Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Developing Knowledge Based System for Predicting HIV/AIDS Stages Using Data Mining Techniques

[1]Zegale Lake , [2] Dr. Shaik Saidhbi
[1]Lecturer, Department of Computer Science, Samara University
[2]Associate Professor, Department of Computer Science, Samara University

Abstract : The Human Immunodeficiency Virus (HIV) is an infectious disease that attacks the immune system cells. HIV disease is the most killer disease among humankind, especially in adult age level. Data mining is the process of extracting previous unknown patterns and knowledge to design predictive models from large patient data. The huge amounts of patient data generated for the prediction of HIV Stages are time-consuming and too complex to processed and analyzed by manual methods. Prediction of HIV/AIDS Stages is a major problem using traditional method faced in hospitals and other clinical centers. The researcher is initiated to develop Knowledge-Based System and identify HIV/AIDS patient's stages using different data mining technique based on the data which is collected from the hospitals. The researcher use Knowledge Discovery in Databases Data Mining methodology for this research study. To identify the stage of HIV patient the researcher used three data mining classification algorithms; J48
Decision tree as well as, PART and JRip rule induction. The PART classification algorithm is selected for use in the development of a knowledge base because it registered the best accuracy of 98.4885 %. The evaluation result of the proposed knowledge-based system achieves a promising result of 89.7959 % accuracy and 76% user acceptance testing. To accomplish this study, the researcher needs an accurate dataset.

Key Words: HIV/AIDS, Knowledge Based System, Diagnosis, Data Mining, Prediction

Introduction

HIV/AIDS stands for human immunodeficiency virus and Acquired Immunodeficiency Syndrome. HIV is a virus distributed through the body fluids that affect specific cells of the immune system, called CD4 cells, or T cells. On the other hand, HIV can damage so many of these cells that the body can't fight off infections and disease. It is sometimes caused due to unprotected sex. It can also be passed on by sharing infected needle material and other injecting equipment, and from mother to child during pregnancy, birth, and breastfeeding. [1]

In the medical field, much data needs to be recorded like symptoms at each stage, background history of the 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 an exhaustive collection of data and facts. [2] With the help of knowledge discovery, we can automatically find the pattern so that a decision can be taken.

Knowledge-Based System is a computer application that performs a task that would otherwise be performed by a human expert; such tasks include but are not limited to making a financial forecast, scheduling routes for delivery vehicles, diagnosing human illnesses, and several others. To design Knowledge-Based System, the domain of the knowledge field is required, so an individual needs to be able to study how the human expert makes decisions and translate the rules used into terms that the computer would understand. The modern solution to address the problem, without the involvement of a human legal expert, is to develop Knowledge-Based System [6].

Knowledge-Based System is one of the most important applications of artificial intelligence. It is a set of programs to manipulate knowledge to solve problems within a specialized domain that necessitates human expertise. Knowledge-Based Systems sometimes also called expert system. The most important components of the Knowledge-Based System are the user interface, data representation, inference, explanations and advantages of the Knowledge-Based System are reduced cost, fast response, increased reliability, intelligent database, reducing errors, multiple expertise, and reduced danger. [6]

The problem of effectively utilizing this huge amount of data is becoming a major problem for all health to make an effective decision. Additionally, inaccurate data leads to a wrong decision. So in order to make an effective and efficient decision it needs a new and more effective technique to extract the hidden knowledge that leads the user to the right decision. Human experts are unable to retain large amounts of data in memory so the purpose of this proposed research is to solve this type of problem.

Related works

According to [39] the study entitled **" Integration of Data Mining and Hybrid Expert System",** their study is focused on the fuzzy rule extraction technique that has achieved satisfactory values for clinical cases of classification of epileptic crises. Therefore, the system was able to translate the encoded knowledge among the connection weights of the NNES into rules. Besides, all examples applied demonstrated that the procedure for reducing the number of redundant rules produced after the learning and refinement stages of the ANN was appropriate. The main goals of this work are to apply some of these tools and techniques to a medical database of breast cancer, so that detection and prediction patterns are discovered, and use the database resulting from the mining process in a hybrid expert system that will help in medical diagnosis. Based on the simulations for clinical cases of classification of epileptic crises, whose results were among 77.8 up to 83% accuracy, Also, the researcher collects six hundred ninety-nine (699) data from the institution and the data have eleven (11) attributes.

According to [2] the study entitled "**An Expert System for HIV Screening Using Visual Prolog**". Their study standard reporting tool to generate a report of all the patients that have used the system and their results will be an improvement and ensure system optimality and efficiency. A user-friendly medical expert system for screening HIV was designed using Visual Prolog, to aid medical practitioners and health care workers in the process of screening individuals of HIV. This would in turn help in solving the challenges faced by people most especially in communities where there is a shortage or unavailability of medical personnel. It provides a very rapid method of prognosis with **much accuracy** and reduces the hours patients spend in hospitals and boring routine tasks associated with the existing method of HIV screening.

In addition to this, [41] the study entitled" **Prediction of Pediatrics HIV/AIDS Patient's Survival in Nigeria: A Data Mining Approach**" states that it identified survival variables for HIV/AIDS pediatric patients, developed predictive model for determining the survival of the patients who were receiving antiretroviral drug in the Southwestern Nigeria based on identified variables, compared and validate the developed model. In this research paper Predictive model was developed using supervised learning technique (The Multi-Layer Perception (MLP)) and the Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the models in which CD4 count, Viral Load, Opportunistic infections, and Nutritional status were used as the independent variables for the prediction. The result showed that The Multi-Layer Perception (MLP)) was suitable for carrying out the task of forecasting the survival of Pediatrics HIV/AIDS patients with an accuracy of 99.07%.

The researcher [43] have developed **Data Mining Based Hybrid Intelligent System for Medical Application**, the purpose of this research paper is a hybrid intelligent system that uses data mining technique for healthcare center especially for Tuberculosis (TB) as a tool for the knowledge acquisition process. For this study, a total of 6330 TB datasets were collected from two different places. 5125 datasets were collected from the Felege Hiwot hospital and 1205 datasets from Bahir Dar health Center. Each of those records consists of 20 different variables

(attributes). For validation purpose, the full dataset is split into 80% (5064) as training dataset and 20 %(1266) as testing dataset, the dataset contains the following attributes: age, weight, Hemoglobin, chronic cough, fever, chest pain, bloody sputum, headache, loss of appetite, night sweating, weight loss, exhaustion, HIV test, shortness of breath, sputum test, X-ray test, FNAC test, CSF test, lymph node swelling, and TB type. They use WEKA for model construction and evaluation, Java NetBeans for integrating data mining results with rule-based reasoning and Prolog for knowledge representation. The researcher uses four data mining algorithms such as J48, BFTree, JRIP, and PART. Finally, the aim of integrating data mining techniques with the hybrid intelligent systems is to reduce the difficulty caused by the 'knowledge acquisition bottleneck' and to obtain low-cost and high-quality knowledge base. The system is evaluated using different evaluation methods and the system has achieved an accuracy of 83.5% on system performance testing and 82.2% on user acceptance testing.

Method and Approaches

Knowledge Acquisition

Manual Knowledge Acquisition

The researcher used both interviews and document analysis to acquire knowledge. The researcher conducted the domain experts' interview with Doctors, Nurses, Laboratory Technicians and Pharmacists who work for HIV/AIDS diagnosis and Treatment and Research Center at different hospitals what the data have been collected.

Automated Knowledge Acquisition

The researcher used the KDD model to acquire knowledge from each hospital dataset using the WEKA data mining tool. Knowledge Discovery in Databases (KDD) denotes that the task of revealing significant connections and regularities in data mining based on the use of algorithms. The KDD process is an iterative fulfillment of the following steps:

I. Data selection and preprocessing, such as checking for errors, removing outliers, handling missing values, and transformation of formats.
II. Data transformations, for example, discretization of variables or production of derived variables
III. Selection of a data mining method and adjustment of its parameters.
IV. Data mining, i.e. application of the selected method.
V. Interpretation and evaluation of the results.

Knowledge Representation

Since the knowledge that the researcher acquired from Data mining classification technique is in the form of rules and the knowledge that the researcher acquired from document analysis and domain experts' interview about diagnosis and treatment of HIV/AIDS stages are full of decision trees and procedures which are easy to convert to rules, the researcher forced to use rule-based knowledge representation method which is the most predominant knowledge representation methods to develop the Knowledgebase.

System development methodology

The prototyping approach is followed to develop the knowledge-based system. Prototyping allows participating users and domain experts for evaluating systems performance and efficiency.

Implementation tool

In order to mine the hidden knowledge from the pre-processed dataset and compare the performance of classifiers, the researcher used WEKA 3.6.9 data mining tool. To represent rules in the knowledge base and construct the prototype of HIV/AIDS Advising Knowledge-Based System, the researcher used Java NetBeans IDE 7.2.1 with JDK 6. It was employed to integrate WEKA results with the Knowledge-based system and develop the GUI of the proposed system.

Evaluation methods

The researcher used Precision, Recall, F-measure and True Positive rate to evaluate the results and the accuracy of the data mining model. The researcher also evaluated the KBS using system performance testing by preparing test cases and users' acceptance testing questionnaire which helps the researcher to make sure that whether the potential users would like to use the proposed system frequently and whether the proposed systems meet user requirements.

Result and discussion

Variable used for the experiment

The goal of attribute selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit of reducing the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand [61].

To select the best attributes for data mining, the researcher uses information gain method which exist in WEKA data mining tool and the domain experts' advice and come ups with 9 attributes which are listed here below in Table 9:

Table 9: Attribute Description

No.	Attribute Name	Value Type	Description
1.	Sex	Nominal	Describes the sex of the patient
2.	Age	Numeric	Describes the patient Age
3.	Marital Status	Nominal	It describes the marital status of the patient
4.	Functional Status	Nominal	Describes the functional status of each patient
5.	Weight	Numeric	It describes the weight of the patient
6.	Symptoms	Nominal	It describes the sign and symptoms of each individual HIV/AIDS Patients.
7.	CD4	Numeric	Describes the CD4 count number of the patient
8.	Original Regimen	Nominal	Describes the type of regimen
9.	WHO Clinical Stage	Nominal	It describes the WHO clinical stage of the patient.

The researcher conducted a discussion with domain experts and reviewed kinds of literature to identify relevant attributes, which have contributing for identifying stage of HIV. The attributes of the above is selected from patient record, some of them are Sex, Age, Marital Status, Functional Status, Weight, Symptoms, CD4 count, Original Regimen and Who Clinical Stage are the most relevant attribute used for model building.

Sex: This attribute describes the gender of the patient and it is classified up on Male and Female.

Age: This attribute describes the age of the patient and the data type is numeric.

Marital Status: This attribute describes the status of the patient i.e. either married means it have wife or husband, widowed means at the last age or old and single means the first age from 1 - 15

Functional Status: This attribute describes the functional status of the patient. The functional status of the HIV patients was categorized into three grades by WHO, which was internationally validated. The grading was as follows-(**W**) Working: Able to perform usual work inside or outside home. (**A**) Ambulatory: Able to perform Activity of Daily Living -ADL, not able to work. (**B**) Bedridden: Not able to perform Activities of daily living.

Weight: This attribute describes the weight of each patient which is affected by HIV in KG.

Symptoms: This attribute describes sign and symptoms of the patient happen on his/her body

CD4 count: This attribute describes the CD4 count of the patient and the data type of this attribute is numeric.

Original Regimen: This attribute describes the type of drug the patient is taken when he/she is on treatment. The data type is nominal.

Who Clinical Stage: This attribute describes the final class for this study and the stages of the patient like Stage 1, Stage 2, Stage 3 and Stage 4.

```
Attribute Evaluator (supervised, Class (nominal): 9 WHO clinical stage):
        Information Gain Ranking Filter

Ranked attributes:
 1.82974    5 Symptom
 0.119509   7 CD4
 0.036803   8 Original Regimen
 0.034558   4 Functional Status
 0.009525   6 Weight in KG
 0.004228   2 Age
 0.002283   1 Sex
 0.000525   3 Maritul Status

Selected attributes: 5,7,8,4,6,2,1,3 : 8
```

Weka experiment result

Table 16: Objective evaluation results

Evaluations	Algorithms		
	J48	PART	JRip
Correctly Classified Instances	6112	6125	6091
Incorrectly Classified Instances	107	94	128
Accuracy (%)	98.2795	**98.4885**	97.9418
Time taken	0.08	1.03	2.13
Av.TP Rate	0.983	**0.985**	0.979
Av. FP Rate	0.01	0.007	0.008
Av. Precision	0.983	**0.985**	0.98
Av. Recall	0.983	**0.985**	0.979
Av. F-Measure	0.983	**0.985**	0.98
ROC Area	0.999	1	0.996

Table 13: Confusion Matrix for the PART classification algorithm

a	b	c	d	Classified as
884	0	0	0	a = Stage 4
0	2445	51	0	b = Stage 3
0	43	1390	0	b = Stage 2
0	0	0	1406	b = Stage 1

System performance result

Table 20: System performance measurement result

No	Performance Measurement	Result
1.	Accuracy	89.7959 %
2.	TP Rate	0.898
3.	FP Rate	0.054
4.	Precision	0.878
5.	Recall	0.898
6.	F-Measure	0.875
7.	ROC curve	0.981
8.	CCI	44
9.	ICI	5

Conclusion

This study showed that developing knowledge-based predictive models for HIV/AIDS Stages uses data mining techniques. The algorithms for these study are J48, PART and JRip algorithms were able to predict the stages of HIV/AIDS using the following relevant attributes which is collected from the hospitals like sex, age, marital status, functional status, weight, CD4, Original Regimen, symptoms and WHO clinical stage as inputs with an accuracy of 98.2795 %, 98.4885 %, and 97.9418 %respectively.

System performance evaluation and user acceptance test has been conducted to evaluate the designed system performance.

As a result, 89.7959 %of accuracies were scored for system performance and 76% were scored for the user acceptance test. So, the research concludes that integrating data mining rules, explicit and tacit knowledge is possible to develop knowledge-based system.

Reference

[1] Amol Joglekar,Dr. G. Prasanna Lakshmi, Maunash Jani, "Prediction of Favourable Rules to Identify Suspected Patients of HIV Using Integration of Expert System and Data Mining," International Journal of Mechanical Engineering andd Information Technology, vol. 04, no. 03, p. 1612, 2016.

[2] B.A. Abdulsalami., T.K. Olaniyi, R.A. Azeez & M.A. Ogunrinde, "An Expert System For HIV Screening Using Visual Prolog," African Journal of Computing & ICT, p. 121, Octobe 2015.

[3] Idowu Peter Adebayo*, Agbelusi Olutola, Aladekomo T. A., "The Prediction of Paediatric HIV/AIDS Patient Survival: A Data Mining Approach," Asian Journal of Computer and Information Systems, p. 87, June 2016.

[4] WHO, INTERIM WHO CLINICAL STAGING OF HIV/AIDS, AFRICAN REGION: Treat Million, 2005.

[5] WHO, "HIV/AIDS," 19 July 2018. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/hiv-aids. [Accessed 17 december 2018].
[6] Nohria, Rimpy, "Medical Expert System- A Comprehensive Review," International Journal of Computer Applications, vol. Volume 130, p. 44, November 2015.
[7] WHO, "World Health Organization," 19 July 2018. [Online]. Available: file:///E:/document/ARBAMINCH/2010-2011/Thesis%20proposal/raquel%20comm/HIV_AIDS.html. [Accessed 1 march 2019].
[8] Mrs. Bharati M. Ramageri, Lecturer, "DATA MINING TECHNIQUES AND APPLICATIONS," Indian Journal of Computer Science and Engineering, vol. 1, p. 304.
[9] Ravleen Singh, Dr. Tariq Hussain Sheikh, "An Overview of Data Mining Applications in Healthcare," International Journal of Advance Research in Computer Science and Management Studies, vol. 4, no. 2, p. 131, 2016.