IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. II (Jul.-Aug. 2017), PP 01-07
www.iosrjournals.org

Pattern Discovery and Association Analysis To Identify Customer Vulnerable To HIV/AIDS: Case of Marie Stopes Gonder Branch Clinic

*Fistume Tamene[1], Fediu Akmel[2], Ermiyas Birhanu[2], Behar Siraj[2]
[1] *(Lecturer, Department of Computer Science, Wolkite University, Wolkite, Ethiopia)*
[2] *(Lecturer, Department of Computer Science, Wolkite University, Wolkite, Ethiopia)*
[2] *(Lecturer, Deparement of Software Engineering, Wolkite University, Wolkite,Ethiopia)*
[2]*(MSc Candidate, Departement of Software Enginnering, AASTU,Addis Ababa,Ethiopa)*
Corresponding Author: Fistume Tamene[1]

Abstract: *In the 30 years since HIV/AIDS was first discovered, the disease has become a disturbing pandemic, taking the lives of 30 million people around the world. In 2010 alone, HIV/AIDS killed 1.8 million people, 1.2 million of whom were living in sub-Saharan Africa. In Ethiopia,HIV/AIDS is one of the key challenges for the overall development of Ethiopia, as it has led to a seven-year decrease in life expectancy and a greatly reduced workforce. Even if there are a number of voluntarily counseling and testing centers that work on HIV/AIDS prevention located in several cities of the country, they didn't change and solve the problem related with HIV/AIDS. In addition in most of Countries counseling and Testing centers ,the data collected is simply put together and maximum used for statics purpose rather than analyzing to discover relevant and interesting previously unknown data characteristics,relationships,dependencies etc . The main objective of this study was pattern discovery and generating interesting hidden association rules from data which is taken from Marie stopes Gondar branch clinic. The contribution of this Study is by analyzing customer's data that did HIV/AIDS test on the clinic, to identify which customer is more vulnerable to HIV/AIDS. It helps counselors in VCT centers in predicting some hidden but interesting relationships among the attributes they use during the course of counseling. For doing this, methodology such as data collection and tool selection was used. After data was collected, the main data preprocessing tasks are applied on data sets to clean data and to make it ready for experiment purpose. Out of 1992 instances of original data 1861 was made ready for the experiment. Weka3.4. tool is used for experiment and the well known association rule mining algorithm Apriori was used to extract those interesting rules from data. In order to get those interesting rules three basic experiment was conducted .Experiment I was conducted by using the whole data set. Experiment II was conducted by considering only those positive classes. Experiment III was done by only considering those positive classes but with the absence of positive class attribute. One of the result of experiments showed that customers that donot use condom during sexual intercourse and non employed person are vulnerable to HIV/AIDS.*

Keywords: *Pattern Discovery , Association Analysis,Vulnerable ,Data Mining , HIV/AIDS*

Date of Submission: 28-06-2017 Date of acceptance: 15-07-2017

I. Introduction

Now a day HIV/AIDS is the most serious problems of every country in the world. Every day in the news we hear about people getting died. Too many people are taken away too early in so many reason. The greatest killer today is AIDS pandemic[1], which takes more lives than others.Because individuals in their most productive years (15-49 years old) are most commonly infected with HIV/AIDS, the disease has a wide socioeconomic impact that threatens development progress in many poor countries. Some estimates suggest that annual GDP growth in highly affected countries can be 2-4% lower than in countries with the absence of AIDS. According to World Bank report number of population estimated over 77 million at 2008 , this implies that the second most populated in Africa next to Nigeria and around one- fifth of are aged between 15-24 years. Ethiopia is one of the countries who lost more than 3 million people by HIV/AIDS[2]. The primary mode of HIV transmission in Ethiopia is heterosexual contact. Young women are more vulnerable to infection than young men, urban women are three times as likely to be infected as urban men, although in rural areas the difference between genders is negligible. Populations at higher risk for HIV infection include sex workers, police officers and members of the military. Currently in Ethiopia, there are a number of Governmental institutions and NGO's that work on HIV/AIDS. Among those voluntarily counseling and testing centers, Marie Stopes international is one of non governmental Institution that work on HIV/AIDS. Even if this institution plays an important role on

free counseling and testing HIV/AIDS, they didn't change and solve problems at all. In addition the collected data is simply put together and maximum used for statics purpose rather than analyzing to discover relevant and interesting previously unknown data characteristics, relationships between those attributes and dependencies.
This study is all about identification of customers that are vulnerable to HIV/AIDS in Gondar city. The most important benefit of this project is to help counselors in VCT centers in predicting some hidden but interesting relationships among the attributes they use during the course of counseling.

STATEMENT OF PROBLEM
Even if they have a number of benifites delived by the organization to the community around gonder ; the following listed problems that we have observed:-

- Lots of information collected from the customers is oftraditional file system format ; due to this it is difficult to extract new knowlwege.
- The clinic data are not integrated for further predictive purposes.
- The clinic does not have any representative model that serves as prediction.

The following are listing of research questions this study attempts to answer :

- Which rules are good for prediction of HIV/ADIS status ?
- How we can integrate predictive model with the orginal data ?

GENERAL OBJECTIVE
The general objective of this study was to apply pattern discovery and Association Rule mining algorithm in order to extract useful and interesting patterns that show important associations among different attributes and to find out a rule for prediction of the customer vulnerable to HIV/AIDS in Gondar city from customer data of Marie Stopes Gondar Branch Clinic.

II. Litreature Review

A study on Pattern discovery and Association analysis on identification of customer vulnerable to HIV/AIDS has been reported in some literature. There are some articles that were written on risks and vulnerability to HIV. According to article, facing the HIV/AIDS Pandemic [3], when HIV/AIDS was first identified in the 1980s, public health officials assumed its spread could be halted by informing people about how the virus was transmitted and how people could protect them from it and by safeguarding blood supplies. This approach to prevention was successful in politically organized communities with access to information and resources; the case of white gay men in North America, Australia and Western Europe.

Another article called Conducting a Situation Analysis of Orphans and Vulnerable Children Affected by HIV/AIDS [4].according to this article the impacts of HIV/AIDS on children, families, communities and countries are products of many interrelated factors and require responses that vary by family, community and country. These factors include the local pattern of the spread of HIV infection, economic activities, service availability, resources, public knowledge and awareness, the social environment, culture, the legal environment, and political leadership. For responses and interventions to be effective with a strategic use of resources, they must be informed by a working understanding of the most significant of these factors and how they relate to each other in terms of causality and mitigation of the devastating impacts.

III. Research Methodology

This study is experimental type, implementing an algorithm and evaluating the performance of the selected algorithm to solve the real-world problems[5]. Here under we listed the basic steps that we followed to accomblish this study.

Study Design
A retrospective study design was employed using data from Marie stopes clinic of Gondar branch which is found the city of Gonder. Records of HIV/AIDS cases over three years from January 2002 (2010) to December 2044(2012) were retrieved from the clinic for the study area. Thus, trends of HIV/AIDS were analyzed by age, sex, marital status, occupation and so on .

Sample Size and Population
All data recorded for HIV/AIDS cases in the time from January 2010 to December 2012 was included in the study. Total patients from city branch all age and sex group who took service for voluntarly test their blood from clinic was included in the study.

DATA COLLECTION

The data used in this study is collected from Marie stopes clinic of Gondar branch. The dataset used for this project is collected in year 2002 (2010), 2003(2011) and 2004(2012) ; which contains customer information who visited the clinic on those mentioned years for the purpose of voluntarily testing their blood.The original data consists of 27 attributes and 1992 records , including information concerning a particular voluntarily counseling and testing center.

EXPERMENTAL TOOLS USED

Before doing a reseach project, selecting the best tool which fits the intended goal is an essential task.since weka is the newly emerging and popular data mining tool and also freeware, the weka 3.4 is used for Association rule mining. The main reason for selecting WEKA was because of open source and platform independence[6]. In addition, Excel 2007 is used for data preparation.

DATA PREPARATION

The first activity after collecting data in any data mining project or research is data preparation. In parallel to the tool used selection, the other activity that has been conducted was preprocessing data in order to make it suitable for mining tool. Even if the data are stored in excel, all collected data are many problems such as missing values, Noisy data. In order to address such kind of problems and also other data related problems, the following data preprocessing tasks was applied on data.

DATA CLEANING

Data cleaning is an important step which allows you to solve problems related with attribute values such as incompleteness, redundacy and also to remove noisy data. Main problems of the original data are:

- **Missing value**: some attribute values are not available. Excel features are used to identify those attributes and experts comment are used in filling attribute values and removing instances. Examples of attributes with missing value: Occupation, expected exposure time.
- **Noisy Data**: some of attributes in original data contains values that are incorrect or invalid or doesn't give any sense. For instance, Customer that never had sex but condom use (this means the customer that doesn't ever had sex in his life but which used condom) and Education level illiterate but occupation manager(this means a customer that didn't read and write but he is the manager).

Since the data set was originally stored in Excel it is easy to apply data cleaning on original data using MS Excel features. In addition to this, experts are involved during data cleaning process. At the end of data cleaning process, some of rows with so many missing values are removed. After data cleaning step, out of 1992 data 1861 are selected.

DATA INTEGRATION

As mentioned on data collection, all data used for this project is collected from Marei stopes clinic of Gondar branch. The only thing done was integration of customer's data that is stored in 2002, 2003 and 2004 into one file.

DATA REDUCTION

Data reduction allows to reduce the size of data set using three basic data reduction strategies. Since the original data after data cleaning process contains 1861 instances, it is not necessary to apply numeriosity or size reduction. Since the size of data set is small it is not necessary to apply data compression. But the collected data set have 27 attributes. However all attributes are not important for the Data mining task to be developed. In order to identify important attributes, relevance analysis has been performed on data so as to remove some irrelevant and least important attributes based on expert's recommendation. In every data mining task, one can get some attribute that has little or no impact on the overall mining output. As mentioned above there are many attributes in the data that will be used for this study. After getting the description of the features, the domain experts were consulted for the selection of appropriate attributes that may help in discovering some patterns about the data. Hence, those 12 attributes that have **yes**relevant value, are selected for this study. The majority of the attributes are excluded automatically based on the recommendations of the domain experts.

DATA TRANSFORMATION

Data transformation plays an important role in converting data with different format into one common format[7]. So, tasks such as discretization and normalization are performed during data transformation. Moreover, in order to make it suitable for the experimentation, the dataset must be transformed into common format.Discretization was done on Maritalstatus, Age , Occupation, PrmeryreasonHere, Education attributes:

After finishing data transformation step, all data are saved in CSV file format. Then the CSV file is opened with Notepad and ARFF file components *@relation, @ attribute <attribute names>*are added with all attribute data type. All data set instances are also added next to @data. After doing this, the notepad data is saved with ARFF file extension so that to make it ready for experiment.

IV. Experimental Setting

As the main goal of this study was extracting useful and interesting patterns using Association Rule mining algorithm. The models were built with Apriori Algorithm using Weka 3.4 machine learing algorthims. Three experiments were conducted for this study with different rule setting.

APRIORI ALGORITHM

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent item sets for Boolean association rules. This algorithm finds frequent item sets using candidate generation[8]. It generates frequently occurring itemsets and finds association rules using minimum support and minimum confidence defined. To get the association among the attributes of the datasets, association rule mining task is performed using the Weka 3.4. *Apriori* algorithms.

In general, **Apriori** -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 in weka 3.4 means that the association rule we want to create has N number of rules, T type of metric, C amount of minimum confidence, D as amount of delta for decrease in minimum support, S as the significance level, U and M as the amount of upper bound and lower bound minimum supports respectively. More detail explantion of Apriori Algorithm paramters is discussed figure 1 shown below.

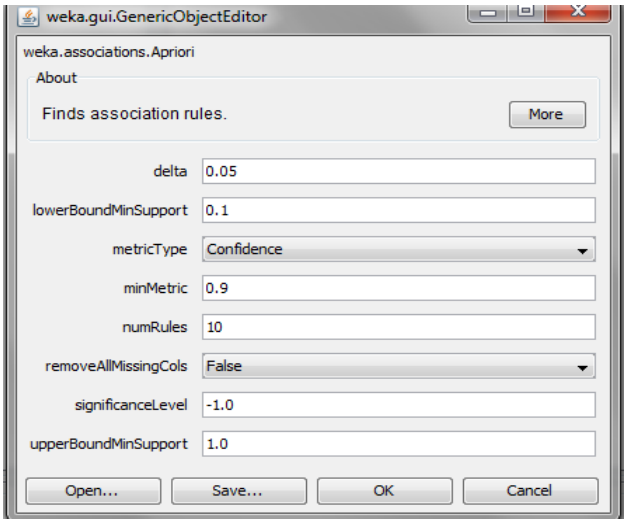

Fig 1: Parameters of Apriori Algorithm

V. Experment And Discussion

Experment 1 : First run

The first experiment is conducted by using the whole data set and the default parameter values of apriori algorithm. **Experiment I first run Analysis using default parameter values (N=10, C=0.9, S=0.3, D=0.05).** The first experiment generated more than 10 best rules using the default apiriori parameter values. But, out of those rules only 5 of them are interesting according to experts comment. i.e. 3, 4, 5, 6, 8.

Rule3:

If the customer is unemployed then with 93% confidence and with 61% support that this customer doesn't use condom.

Rule 4:

If the customer is female and if she unemployed then with 93% confidence and with 36% support that she doesn't use condom.

Rule 5:

If the customer age is between 18 and 25 and if the customer is unemployed then with 92% confidence and with 42% support this customer does not use condom.

Rule 6:

If the customer marital status is single and this customer had not made any sex, then with 92% confidence and with 33% support that this customer is free from HIV.

Rule 8:
If the customer is unemployed and if marital status of the customer is single then with 92% confidence and with 37 support that this customer doesn't use condom.

Experiment I: Second Run (using Number of rules 20)
In order to extract best rules, the second experiment was conducted by setting Number of rules 20. From those 20 Best rules, only the following are interesting according to Experts comment.i.e 10, 11, 13 and18.

Rule 10
If the customer age is between 18 and 25 and if this customer is female and if she is unemployed then with 94% confidence and 25% support that she doesn't use condom.

Rule 11
If the customer age is between 18 and 25 and if the customer primary reason for test is future plan then with 93% confidence and 25% support that this customer is free from HIV.

Rule 13
If the customer age is between 18 and 25 and if the customer primary reason for test is future plan then with 93% confidence and 25% support that this customer doesn't use condom.

Rule 18
If the Customer marital status is single and this customer had not made any sex and if the customer doesn't use condom then with 92% confidence and with 27% support that this customer is free from HIV.

Experiment I: Third run (using number of rules 30)
In this experiment the number of rule we used was 30. Out of 30, only the following are interesting according to expert's comment.i.e 21, 22 and 29.

Rule 21
If the customer residence place is urban and if the customer marital status is single and if this customer had not made any sex then with 92% confidence and with 27% support that this customer is free from HIV.

Rule 22
If the customer is female and if her residence place is rural then with 92 % confidence and with 25 % support that she doesn't use condom.

Rule 29
If the customer age is between 18 and 25 then with 91%confidence and with 56 % support that this customer doesn't use condom.

Experiment II: Positive Class
From all above experiments, the outputs don't show any explicit rule in connection with the customers having Positive HIV test result. Therefore, to get any rule that shows any interesting associations concerning *Positive* results and at the same time that fulfills the minimum support and confidence, the experiment has been done by considering only those records with Positive class labels. So, experiment II was done only by considering positive classes in search of any interesting patterns.

Experiment II: first run by considering only positive class
Experiment 2 which considered only positive class generated the above 10 best rules. Out of 10, only the following are interesting.i.e 1, 4,5,6,9 and 10.

Rule 1:
If the customer doesn't use condom then with 77% support and with 100% confidence that this customer is HIV positive.

Rule 4:
If the customer is unemployed then with 58 % support and 100% confidence that this customer doesn't use condom and this customer is HIV positive.

Rule 5:
If the customer is unemployed and doesn't use condom then with 58% support and 100% confidence that this customer is HIV positive.

Rule 6:
If the customer is unemployed and if the HIV result is positive then with 58% support 100% confidence then this customer doesn't use condom.

Rule 9:
If the customer had sex then with 57% support and with 100 % confidence that this customer is HIV positive.

Rule 10
If the customer lives in rural then with 54% support and 100 %confidence that this person is Positive.

Experiment II: second run (using number of rules 20)
From above 20 best rules generated, only the following are interesting.i.e 11,14,16,19 and 20.
Rule 11:
If the customer age is between 18 and 25 and if the customer doesn't use condom then with 53% support and with 100% confidence that this customer is HIV positive.
Rule 14:
If the customer is technician then with 48% support and with100 % confidence that this customer is HIV positive.
Rule 16:
If the customer is technician and if the customer doesn't use condom then with 46% and 100 % confidence that this customer is HIV positive.
Rule 19:
If the customer is technician then with 46% support and 96 % confidence that this customer is HIV positive and do not use condom.
Rule 20:
If the customer is technician and customer is HIV positive then with 46% support and with 96% confidence that this customer doesn't use condom.
The above two runs using positive class generated those selected interesting rules.since the aim of this project is to identify which customer with what kind of characters is more vulnerable to HIV, experiment III was conducted for those instances with positive class but with absence attribute HIVresult.

Experiment III: Positive classes with the absence of HIVresult attribute.
The goal of this particular experiment is to find out association between/among the features that lead to HIV positive result.
From the above ten best rules, the following are selected as interesting rules according to specialists comment.i.e 1,3,4,5,6,9,10.
Rule 1:
If customer is unemployed then with 58% support and with 100% confidence doesn't use condom.
Rule 3:
If customer age is between 18 and 25 and if customer is unemployed then with 41% support and with 100% confidence this person doesn't use condom.
Rule 4:
If customer is unemployed and a technician then with 36% support and with 100% confidence this person doesn't use condom.
Rule 5:
If customer is unemployed and had a sex then with 32% support and with 100% confidence this person doesn't use condom.
Rule 6:
If customer is unemployed and primary reason of customer for test is symptom then with 32% support and with 100% confidence this person doesn't use condom.
Rule 9:
If customer age is between 18 and 25 and a technician then with 32% support and with 98% confidence this person doesn't use condom.
Rule 10:
If customer age is between 18 and 25 and lives in rural then with 41% support and with 97% confidence this person doesn't use condom.

VI. Conclusion And Future Works

CONCLUSION
At the end of conducting those three above experiments and after a analyzing interesting association rules found, the following findings and the conclusions are given accordingly:

- The first important finding is that according to rule 3,5 and 8 in experiment I, most customers that are not employed are most of the time doesn't use condom. On other hand, according to Experiment II rule 1, most people who are HIV positive are found to be non-condom users. In addition, people who don't use condom and are HIV positive comprise 19.13% of the total customers. Not using condom has a significant contribution for being HIV Positive. So, customers that do not use condom during sexual intercourse are vulnerable to HIV.
- According to rule 6 of Experiment I, the majority of customers that are single and had not made sex are found to be HIV negative. In other words peoples that are abstinence are safer than those that did sex.these

result found preferring to abstain is one of the prevention mechanisms of HIV prevalence. Among those people who are single, who never had sexual intercourse and are HIV negative comprise 33.6% of the total customers. On the other hand, according to experiment II rule 9, most customers that had sex are found to be HIV positive. So peoples that are single and had sexual intercourse are more vulnerable to HIV.

- According to rule 11 of experiment I, those customers whose age is between 18 and 25 and whose reason for test is future plan are found to be HIV negative. On other hand according to experiment II rule 11, those customers whose age is between 18 and 25 and who are non condom users are found to be HIV positive. This reflects that being at this age range alone does not guarantee being HIV negative. But, this result doesn't seem sound because this may be the reflection of the above finding; i.e. customers that do not use condom are vulnerable to HIV.
- According to rule 22 of experiment I, most female that lives in rural use condom rarely. On other hand as per rule 10 of experiment II most peoples that live in rural are found to be HIV positive. This reflects that those females that live in rural area are more vulnerable to HIV.

FUTURE WORKS

- As in all most experiment shown, all of HIV victims are not condom users.Therefore an awareness creation program should be strengthened to encourage people to use condom specially for peoples that lives in rural and also free distribution of condoms should be initiated. So the concerned bodies should support those activites financially.
- The customers in the age below 18 are found HIV negative, as it is stated above. This is not a result of being in this age range; rather it is because of the fact that the majority of them did not start sexual intercourse. Therefore, abstinence, especially at early age, should be encouraged and continuous education should be offered for people in this age.
- It is discovered that almost all the customers who are HIV positive are not condom users and are unemployed. This strengthens what is just recommended above. Hence, efforts should be done to increase the rate of condom usage.
- As the experiment shows, there are some determinant features for vulnerability of HIV/AIDS. Therefore, a special emphasis should be given for those features while dealing with HIV/AIDS related activities and other similar studies.
- In this experiment, it is found that customers at the age range of 18-25 do not use condom. So, it is recommended that further studies can be done to find out the rationale behind it.

Acknowledgments

First of all we would like to thanks our instructor Dr.Mellion Meshesha for his guidance and suggestions to comlete this study well. Then, we are happy to acknowledge Ato Dejene Seyoum who is head of the Marie Stopes Gondar Branch Clinic for providing all necessary information and data. Lastly we would like to thanks sister Hayat and laboratory technicians' Ato Meakuanint who provide their important comment in every point that need specialist knowledge.

References

[1] Comfort A, Patricia A, and Emmanuel I: 'Acceptability of voluntary counselling and testing among medical students in Jos, Nigeria', 2010, 4, (6), pp. 357-361
[2] K, T.G.H.A.P.T.H.E.W.R.L.D.B.A.N.: 'HIV / AIDS in Ethiopia An Epidemiological Synthesis', in Editor (Ed.)^(Eds.): 'Book HIV / AIDS in Ethiopia An Epidemiological Synthesis' (2008, edn.), pp.
[3] Lamptey, P.: 'Facing the HIV/AIDS Pandemic', 2002, 57, (3), pp. 38-42
[4] Williamson, J.: 'Conducting a Situation Analysis of Orphans Vulnerable Children Affected by HIV/AIDS', in Editor (Ed.)^(Eds.): 'Book Conducting a Situation Analysis of Orphans Vulnerable Children Affected by HIV/AIDS' (2004, edn.), pp. February
[5] MIKIYAS G: 'Content-Based Classification Of Ethiopian Nations Music Video Clip '. Msc, Bahir Dar University, 2016
[6] Ermiyas B, and Feidu A: 'Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University', *International Journal of Advanced Research in Computer Science and Software Engineerin*g, 2017, 7, (2), pp. 46-50
[7] Dametaw, A.: 'Designing A Predictive Model for Heart Disease Detection Using Data Mining Techniques', Addis Ababa University 2011
[8] Jiawei Han, and Kamber, M.: 'Mining Frequent Patterns, Associations and Correlations' (2006 2006)

Fistume Tamene. "Pattern Discovery and Association Analysis To Identify Customer Vulnerable To HIV/AIDS: Case of Marie Stopes Gonder Branch Clinic." IOSR Journal of Computer Engineering (IOSR-JCE) 19.4 (2017): 01-07.