A Bayes-Based Model for HIV Prediction Extinction

Oguns Yetunde J.
Department of Computer Science,
The Polytechnic Ibadan, Nigeria

Ogundele Tunde J., Thompson Aderonke F.
Department of Computer Science,
Federal University of Technology Akure, Nigeria.
ogundele.tj@gmail.com

Adu Ayooluwa S.
Equitable Health Access Initiative
Alagbaka, Akure.

Abstract

The human immunodeficiency virus (HIV) is one of the most serious and deadly diseases in human history. It is an infectious agent that causes Acquired Immuno Deficiency Syndrome (AIDS), a disease that leaves a person vulnerable to life threatening infections. Though, there have been increase in the level of HIV awareness throughout the world and a lot of governmental and non-governmental organizations have invested huge funds, energy, and other resources into reducing the virus across the globe, but these alone cannot be enough for its extinction. In this paper, a bayes-based model technique is used to develop a predictive model for extinction of HIV/AIDS, our method is based on generating a dataset which is gotten by administering questionnaires as a means of eliciting responses from people or respondents, we used bayes-model to analyse this data. The result shows that in some years' time, there will be extinction (or reduce to control level) of HIV/AID if certain factors are carefully considered by all.

Keywords: Data mining, Bayesian classifier, Human immunodeficiency virus

Introduction

Humans have been "manually" extracting information from data for centuries, but the increasing volume of data in modern times requires the need for more automated approaches [1]. As datasets and the information extracted from these large data has grown in size and complexity, direct hands-on data analysis has increasingly been supplemented and augmented with indirect, automated data processing using more complex and sophisticated tools, methods and models. The availability of huge amount of data and the imminent need for turning such data into useful information and knowledge brought about the concept of data mining.

Data mining identifies trends (or knowledge) within data that go beyond simple data analysis which can be used for so many applications [9]. Data mining is the process of using computing power to apply methodologies, including new techniques for knowledge discovery from a large data [10]. The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction [7]. Knowledge discovery provides explicit information about the characteristics of the collected data. Predictive modeling provide predictions of future events, and the processes may range from the transparent (e.g., rule-based approaches) through to the opaque (e.g., neural networks).

Human immunodeficiency virus (HIV) is one of the most serious and deadly diseases in human history. The virus is an infectious agent that causes Acquired Immuno Deficiency Syndrome (AIDS), a disease that leaves a person vulnerable to life threatening infections (Emuoyibofarhe et al, 2011). Recently, HIV/AIDS has become a global enemy that every people, citizens, governments, and non-governmental organizations has been fighting to kick it out of their society if possible, to reduce it to the bearest minimum by increasing the level of people's sensitization against this infection. In this paper we developed a model to reduce to a control level and kick out this disease over time.

The remaining part of the paper is arranged as follows: In section 2, we discussed statement of problem, section 3 discusses the bayesian classifier. Overview of the proposed scheme is presented in section 4, and section 5 is conclusion.

Statement of Problem

HIV/AIDS has gained popularity and sufficient research time in the last two decades. Research has shown that it is most predominantly in people between the ages of 15-50. A lot of government and non-governmental organizations have been actively involved in eradicating the disease. Hitherto, there is no clear relevant predictive service available in the eradication of HIV/AIDS, Imagine a world without HIV/AIDS.

Bayesian Classifier

The Bayesian approach provides a mathematical rule explaining how one can change one's existing beliefs in the light of new evidence [2], bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.

Mathematically, Bayes' rule states that

$$posterior = \frac{likelyhood * prior}{marginal\ likelyhood} \qquad 1$$

Equation 1 can be written as

$$P(R = r|e) = \frac{P(e|R = r) * P(R = r)}{P(e)} \qquad 2$$

Where $P(R = r|e)$ denotes the probability that random variable R has value r given evidence e. The denominator is just a normalizing constant that ensures the posterior sums up to 1; it can be computed by summing up the numerator over all possible values of R, that is,

$$\mathrm{P(e) = P(R = 0, e) + P(R = 1, e) + \ldots} \qquad 3$$

$$\mathrm{P(e) = sum\_r\, P(e \mid R = r)\, P(R = r)} \qquad 4$$

This is called the marginal likelihood (since we marginalize out over R), and gives the prior probability of the evidence [3].

An instance is considered thus; suppose someone have been tested positive for a disease; what is the probability that the person actually have the disease? This depends on the accuracy and sensitivity of the test, and on the background (prior) probability of the disease. Formally to solve the above problem, let

$$P(Test = +ve \mid Disease = true) \ = \ 0.95 \qquad 5$$

so the false negative rate, $P(Test = -ve \mid Disease = true) \ = \ 5\%$.

Let

$$P(Test = +ve \mid Disease = false) \ = \ 0.05 \qquad 6$$

so the false positive rate is also 5%. Suppose the disease is rare: $P(Disease = true) \ = \ 0.01 \ (1\%)$. Let D denote Disease (R in the equation 2 above) and "$T = +ve$" denote the positive Test (e in the equation 2 above). Then,

$$P(D = true|T = +ve) = \frac{\mathrm{P(T = +ve|D = true) \ * \ P(D = true)}}{\mathrm{P(T = +ve|D = true) \ * \ P(D = true) + \ P(T = +ve|D = false) \ * \ P(D = false)}} \qquad 7$$

$$P(D = true|T = +ve) = \frac{0.95 \ * \ 0.01}{0.95 * 0.01 \ + \ 0.05 * 0.99} \qquad 8$$

$$P(D = true|T = +ve) = \ 0.161 \qquad 9$$

Thus, the probability of having the disease, given that the person tested positive is just 16%. This seems too low, but here is an intuitive argument to support it. Out of 100 people, we expect only 1 to have the virus, and that person may probably test positive. But we also expect about 5% of the others (about 5 people in total) to test positive by accident. So, out of the 6 people tested positive, we only expect 1 of them to actually have the virus; and indeed $1/6$ is approximately 0.16. In other words, the reason the number is so small is that, it is believed that this is a rare disease; the test has made it 16 times more likely that the person have the $disease \ (p(D = 1|T = 1)/p(D = 1) = 0.16/0.01 = 16)$, but, it is still unlikely in absolute terms. If one want to be "objective", one can set the prior to uniform (that is, effectively ignore the prior), and then get

$$\mathrm{P(D = true|T = +ve)} = \frac{\mathrm{P(T = +ve|D = true) * \ P(D = true)}}{\mathrm{P(T = +ve)}} \qquad 10$$

$$\mathrm{P(D = true|T = +ve)} = \frac{0.95 \ * \ 0.5}{0.95 * 0.5 \ + \ 0.05 * 0.5} = \ 0.95 \qquad 11$$

This, of course, is just the true positive rate of the test. However, this conclusion relies on the belief that, if one did not conduct the test, half the people in the world have the disease, which does not seem reasonable. If they all show up positive, then the posterior will increase. For example, if we conduct two (conditionally independent) tests T1, T2 with the same reliability, and they are both positive, we get

$$P(D = true|T1 = +ve, T2 = +ve) = \frac{\mathrm{P(T1 = +ve|D = true) \ * \ P(T2 = +ve|D = true) \ * \ P(D = true)}}{\mathrm{P(T1 = +ve, T2 = +ve)}} \qquad 12$$

$$= \frac{0.95 \ * \ 0.95 \ * \ 0.01}{0.95 * 0.95 * 0.01 \ + \ 0.05 * 0.05 * 0.99} \qquad 13$$

$$= \frac{0.009}{0.0115} \ = 0.7826 \qquad 14$$

The assumption that the pieces of evidence are conditionally independent is called the naive Bayes assumption. This model has been successfully used for classifying email as spam $(D = true)\ or\ not\ (D = false)$ given the presence of various keywords $(Ti = +ve\ if\ word\ i\ is\ in\ the\ text, else\ Ti = -ve)$. It is clear that the words are not independent, even conditioned on spam/not-spam, but the model works surprisingly well nonetheless. [4]

Proposed Scheme

In this work, the descriptive survey type was used, where a cross section of people (both male and female) from a local government in Ibadan city, Nigeria. We chose Ibadan (the capital city of Oyo State) as a case study because it is the third largest metropolitan area, by population, in Nigeria (after Lagos and Kano), with a population of 1,338,659 according to the 2006 census [8]. Ibadan is also the largest and most populous city in the country and the third in Africa after Cairo and Johannesburg [8].

We administered questionnaires and gave them to the people of this town. Twenty (20) items/questions was constructed and used in the study. The questionnaire was divided into two sections (section A and B). Section A was on personal bio-data information of the respondent, the important section for this work is section B which was designed to handle questions that will facilitate the analysis of people's view as regard HIV/AIDS infection and a dataset was generated from the answers provided to these questions.

We carefully selected four important attributes for our dataset collected via the questionnaire and these data were analysed, the attributes are age, precaution, stigma, and awareness. The table 1 summarized the response from our respondents indicating the values for precaution, stigma, and awareness while table 2 shows the distinct value for attribute age.

Table (1): table showing precaution, stigma, and awareness attributes

	Precaution	**Stigma**	**Awareness**
High	61%	58%	31%
Medium	39%	42%	69%
Low	0%	0%	0%

Table (2): table showing age attribute

	Age of infected people
Adult	23%
Youth	67%
Child	10%

Attributes

Precaution, is used to record the rate in percent at which people take to precautionary measures of HIV/AIDS infection.

Stigma, is used to record the rate in percentage at which infected people are being stigmatised.

Age, is used to record the age of infected people in percentage.

Awareness, is used to record in percentage, the rate at which people know about HIV/AIDS infection.

The class value, is derived by assigning weight to each attribute values such as high, low, child etc. Research has shown that the age that any infected person can live without any prejudice/factor trying to undermine the person existence is twenty four (24) years [6]. With a view to that above fact, weight is assigned to each attribute which in-turn help in estimating our class value for all probable combination of attributes values (for instance, precaution=high, awareness=medium, age=adult and stigma=medium).

The weight assigned to each attributes is as shown in the table 3 and 4.

Table (3): Weighted value for precaution, stigma, and awareness attributes

	Precaution	**Stigma**	**Awareness**
High	100% of 24years	10% of 24years	100% of 24years
Medium	50% of 24years	50% of 24years	50% of 24years
Low	10% of 24years	100% of 24years	10% of 24years

Table (4): Weighted value for age attribute

	Age of infected people
Adult	10% of 24years
Youth	50% of 24years
Child	100% of 24years

As shown in the proposed scheme and the analysis of data, if the level at which people take precaution is high (61%), stigmatisation is high (58%), awareness is medium (69%) and the age category of infected people is youth (67%), then in the nearest future HIV/AIDS can go out of existence.

Conclusion and Future Work

Most of the present effort to prevent or reduce HIV infection have typically focused on individuals at high risk for HIV infection rather than on PLHIV (people living with HIV/AIDS). But, there remains a relative paucity of programs to help ensure that HIV/AIDS can go out of existence.

Consequently, from our findings, we have shown that HIV/AIDS can go out of existence if certain factors are carefully considered by all in the future. Future work can be done in medical line especially with the latest information on the outbreak of ebola virus and continuing increase in the level.

References

1. Data Mining - http://en.wikipedia.org/wiki/Data_mining
2. The groups mostly affected by HIV/AIDS. Retrieved July 20, 2013 from http://www.aidsonline.org/india/the-groups-most-affected-by-hiv-aids.php
3. In praise of Bayes, (9/30/2000). Retrieved July 16, 2013 from http://www.cs.ubc.ca/~murphyk/Bayes/economist.html
4. **Alka Gangrade, Ravindra Patel 2012.** Privacy Preserving Naïve Bayes Classifier for Horizontally Distribution Scenario Using Un-trusted Third Party , IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727, Volume 7, Issue 6 (Nov. - Dec. 2012), PP 04-12
5. **Kevin Murphy 2013**. A brief introduction to Bayes' Rule. http://www.cs.ubc.ca/~murphyk/Bayes/bayesrule.html. Retrieved July 16, 2013

6. Living with HIV. **2013**. http://www.hivpositivemagazine.com/living.html. Retrieved July 16, 2013.
7. **AMM Tudorache 2008**. Data Mining and the Process of Taking Decisions in E-Business. Romanian Economic and Business Review – Volume 3, Issue 4, pages 111-116
8. http://en.wikipedia.org/wiki/Ibadan
9. **Jiawei Han, Micheline Kamber 2006**. Data Mining Concept and Techniques, second edition. ISBN 978-1-55860-901-3
10. **Cruz J.A, Wilshart D.S 2006**. Application of Machine Learning in Cancer Prediction and Prognosis. Cancer Infomatic 2(2) 59-77.