International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

Prediction percentage of severity in HIV Patients using data mining algorithm

Amol Joglekar, Research Scholar Pacific University, Udaipur 1; Dr. G. Prasanna Lakshmi, Guide , Mumbai 2; Maunash Jani, Member , Mumbai 3

Abstract :Data mining consists of analyzing large amount of data from various sources and fetches the best knowledge out of it. Medical science has huge information and if it is processed in an appropriate manner we can able to predict dangerous disease like HIV in no time so that early treatment can be started. We have proposed an innovative data mining algorithm which helps doctors or trainer doctors to identify the severity level of HIV/AIDS in patients. With the help of proposed algorithm one can study different symptoms associated with the disease and the severity of symptom level. The proposed algorithm has been detailed in paper.

Keywords: HIV/AIDS, data mining, item sets and attributes.

Introduction

The origin of HIV/AIDS was come into picture in early 1980's and from that time onwards research is ongoing in variety of aspects. HIV is lentivirus i.e. it attacks the immune system of body slowly. It attacks on human body and destroys each and every body part resulting into death. HIV has been found on many animals like cat, sheep, horses etc. In India many people die because of HIV/AIDS and it has become the third country having HIV infected people.[1,2] Like other diseases HIV shows symptoms at different stages and therefore we can predict the percentage of infection that patient may have. The task of identifying such type of disease is complicated and needs to be handled accurately. If the treatment is given at early stage patient can live a normal life up to many years without any problem. With the help of data mining and its impact on medical field we can detect this at early stage so that proper treatment can be started. Health informatics is a rapidly developing field which is concern with developing Computer Science and Information Technology to medical field and organization. It is computerization of health information to support and optimize administration of health services, clinical care research.

Data mining is a model which analyzes data based on variety of parameter like history and finds out knowledge or actual facts out of it. It has a variety of technique to discover data.[4,5] Therefore an automated predictive system is needed to predict the percentage of HIV disease.

Background and Related Work

The proposed thesis paper is about developing an algorithm using data mining predictive tools and techniques. There are many diseases like cancer, heart attack etc. are having symptoms. Due to this reason there is a need of studying current techniques in order to propose futuristic scope. Medical science has great potential for finding out hidden information or domain. The data should be extracted in a particular format so that it can be used to dig out the knowledge.

Ronaldo Cristiano Prati, Maria Carloin Monard and Andre C.P.L.F.de Carvalho [6] presented different way to extract knowledge rules from database which is HIV infected. Their main feature was to incorporate exceptions into the representations used by system or machine understandable format. That method had two steps: training of common sense rules and checking exceptions. In order to implement there were needed real world dataset where a viral protease cleaves HIV viral poly protein amino acid remains. A method was to find general rules which were suitable for analysis. It allows more easy way to help people and understand the process

Sengul Dogan and Ibrah Turkoglu [7] suggested a new approach to find association rules which is an effective method for discovering Hyperlipidemia. They presented a model inherited from biochemistry blood parameter which is very helpful to make everything easier experts in the diagnosis of Hyperlipidemia. The basic feature of lipid parameter was LDL, HDL and VLDV were taken into consideration and based on these results were calculated. They observed the results were matching perfectly with physician's decisions.

Jesmin Nahar, Tasadduq Imam, Kevin S Tickle and Yi-Ping Phoebe Chen [8] presented a rule extraction model on heart disease using Apriori and Predictive Apriori techniques. They had considered some attributes like age, sex, chest pain, old peak, ca, Thal etc. Based on these attributes they had developed rules for male and female which tells the healthy patterns observed in them. They had demonstrated rule mining to determine interesting knowledge which would analyze the factors causing heart disease.

Sean N. Ghazavi, Thunshun W. Liao [9] presented a study of medical data using mining techniques like fuzzy modeling

ISSN:2319-7900

method for digging out features. The emphasis was on why selection of feature was more important in medical science which reduces time to identify the disease and provides accurate results with good accuracy. They had considered two types of disease cancer and diabetes. They had provided a feature based solution which was not available with ready-made data mining software.

I.S.Jenzi, P. Priyanka, Dr. P. Alli [10] proposed a new system based on Data Mining for predicting Heart Disease. They had collected patterns from medical data for finding heart disease. They did user friendly application for predicting the disease. They found that decision tree was easy to interpret and had a good accuracy. They found two difficulties like large dataset and user interface should be efficient enough to support data in different format instead of ARTF which can be taken care of in next research.

Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddique [11] focuses on predictive analysis of diabetic treatment using regression based data mining technique. Authors had collected data from Saudi Arabia consisting of variables like drug, weight, smoke intake and age. They designed a GUI based application which was easy to operate and helpful for providing training to doctors is and medical assistants.

Dr. K. Rameshkumar [12] developed a model using ARM (Association Rule Mining) to extract valuable information from database. Author has proposed a new algorithm which would take care of missing values for detecting HIV AIDS. With the help of this proposed algorithm author could able to extract information about CD4 cell counts, RNA levels and treatment given for various patient. The model is lacking of handling data with a very good accuracy.

Research Methodology

A. System Design

The proposed system architecture will have three layers with technology and operating system context.

- Application layer
- Business logic
- Storage unit.

The application layer is a user interface module which is used for communicating and accepting details of patients in a friendly manner. The business logic as a name indicates contains different procedures, algorithms which are used for writing logic for the application. It is a part of software which shows how the real world data can be shown and used for analysis. The storage layer or commonly known as database layer is a house for information storage. Whatever user will ask with the help of GUI interface can be extracted with the help of queries and result can be displayed to user in understandable format.

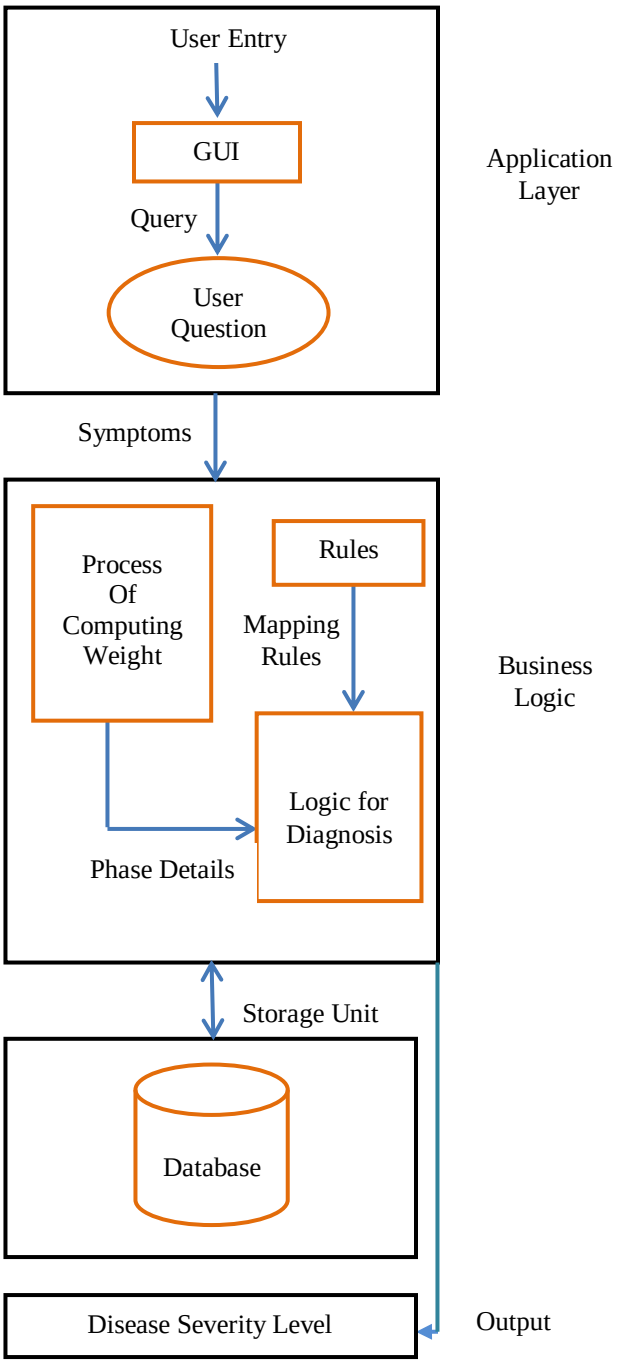

Figure 1: Component diagram of a system.

ISSN:2319-7900

B. Methodology (Process Flow) used in Proposed Algorithm

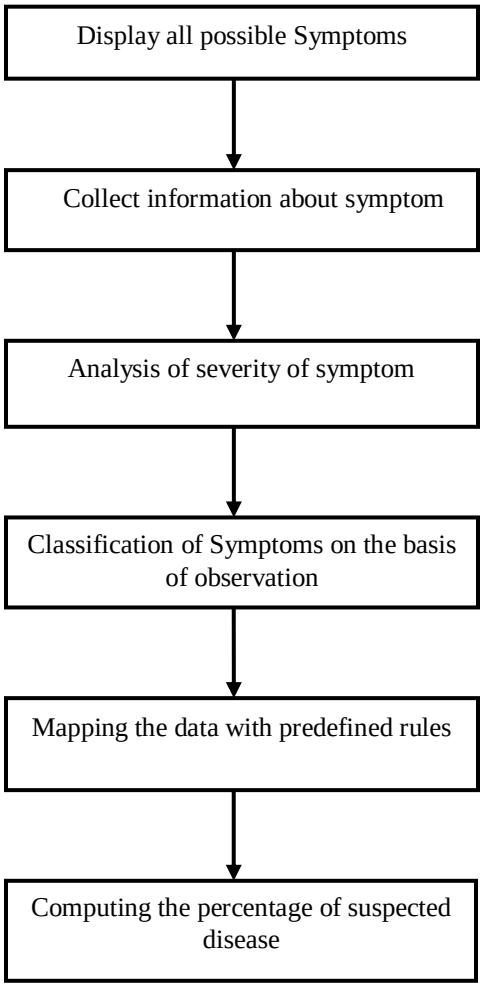

Figure 2: Process flow of methodology.

The process flow has 6 functional units consisting of analysis and computing the occurrence of HIV infection.

1) The process boundaries are defined as:

a. Display of all possible symptoms – represented as starting point.
b. Display the result – represented as ending point.

2) Each step in the flow after the start process is defined as below:

a. All the information about the symptoms is gathered which is taken from the set of inputs.
b. The above information is then classified on the basis of observation.
c. The analysis of severity of the symptoms collected is done by applied levels of severity as input.
d. All the above data is then mapped with predefined rules.
e. A percentage of suspected disease is computed from the mapped data.
f. The final result is then displayed.

C. Experimental Data

The proposed algorithm uses four data sets called – Symptoms, Disease, Severity Level and Common Symptoms. There are numerical values allocated to each symptom and severity. The aim is to classify the diseases and predict the presence of HIV infection into the patient by analyzing and computing the small and large item sets on the basis of proposed algorithm. The small item dataset has 17 numerical values containing the symptom. The large item dataset has 8 attributes containing the combination of symptoms commonly found in patients. The severity level dataset has 5 numerical values containing the severity level for the symptoms present in the patient during the analysis process.

Table 1 : Small Item Set - Symptoms, L:

Symptom	Weight	Stage
Vomiting	3	S1,S3
Fever	2	S1
Joint Pain	2	S1
Muscle Pain	2	S1
Headache	3	S1
Swollen Gland	4	S1
Diarrhea	3	S1
Tiredness	2	S1,S2
Swollen Lymph Nodes	4	S2
Weight Loss	4	S2
Rashes	3	S2,S3
Stiffness	2	S2
Sores	2	S2
Depression	2	S2
Mouth Infection	3	S2
Vision Loss	2	S3
Memory Loss	3	S3

ISSN:2319-7900

Table 2: Attributes - Disease, L:

Disease	Stage
Migraine	S1
Meningitis	S1
Influenza	S1
Infectious Diarrhea	S1,S2,S3
Alcohol withdrawal Syndrome	S1
Dehydration	S1
Common Cold	S1
Pneumonia	S1,S2,S3
Typhoid	S1,S2
Acute Bronchitis	S1
Chicken Pox	S1,S2
Dengue Fever	S2
Malaria	S2
Tuberculosis	S2,S3
Celiac Disease	S2,S3
Salicylate Poisoning	S3

Proposed Algorithm

The proposed approach is helpful in identifying the presence of HIV infection in a patient. As a result, medical conclusions, treatment procedures and decisions can be made by practitioners accurately.

Algorithm

1. Input L for S
2. Select S_v for L
3. Scan all D ϵ L
4. Choose W for L and accumulate it with S_v , W ϵ L * S_v
5. for S=1 to n
6. do
7. for all D ϵ S_n
8. do
9. for all L ϵ D
10. do
11. t1= $\sum_{i=1}^{k} W * Sv$
12. for each subset l_i ϵ L
13. t2= $Max(\sum_{i=1}^{m} li) + Lw$
14. end for
15. for each S
16. do
17. Suspection = (t1 / t2) * 100
18. end for
19. end for
20. end for

S: Stage
D: Disease
L: Symptom
W: Weight of Symptom
S_v: Scaling Variable
t1,t2: Temporary Variable
l_i: Scaling value of Symptom

Conclusion and Future Scope

The above thesis paper has proposed a new algorithm for identifying HIV infections in patients. The future work will be an analysis and testing of an algorithm with the help of test data and developing a predictive model which can predict the disease in no time. With the help of algorithm rules can be constructed. Also to test the accuracy of proposed algorithm on different platform will be a challenge.

Acknowledgments

The authors are thankful to IJACT Journal for the support to develop this document.

References

[1] www.avert.org
[2] www.health.com
[3] www.aidsprogramme.ukzn.ac.za
[4] Han J Kamber, "Data Mining Concepts and Techniques", Morgan Kaufman Publishers 2006.
[5] "Data Mining : Introductory and Advanced Topics" Margaret H. Dunham
[6] Ronaldo Cristiano Prati, Maria Carolina Monard and Andre C.P.L.F de Carvalho. "Looking for exceptions on knowledge rules induced from HIV cleavage data set." International Journal Genetics and Molecular Biology, Vol.27 , Issue.4, pp.637-643,2004.
[7] Sengul Dogan, Ibrahim Turkoglu. "Diagnosing hyper lipidemia using association rules." Mathematical and Computational Applications, Vol. 13, No.3 , pp.193-202, 2008.
[8] Jesmin Nahar, Tasadduq Imam , Kevin S. Tickle , Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in male and females" Expert Systems and Applications 40,pp. 1086-1093, 2013
[9] Sean N. Ghazavi, Thunshun W. Liao. "Medical data mining by fuzzy modeling with selected features"

ISSN:2319-7900

Artificial Intelligence in Medicine 43,pp. 195-206 , 2008.

[10] I.S.Jenzi , P.Priyanka, Dr. P. Alli "A reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction" International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 3, pp. 20-24, March 2013

[11] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddique "Application of data mining: Diabetes helth care in young and old patients" Journal of King Saud University- Computer and Information Sciences,25 pp. 127-136, 2013.

[12] Dr. K Rameshkumar "Association Rules Mining from HIV/AIDS patient's case history database with missing values" International Journal on Data Mining and Intelligent Information Technology Applications" Vol.2 , No. 1,pp. 18-24, March 2012.

[13] K. Shrinivas, B. Kavitha Rani, Dr. A Govardhan "Application of Data Mining Techniques in Helthcare and Prediction of Heart Attacks" IJCSE vol. 02 ,2010, 250-255.

[14] Ahmad LG, Eshlagby A T, Poorebrabimi A, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence" Health and Medical Informatics vol.04 issue 02 1000124. Open access journal 2013.

[15] Gitanjali J,C.Ranichandra, M.Pounambal "APRIORI algorithm based medical data mining for frequent disease identification" IPASJ International Journal of Information Technology(IIJIT) Vol.2, Issue 4, April 2014, ISSN 2321-5976.

[16] Priyanka Sharma, DBV Singh, Manoj Kumar Bandil, Nidhi Mishra "Decision Support System for Malaria and Dengue Disease Diagnosis(DSSMD) " International Journal of Information and Computation Technology ISSN0974-2239 ,Vol.3, Number 7(2013) pp.633-640.

[17] D. Senthil Kumar, G.Sathyadevi, S.Sivanesh "Decision Support System for Medical Diagnosis Using Data Mining" International Journal of Computer Science Issues (IJCSI), Vol.8, Issue 3, No.1 May 2011, ISSN: 1694-0814.

[18] Amol Joglekar, Dr. G. Prasanna Lakshmi "Detection of HIV and AIDS by Advance Algorithms" International Journal of Electronics Communication and Computer Engineering" (IJECCE) Vol.5, Issue 3, ISSN 2278-4209.

Biographies

1) **AMOL JOGLEKAR** Passed M.Sc (Computer Science) from Mumbai university and completed M.Phil (Computer Science) from Madurai Kamraj University Currently he is pursuing P.Hd from Paciffic University , Udaipur.

2) **DR. DR.G. PRASANNA LAKSHMI** She has completed M.Phil and Phd in Computer Science from GITAM university. She was acting as a In-charge Principal in Wilfred Institute of Technology , panvel. She has overall 15 years of experience.

3) **MR. MAUNASH JANI** working as a freelancer in Mumbai and his area of interest is programming languages and core topics of Computer Science.