSTRACT
HIV is a disease which is hard to analyze and predict and therefore complex operations are needed to diagnose to it. Many times the detections or observations done by doctors differ in many ways. In order to automate the system and improve accuracy of the same this research paper proposes an idea of integrating expert system with data mining. Basically expert system is a technology with artificial intelligence helping us to solve problems where human expertises are needed. Data mining tools helps us to choose the correct pattern from the huge medical database.

Keywords: Expert System, Data Mining, HIV, Prediction.

INTRODUCTION
Data Mining and Expert Systems
In medical field much data needs to be recorded like symptoms at each stage, background history of patient, family history, blood reports and so on. Using data mining which is a collection of techniques that provide necessary actions to retrieve and gather knowledge from exhaustive collection of data and facts. The concepts based on data mining are emphasis on discovering knowledge, predicting trends and removing unwanted data. With the help of knowledge discovery we can automatically find the pattern so that decision can be taken.

Decision taken by machines should be very specialize as day by day there is an increase in number of diseases. The process of identifying disease and analyzing starts when patient comes to a doctor with some symptoms. With the help of questions doctors try to identify the symptoms. In case of HIV there could be the possibility of some personal questions which are very sensitive and hence patient may lie or may not give correct answer. So by introducing an Expert system we can minimize this problem. An Expert System is a computer operated system which helps to take decision like a human being. They are designed to solve complex problems with the help of if-then rules. This type of system includes DBMS and Artificial Intelligence. Medical diagnosis is used to store and analyze patient’s information. It also helps doctors to take proper decisions based on symptoms.

In healthcare industry along with data mining expert systems plays a vital role. It makes the organization more accessible providing easiest way to reach more number of people. As medical knowledge is more complex and it is very difficult to tackle. Therefore there is a need of tool which can store information and its reason. This system can be used to generate a warning for patients so that appropriate action can be taken.

HIV/AIDS
HIV stands for human immunodeficiency virus. HIV is a virus spread through body fluids that affects specific cells of the immune system, called CD4 cells, or T cells. Over time, HIV can destroy so many of these cells that the body can’t fight off infections and disease.

HIV is a virus which attacks the immune system,
and weakens your ability to fight infections and disease. It is most commonly caused due to unsafe sex. It can also be passed on by sharing infected needles and other injecting equipment, and from an HIV-positive mother to the child during pregnancy, birth and breastfeeding. The virus attacks the immune system, and weakens your ability to fight infections and disease.

LITERATURE REVIEW
The proposed research paper is about predicting favorable rules for identifying HIV disease. There are many diseases which have symptoms and they need to be recorded based on their attribute or values. There may be different stages for various diseases. Therefore we need to study all different types of disease, how the data is collected and in what format so that attributes and parameter for symptoms can be taken into consideration. Data can be extracted and rules can be formed based on symptoms.

K.R. Lakshmi, M Veera Krishna and S. Prem Kumar [2] did checking the performance of data mining algorithms with respect to accuracy by computing time and cross validation error. They considered ten algorithms and tested with a huge medical data set of 2268 records related to heart disease. Authors claimed that PLS-DA classifier is a best suitable for Heart based disease. The testing was done using Tangara tool.

Chaitrali S Dangare, Sulbha A Apte [3] has evaluated prediction system for Heart Disease using increasing number of input attributes. The data mining classification techniques like Decision Tree, Naïve Bayes and Neural Networks were tested for Heart Disease. The performance of these techniques were tested and compared. Authors claimed that Neural Network gives highest accuracy.

Karthik S, Priyadarshani A, Anuradha J and Tripathi B. K [4] were applied a soft computing technique for intelligent diagnosis of liver disease. They have implemented classification in three phases. Artificial Neural Network (ANN) was used in first phase of implementation. In the second phase they used classification rules with rough set rules using Learn By Example (LEN) algorithm. The last phase was tested with fuzzy rules for identifying different liver types.

Chandraveer S. Deora, Sana Arora, Zaid Makani [5] discovered that when data mining algorithm used for digging out knowledge we may encounter many different rules. Association rule may not be useful to user. Therefore again we need to filter some rules so that unwanted rules based on situations can be removed and only correct few rules can be shown. Authors had used association rule mining to find out interesting measures i.e. Irule. They started with supper and confidence but landed with one question how to choose the best one from the given list. They did survey on different interesting measures and proved that Irule gives the best probable rules.

A. B. Rajesh Kumar, G. V. Ramesh Babu, C. Phani Ramesh, P. Madhura, and M. Padmavatham [6] proposed a new idea of using centralized database for accessing patient’s information in a secured way. There are many diseases like HIV where data is more sensitive and hence it should be kept secret. They found a decision tree can be used to predict the disease. Expert system was proposed to take some decisions based on patient’s information which helps doctors to take decision and provide proper treatment and at the same time data is made secure using cloud computing.

K. Haafeeza, R. Mohanraj [7] proposed a hybrid technique for diagnosing multidisease. The authors identified a gap between disease prediction, treatment and analysis. Their main aim was to reduce the number of deaths caused by various disease. They developed a layered approach for taking decisions and improve medical results. They used Naïve Bayes, Decision Tree, Neural Network, Association rule and Regression techniques. Integrating all these techniques they found out the risk of patients down with the disease.

Wilbert Sibanda, Philip Pretorius [8] did a review on one of the data mining algorithm i.e. Neural Network which was proposed by many authors for
identifying HIV disease. They mentioned that more than 30 papers been published between 1997-2012. Each paper was having an idea of using Neural Network and training datasets for identifying HIV. They claimed that it is a powerful and efficient tool to predict HIV disease.

RESEARCH METHODOLOGIES
A. Proposed architecture using expert system

![Figure 1: Architecture Flow of Proposed System.]

Here in the proposed architecture we have a database of symptoms which may come at different stages. The system needs to extract knowledge from patients regarding their history, symptoms and other details like age, habits etc. This process is known as Knowledge Extraction using Expert System. The process is carried out with the help of questionnaire in which we can use variables and score of the respondent as shown in the table.

**Table 1: Table Format for Expert System Consisting of Likert Scale and Value.**

<table>
<thead>
<tr>
<th>Likert Scale</th>
<th>Very high</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Very Low</th>
<th>Not available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When patient visits the clinic they need to tell what kind of symptoms they observed so that based on the data we can choose rules and predict the probable disease. Depending upon symptoms many rules can be selected but here we may need some system which will filter some rules which may not require and hence it will be easy for prediction.

B. Technical Model

**B. Technical Model**

In order to implement an automated system model we need a data mining algorithm implementation. Here we propose an association rules which are based on if-then pattern. Once the implementation of algorithm is done we create patterns for each symptom. Pattern formation is the process of combining symptoms from various rules and showing possibility of diseases in terms of percentage. The expert system helps us to represent knowledge so that it can easily infer the decision taken by the system.

C. Formulation of Questionnaire

![Figure 3: Process of Evaluating Questions.]

Doctors are supposed to create questions for a disease and they will be recorded in a database. Based on query fired patients has to respond in the
form of scalar representation. The values of those responses will be recorded in a database. Depending upon the stages available the same set of questions can be asked or there will be new formulation of questions. The idea of using scalar representation is patients need not have to type they just have to choose the scale with the help of mouse which will be an easy process.

D. Detecting a Disease

![Diagram of Decision Making Process of Proposed System]

There will be set of questions generated by the system based on each symptom. We record the response of each answer and based on calculation we display the list of disease. The percentage calculation is based on formula. If the percentage is high for a particular disease then we can start treatment at earliest.

E. Proposed Algorithm

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys bread, he is 80% likely to also purchase butter."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are useful for analyzing and predicting behavior and pattern. They play an important part in data analysis, product clustering, and design.

![Diagram of Algorithm Flow of Proposed System]

The proposed algorithm is used for the computation of association rules to diagnose the diseases in a patient.

This algorithm uses parameters to measure the severity level mentioned below:
Very high = 10, High = 8, Medium = 6, Low = 4, Very low = 2

**Step 1:** Display the list of all the symptoms from the database for stage one.

**Step 2:** Select the particular symptom to be diagnosed.

**Step 3:** Select the level of severity for each selected symptoms.

**Step 4:** For each selected symptoms, search the database for all the diseases that has the symptom and then select the weight of the symptom and multiply it by the appropriate scaling variable.
Step 5: Actual weight of symptoms = Weight of symptom + Sum of Max of all Scaling values of levels of security.
FOR EACH Selected symptom in All Disease for stage one
IF Symptom is present THEN
Suspected Result = Sum of Weight of all symptoms * Sum of scaling weights of severity levels
Suspected Result % = (Suspected Result / Actual weight of symptoms) * 100
END IF
END FOR EACH

Step 6: Select the disease as per the symptoms with correlation of specific information of the symptoms and the diseases. This is the diagnosed disease.

Step 7: Display disease; after the diagnosis of the disease, the diagnosed disease is displayed in the application.

Dataset for HIV Disease
Predictable attributes based on diagnosis and key attributes.

1. Diagnosis

Table 2: Key Attribute

<table>
<thead>
<tr>
<th>No HIV</th>
<th>HIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>Value</td>
</tr>
<tr>
<td>0</td>
<td>&lt;50%</td>
</tr>
<tr>
<td>1</td>
<td>&gt;50%</td>
</tr>
</tbody>
</table>

2. Patient ID – Patient’s identification number

Table 3: Key Attributes of HIV

<table>
<thead>
<tr>
<th>No</th>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>1: Male; 0: Female</td>
</tr>
<tr>
<td>2</td>
<td>Depression</td>
<td>1: Major Depression 2: Atypical Depression 3: Dysthymia 4: seasonal affective disorder</td>
</tr>
<tr>
<td>3</td>
<td>Diarrhea</td>
<td>1: Yes; 2: No</td>
</tr>
<tr>
<td>4</td>
<td>Burning Tingling of the feet &amp; hands</td>
<td>1: Yes; 2: No</td>
</tr>
<tr>
<td>5</td>
<td>Weight Loss</td>
<td>1: Yes; 2: No</td>
</tr>
<tr>
<td>6</td>
<td>Vomiting</td>
<td>0: No, 1: Normal, 2: High</td>
</tr>
<tr>
<td>7</td>
<td>Fatigue</td>
<td>0: No, 1: Normal, 2: High</td>
</tr>
<tr>
<td>8</td>
<td>Age</td>
<td>Value</td>
</tr>
</tbody>
</table>

CONCLUSION
The research paper gives a new technique which works with the combination of data mining and Expert System in medical data to predict the disease with percentage.
The justification of developing expert system is discussed in the paper. Expert Systems play a dominant role in the realm of overwhelming technology. Proposed diagnosis system can play a vital role for the users of the system. The system provides the users to determine a probable diagnosis of HIV very quickly with the aid of a knowledge based expert system.

RESULTS AND FUTURE SCOPE
Figure 7: Graphical Representation of Disease

RESULTS
The focus on this research paper was to give a glimpse on GUI model of the system. The system is not tested with real life data obtained from hospital. The research work will be tested with a real time data so that 100% accuracy and prediction can be achieved.

REFERENCES
9. www.avert.com
10. www.health.com
11. www.aidsprogramme.ukzn.ac.za
12. Han J Kamber “Data Mining Concepts and Techniques”