Prediction of HIV Status in Addis Ababa using Data Mining Technology

Tewodros Zewdu
HiLCoE, Computer Science Programme, Ethiopia
tedimel2007@gmail.com

Tibebe Beshah
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

HIV/AIDS continues to be a major global health priority. Knowledge about HIV status helps both the individual and the community. In spite of the widely and freely available VCT centers in Addis Ababa, most people often do not know their HIV status. One of the solutions for this problem is to predict the HIV status of the population using data mining techniques to identify the most affected part of the population to support prevention programs. The purpose of this paper is to construct and implement HIV status predictive model to scale up the knowledge of HIV status in Addis Ababa.

The general approach of the methodology is the CRISP-DM methodology which includes the following six steps: business understanding, data understanding, data preparation, modeling, evaluation and deployment. Voluntary Counseling and Testing (VCT) centers clients' data in Addis Ababa is used. Microsoft Excel and WEKA 3.6 tool, respectively, were used for further preparation of the data and as data mining tool to implement experimentations using algorithms such as J48, PART and Naïve Bayes. Moreover, WEKA 3.6 was also used to balance the imbalance data as HIV positive clients' data are out weighted by HIV negative clients' data.

The conventional algorithms such as PART, J48 and Naïve Bayes have performed poorly to predict correctly the minority class (HIV positive clients). Therefore, this problem is solved by balancing the data. As a result, pruned J48 classifier that predicts HIV status with 93.95% accuracy is constructed. The paper concluded by identifying HIV status determinant factors, developing HIV status predictive system (HSPS) showing socio-demography, behavioral and clinical result attributes are sufficient enough to predict HIV status of an individual.

Keywords: Prediction; HIV Status; HIV/AIDS; Classification; Classifiers; Prototype

1. Introduction

The emergence of the HIV epidemic is one of the biggest public health challenges the world has ever seen in recent history. Ethiopia is among the sub-Saharan countries most affected by the HIV epidemic. According to the Country Progress Report on HIV/AIDS response in 2012 [1], with an estimated adult prevalence of 1.3%, there are a large number of people living with HIV (approximately 800,000).

Despite these mounting challenges, the global response has been a reason for hope and optimism in fighting the epidemic.

HIV/AIDS Voluntary Counseling and Testing (HCT) services were started in Ethiopia following the endorsement of the national AIDS policy in 1998. This helps a national plan of action to increase access to HIV services in HIV/AIDS care and emphasizing universal access to prevention, treatment, care, and support services [1]. Accordingly, over the past years, the health sector response to HIV focused on scaling up HCT services to new sites and strengthening existing ones to provide optimal services at all levels.

The World Health Organization (WHO) [2] remarked that one of the priority intervention areas is HIV/AIDS Voluntary Counseling and Testing (HCT). The HCT guideline also emphasizes HCT as

a crucial intervention component of the HIV/AIDS prevention, care and support program.

The Knowledge of HIV status of the population helps domain experts, policy makers and service providers by guiding them how to design and where to implement their programs.

The World Health Organization [2] asserted that greater knowledge of HIV status within a community is critical to expand access to HIV treatment, care and support in timely manner as it offers people with HIV an opportunity to receive information and tools to prevent HIV transmission to others.

One of the underlying concerns of HCT policy makers and/ or service providers is the scaling up of the knowledge of HIV status. Knowledge about HIV status is only through HIV testing. In spite of the widely and freely available VCT centers in Addis Ababa, most people often ignore the benefit of HIV testing and hence a lot of people do not know their HIV status.

One of the solutions for this problem is to predict the HIV status of the population from the available data in these VCT centers.

Domain experts use the statistical report in reporting the number of HIV positive and negative clients in terms of socio-demographic and behavioral variables. These traditional methods of data analysis have limited capacity to discover new and unanticipated relationships that are hidden in conventional databases [3].

The information stored in HCT centers can be used beyond for the purpose of monthly, quarterly or annual report. These input data can also be used as determining factors in predicting the HIV status of the population. The identification of determining factors provides a foundation up on which special intervention programs can be designed and/or existing programs can be improved to increase the response of clients.

Data mining provides the methodology and technology to transform these massive data into useful information for decision making and problem solving [3]. Data mining is a dynamic research capable of extracting hidden relationships from these input variables to identify factors that determine the HIV status of clients. This research has attempted to identify determinants of HIV status and predicts the HIV status of the population by analyzing VCT data pattern using recent data and more variables so as to support the scaling up of knowledge of HIV status. The prediction is based on the individual HIV status prediction. Predictive data mining methods are used as pattern recognition tools in data mining to classify HIV status of individuals based on demographic and socio-economic characteristics.

Finally, it has been attempted to obtain answers for the following main research questions:

- What are the main determinant risk factors that cause HIV infection in Addis Ababa?
- Which data mining technique is more appropriate to identify these determining factors?
- How the prototype of HIV virus prediction system can be developed?

2. Literature Review

The paper by Rosma *et al.* [4] described the feasibility of applying data mining technique to predict the survival of AIDS. An adaptive fuzzy regression classification technique, FuReA, was used to predict the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. The authors revealed that neural network model was able to predict the survival of AIDS with an accuracy of 60% to 100% based on selected dependent variables such as CD4, CD8 and viral load counts.

Particularly, a similar work to this study was conducted by Taryn [3]. According to the Author, HIV status can be predicated using Neural Networks and demographic factors. The primary objective of the paper was to use artificial intelligence methods, namely, neural networks to perform knowledge discovery and data mining on HIV clinical and demographic data, resulting in a classifier of HIV status of a patient based on demographic inputs. In

this study, supervised learning was used to train multilayer perceptions and radial bases networks to classify the HIV status of an individual, given certain demographic factors. The target population is women who are pregnant. The variables obtained in the study are: race, region, age of the mother, age of the father, education level of the mother, gravidity, parity, province of origin, and HIV status.

According to this study, all neural network architectures produced similar results but the average accuracy was between 61 and 62% and the author, finally, concluded that demographic data is not sufficient to accurately predict HIV status and this value is inadequate for medical classification.

However, this study has disproved the result and shown that HIV status can be predicted using demographic, behavioral and clinical data.

3. Data Preprocessing

The source of data for this research has been collected from HCT centers database in Addis Ababa. Database in VCT centers is manipulated using Epi-info software, in REC format. It has been exported from Epi-info to Microsoft Excel. Initially, there were 125,378 records with 80 attributes. The summary of clients by region is shown in Table 1.

Table 1: Summary of Clients by Region

Region	*Frequency*
Amhara	900
Afar	528
Dire Dawa	108
SNNP	2600
Tigray	483
Addis Ababa	120,756
Missing value	4
Total	*125,378*

Only relevant attributes for this paper have been selected: Age, Sex, Education Level, Employment, Marital status, Condom use, Had ever sex, Previously tested, Casual partner, Steady partner, and HIV status (target variable). Simple statistical summary is performed to verify the quality of the data set such as missing values, outliers and to obtain high level information regarding the data and these defects have been handled well.

After preprocessing and cleaning of data had been done, the data was found imbalance.

- 65,422 (84%) HIV negative clients
- 12,902(16%) HIV positive clients
- Total of 78,324 records with 11 attributes

However, such size is considered so imbalanced that the dataset misleads classification performance and biased to the majority class.

Lokanayaki and Malathi [5] remarked that a well balanced dataset is very important for creating a good prediction model. Medical datasets are often not balanced in their class labels. Most existing classification methods tend to perform poorly on minority class examples when the dataset is extremely imbalanced.

Zhai [6] also explained that with imbalanced datasets conventional way of maximizing overall performance will often fail to learn anything useful about the minority class.

According to Laza [7], sampling strategies have been used to overcome the class imbalance problem by either eliminating some data from the majority class (under-sampling) or adding some artificially generated or duplicated data to the minority class (over-sampling).

However, according to Chawla [8], random over-sampling leads to over-fitting as there will be multiple copies of minority examples and random under-sampling may cause the classifier to miss important concepts. They proposed an over-sampling approach to overcome the over-fitting and broaden the decision region of minority class examples. This approach is Synthetic Minority Over-sampling Technique (SMOTE).

In this technique, the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with duplicated real data entries. Therefore, in this paper SMOTE technique in WEKA

tool has been used to correct the imbalanced nature of the data.

4. Experimentation and Modeling

Classification algorithms such as J48, PART and Naïve Bayes were trained using stratified 10-fold validation with the given dataset. While training these classifiers, the values of the variables including the target variable were given. Where as, during prediction, given the values of the other 10 variables (attributes), the value of the target attribute (HIV status) has been predicted.

Experiments on the two classification algorithms namely J48 and PART were performed by modifying parameters such as confidence factor, number of instances per a leaf, pruning and unpruning using WEKA tool. These parameters can increase or decrease the complexity and accuracy of the generated rules. Pruning may decrease complexity but at the cost of accuracy. Reducing confidence factor helps to identify relevant attributes and reduces complexity. Moreover, increasing number of instances per leaf reduces the complexity of the generated rules but at the cost of the accuracy of the prediction. Naïve Bayes classifier was trained with default parameters as it doesn't have the above options.

The optimized result is a model with an excellent prediction performance and less complex and sensible rules. Therefore, according to these options, 11 experimentations were performed and evaluated. Overall accuracy is not a good measure in case of imbalanced data as it only considers the majority class of the data. Therefore, Confusion Matrix and ROC curve were also taken as evaluation measures of predictive performance. The confusion matrix shows the predictive performance of the model both on minority and majority classes. The ROC curve shows the trade off between true positive and false negative prediction. Moreover, complexity and acceptability of rules should be considered during comparison of models. The result of each experiment is shown in Table 2.

The overall accuracies of the models are found comparable. However, only the two models could generate less complex rules and a one mode, Pruned J48 with 100 numbers of instances per leaf and confidence factor of 0.25 could yield more sensible rules. The overall prediction performance of this model is 93.95 %.

Moreover, it has registered 98.1% of average ROC curve. This means that the ROC coverage area of this model is 98.1% which is near to perfection (100%). This shows the model predicts HIV positive correctly as positive but not at the cost of predicting HIV negative as positive. Similarly, it predicts HIV negative correctly as negative but not at the cost of predicting HIV positive as negative.

Table 2: Experiment Results

Exp	*Scheme*	*NL*	*NR*	*Acc.%*	*WROC*
1	J48-C0.25-M2	456	-	95.11	98.8
2	*J48-C 0.25-M 100*	126	-	93.95	98.1
3	J48-C0.1-M 2	379	-	95.02	98.7
4	J48-U-M 2	875	-	95.20	98.9
5	J48-U-M 100	180	-	93.98	98.4
6	PART-M 2-C0.25	-	285	95.16	99.1
7	*PART-M100-C0.25*	-	72	94.01	98.3
8	PART-M2-C0.1	-	262	95.11	99.0
9	PART-U-M 2	-	1002	95.20	99.0
10	PART-U-M 100	-	289	94.07	98.8
11	Naïve Bayes	-	-	80.77	89.6

Key: Exp = Experiment Number, M = minimum number of instances per leaf, C = Confidence factor, U = Unpruned, NL = Number of Leaves, NR = Number of rules, Acc. = Accuracy, WROC = Weighted Average ROC Area. "-" = there is no information.

The confusion matrix shows that out of the total of 65,417 actually HIV negative clients, 61873 (94.58%) clients are classified as HIV negatives and the rest are misclassified as HIV positive. And out of the total of 64,535 actual HIV positive clients, 60,223 (93.31%) clients are classified as HIV positive and the rest are misclassified as HIV

negative. These figures show that the model has shown almost equal prediction performance in terms of correctly classifying HIV negatives and HIV positives.

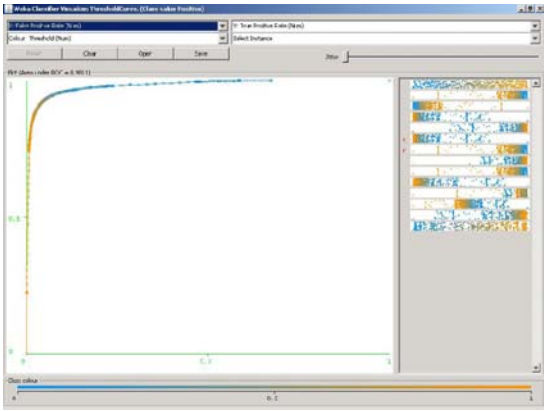

Figure 1: ROC Curve of the Selected Model on the Modified Data

In order to check if the classifier performs well on the imbalanced data, it had been trained on the dataset before sampling was performed. The proportion of the actual data size is 65,417 (84%) HIV negative and 12,907 (16%) is HIV positive. The overall accuracy of the model is found to be 90.84% and it registered 88.8% of ROC curve, as shows in Figure 2.

As shown in Figure 1, the ROC coverage of this model is 88.8%. This shows that the model predicts HIV negative correctly as negative but it is at the cost of predicting HIV positive as negative.

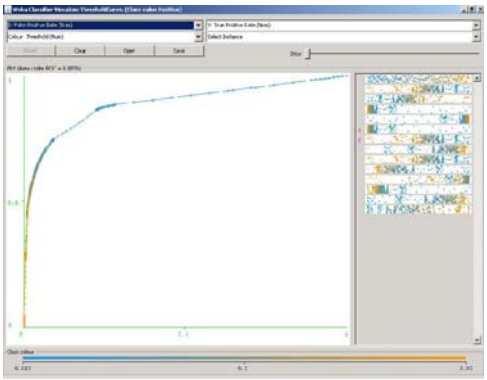

Figure 2: ROC Curve of the Selected Model on the Imbalanced Data

Moreover, the confusion matrix shows the same result that out of the total of 12,907 actual HIV positive clients, only 7331 (56.79%) are classified as HIV positive and the rest are misclassified as HIV negative and out of the total of 65,417 actual HIV negative clients, 63,825 (97.56%) are classified as HIV negative and the rest are misclassified as HIV positive.

Generally, the evaluation measures show that the classifier performs poorly on the minority class of imbalanced data.

5. Result and Discussion

Since the purpose of this research is to build and implement HIV status predictive model, determinant factors for vulnerability of HIV virus have also been found. The results are obtained using the application of Pruned J48 classifier that could perform classification with an overall accuracy of 93.95% in predicting the HIV status of an individual.

Similar study conducted by Taryn [3] showed that the average accuracy was between 61 and 62% to predict HIV status of an individual and concluded that demographic data is not sufficient to accurately predict HIV status, and the value is inadequate for medical classification. However, the study has weakness to achieve the best result. First, the method used in the paper was neural network. This method is mostly used for dataset which has complex relationship among attributes. There is no complex relationship among variables of the data in this study. Second, neural network doesn't allow to update and interpret the rules generated by it. It is a black-box that only shows the prediction performance of the model. The rules are not displayed to a human to update, interpret and get important information. Third, the dataset used in this paper was small in size (around one thousand). This reduces the prediction accuracy of the model as more data the algorithm get to learn, the better the accuracy will be. Fourth, despite the fact that the two classes (negative and positive HIV status) are extremely imbalanced in number, i.e., the number of HIV negative out-weighted the number of HIV positive clients' data. Equal importance should be given to positive HIV status and negative HIV status prediction. In order to have a model that treats the two classes equally in

imbalanced dataset, either the algorithms or the dataset itself should be modified using some techniques. Therefore, the sampling technique used in the above paper was random over-sampling which works by duplicating the existing data. However, this sampling technique results in over-fitting. When over-fitting occurs, the accuracy on the test data is reduced. This also reduces the prediction accuracy of the model. Finally, the paper stated that the method used to handle missing values was not good when there are multiple missing values and as a result this reduces the prediction accuracy of the model. Because of these weaknesses of the paper, the prediction accuracy was reduced.

Our work disproved the result of the above paper by showing that it is possible to predict HIV status of an individual using data mining technique with 93.5% prediction accuracy and demography data is sufficient enough to accurately predict HIV status of an individual. This result has been achieved by focusing on the weaknesses of the above study and taking appropriate approaches as discussed below.

Decision tree and rule based methods are used in this paper. These methods are easy and human interpretable. Unlike neural network algorithm, the rules (output) of these algorithms are displayed to interpret. As a result the rules can be modified according to comments from domain experts and personal judgments. The dataset used in this paper was also large enough in size (78324) to increase the prediction accuracy of the models. Moreover, as these conventional algorithms are not good enough to construct a good model using imbalanced dataset, appropriate sampling technique (SMOTE) was used to balance this imbalanced dataset. Unlike the sampling technique used in the study above, this sampling technique works by creating synthetic samples rather than duplicating the existing ones. This avoids over-fitting and increases the prediction accuracy of the model. In addition to these very important amendments of the above study, inconsistent, missing and outlier values in this paper were handled by simply deleting these values as the size of the data was large enough. This has resulted in higher prediction accuracy of the constructed model. Finally, unlike the above paper, in this study, parameter settings that result in optimum result are identified and a prototype of HIV status predictive system has been developed.

6. Prototype Development and Implementation

In this section, HIV status prediction system (HSPS) is developed that can assist HIV/AIDS intervention program domain experts, policy makers or service providers in predicting HIV status of the population based on the socio-demography, behavior related and clinical result attributes. The system is developed in a Java environment. The HSPS converts codes written in Java and integrates the results on to Java interface.

Figure 3: The HSPS Interface

The main functions of the HSPS interface shown as Figure 3 include: input clients' data section where users input 10 pieces of attributes.

a. Predict button: users click the button to get the result.
b. Clear button: users click the button to clear the previous input
c. Exit button: users click the button to leave the interface
d. Prediction result: this text box shows the prediction result of the provided data
e. And text area that displays determinant factors.

The prediction result displays "Vulnerable" if the client is vulnerable to the virus and "Not Vulnerable" if the client is not vulnerable to the virus. Moreover,

it displays "No such rule" if no rule was generated for such a pattern.

7. Conclusion and Recommendation

In this paper, supervised learning was used to train the three classification techniques to classify the HIV status of an individual, given certain socio-demographic, behavioral clinical result factors.

The paper is concluded by stating the following important points.

First, although the two classes (HIV negative and HIV positive) are imbalanced in size, equal importance should be given to their predictions. Even though the cost of predicting HIV positive as HIV negative seems more serious than the cost of predicting HIV negative as HIV positive, equal importance has been given to each class prediction. As a result, the poor prediction performance of models on the minority data should not be ignored. Therefore, by using proper sampling technique, optimum result has been achieved.

Second, the paper disproved the work of Taryn [3] who said that demographic, behavioral and clinical data are not sufficient to predict HIV status. This research has shown that HIV status can be predicted using demographic, behavioral and clinical data with prediction accuracy of 94%.

Third, in addition to HIV status predictive model, the main findings of this paper are the determinant factors and parameters that result in optimum output, and the prototype of HIV status predictive system.

Fourth, the model works for indefinite time but can be modified by running the same model on the new modified data and the prototype can be enhanced in the future to make it complete.

Fifth, this study has several contributions such as it provides useful insights to HIV/AIDS prevention programs for policy makers, domain experts and service providers. The clients (who need the service) of VCT centers or other prevention programs can benefit if VCT centers are installed or modified and implemented according to the output of this paper. The paper can also be used as additional creative reporting mechanism to statistical report. Moreover, it can invite other interested researchers to explore more in related and similar areas.

Finally, the paper has recommended other researchers to conduct a research on the same topic using relevant and different attributes. Besides, the dataset in this domain is usually too imbalanced for conventional algorithms to perform equally and correctly on the two classes. Therefore, research areas are recommended to be conducted by modifying these conventional algorithms to perform well and equally on both classes when the data is imbalanced.

References

[1] FHAPCO, "Country Progress Report on HIV/AIDS Response", GAP Report, Addis Ababa, Ethiopia, March 31, 2012.

[2] WHO, "Scaling up Priority HIV/AIDS Interventions in the Health Sector," Progress Report, Sept. 2010.

[3] Taryn Nicole Ho Tim, "Predicting HIV Status Using Neural Networks and Demographic Factors ", Johannesburg, April 2006.

[4] Rosma, Sameem, Kareem, Basir, and Adeeba Annapurni, "The Prediction of AIDS Survival: A Data Mining Approach," In Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Applications in Science and Engineering, Vol. 2, 2007.

[5] Lokanayaki and Malathi, "Data Preprocessing for Liver Dataset Using SMOTE," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 11, Nov. 2013.

[6] Zhai, "An Effective Over-sampling Method for Imbalanced Data Sets Classification," Chinese Journal of Electronics, Vol. 20, No.3, Jul. 2011, pp. 489-494.

[7] Laza, "Evaluating the Effect of Unbalanced Data in Biomedical Document Classification", Journal of Integrative Bioinformatics, Vol. 8, No. 3, Sep.2011, pp. 177.

[8] Chawla, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321–357.