**PREDICT HIV PROGRESSION**

Team B

Ning Zhong (zhongn2@miamioh.edu)

Akshaya Ravichandran (ravicha@miamioh.edu)

Department Name, Miami University

501 East High Street, Oxford, Ohio 45056, USA.

***Abstract***

*In this project, we aim to predict the proliferation of HIV virus in a patient to see if the infection becomes less severe using various Machine Learning Algorithms like Linear regression with one variable, Linear regression with multiple variables, Logistic regression and SVM algorithm with Gaussian Kernel. This project is developed in response to a real-life competition posted on Kaggle with a training and test dataset having the same 6 features. We train the Machine Algorithms with the training data and test it with the given test dataset and calculate the prediction accuracy of each algorithm. Later, we compare the results of each algorithm and realize that SVM algorithm has the best training accuracy of 98.70 and test accuracy of 98.55. We conclude that SVM algorithm with Gaussian Kernel is the best suited algorithm for predicting the progression of HI virus in a patient.*

**1 Introduction**

We plan to predict the likelihood that an HIV patient's infection will become less severe, given a respectable dataset and clinical information. This can be done by having a better understanding of the virus causing this infection by getting a genetic handle of its blueprint. Before getting into the details, it is essential that we give a brief description about HIV, AIDS and its consequences to illustrate the importance of the project.

**1.1 HIV and AIDS**:

Human Immunodeficiency Virus (HIV) is a virus which causes Acquired Immuno Deficiency Syndrome (AIDS). AIDS causes the immune system of human beings to fail. If the immune system fails, the body becomes susceptible to other deadly diseases [1]. If the virus is only suppressed and not cured completely, there are chances that the virus would evolve with time by mutating and become resistant to the treatment that was previously effective. Till date, a completely cure for AIDS has not been found [1]. Hence, it is very essential to find a way to predict the likelihood that an HIV patient's infection will become less severe. In order to do so, an understanding of the genetic blueprint of the virus causing the infection is necessary.

**1.2 DATASET and FEATURES**:

The Dataset is obtained from a real-life competition posted on Kaggle [2]. Here’s the link for the data set: <https://www.kaggle.com/c/hivprogression/data>

The features available in the dataset are Patient ID, Responder Status (binary variable determines whether the patient recovers or not), Protease (PR) sequence, Reverse Transcriptase (RT) DNA sequence, Cluster of Differentiation (CD) 4 count and Viral Load. At this point, we do not know which feature or combination of features would give us the best performance. So, we plan to use some combinations of features in the machine learning algorithms that are implemented for this project.

Significance of each feature in the data set:

Reverse Transcriptase (RT) is an enzyme that is responsible for the replication of retroviruses like HIV. Viral Load: This variable is used to determine whether a treatment is working on a patient for a given disease. CD4 count: Higher count represents both a healthier patient and a higher amount of HIV reproduction. Responder Status: 1 – Patient recovers; 0 – patient doesn’t recover. Protease Sequence - Protease activity initiates virion maturation [3].

**2 Related Works**

“An Analysis on the Prediction of HIV Progression” [3] inspired us much by giving a very comprehensive description on the topic. It is interesting to notice that it uses the same dataset posted in the Kaggle competition. Several Machine Learning algorithms such as Multinomial Bayes Classifier, Less Naïve Generative Bayes Classifier, Decision Trees, are employed to predict the change in severity of the infection. In the end, the authors conclude that viral load, CD4 count and the RT sequence were the features that were related to the replication of the HIV. Looking at the results provided by this paper, we decided concentrate on the features addressed in this paper for training out dataset. An important result to be noted in this paper is that, the authors were successful in showing that there was 62.1% correlation between the viral load and patient’s health.

“Applying machine learning to predict patient specific current CD4 cell count in order to determine the progression of human immunodeficiency virus (HIV) infection” [4] also did a similar work. In this paper, machine learning is applied to predict the CD4 count using PR & RT sequence, viral load and time. The authors created three different groups of inputs and fed into SVM and neural network algorithm separately and observed that SVM performed well by predicting the CD4 cell count with an accuracy of 95% when the viral load, time and PR & RT sequences were taken as input. The dataset used in this paper is different from what we are using.

“Multi-Task Learning for HIV Therapy Screening” [5] tried to predict the outcome of a therapeutic treatment for a patient infected with HIV virus using a machine learning technique named multi task learning. The prediction accuracy is improved substantially. These authors did not consider how changes in the PR and RT sequence affected the patient with HIV virus. Also, this paper deals with the effect of a drug treatment on a patient which is different from what we are doing.

In the paper of “Support vector machines to forecast changes in CD4 count of HIV-1 positive patients” [6], the authors built a SVM model with the input of genome, viral load and time to predict the range of CD4 change and provided an accuracy of 83%.

After gaining insight by reading the papers cited above, we plan to use four widely known machine learning algorithms in our project to predict the proliferation of HIV virus, which is explained in the next section.

**3 Approaches**

The dataset collected from Kaggle website contains two parts: training data with 1000 entries and test data with 692 entries and both are in CSV (comma-separated Value) files. There are total six features in both the training and test data sets. They are: patient ID; responder status (1 if the patient recovers and 0 otherwise), Protease nucleotide sequence (PR), Reverse Transcriptase nucleotide sequence (RT), viral load at the beginning of therapy (VL) and CD4 count at the beginning of therapy (CD4). Our dataset has 6 features and by reading the papers cited in the related work section, we use features like CD4 count, Viral Load and responder status to predict HIV progression as those features seem important.

We plan to use Linear Regression as our first attempt for prediction. In our first attempt, we combine features like Viral Load (VL) and CD4 count and use that to predict the responder status (RS) by applying linear regression algorithm on those features. By implementing gradient and cost function, we can come up with an accuracy result for HIV prediction.

Secondly, our idea is to combine both VL and CD4 features and apply linear regression with multiple variables. By calculating the cost function and gradient descent, the probability that a person with HIV improves/does not improve is calculated using this algorithm.

Thirdly, we will use logistic regression with three features: VL, CD4-count and resp status. By applying cost function and creating the model, we will use the prediction generated from our model and compare the results with test data to obtain the accuracy.

Finally, SVM with Gaussian Kernel is widely-used in most non-linear problems. Firstly, we will process our raw data in the Gaussian Kernel and then train the processed data as our model. After obtain adequate information from the model, we will predict the probability for each entry and then, compare it with raw data to get the accuracy as well.

After implementing the algorithms, all four algorithms are compared based on their accuracy of HIV prediction to determine the algorithm that provides the best result. Since, we are using four different Machine Learning Algorithms to predict the proliferation of HIV virus; we are hoping that at least one of these algorithms would outperform the results obtained by other competitors mentioned in the related work section.

The algorithms are implemented and coded using Octave, a high level language used for numerical interpretations that provides capabilities for the numerical solution of linear and nonlinear problems.

**4 Results**

We implemented four algorithms to predict the progression of HIV virus in the given datasets. After implementing all four algorithms, we ran those on the given training and test data set and got the following results:

Linear Regression with one variable provided a training and test accuracy of 63.83 and 41.84 respectively. Using Linear Regression with multiple variables we obtained a training and test accuracy of 79.42 and 41.84 respectively. Running Logistic Regression on the data set provided us a training and test accuracy of 79.02 and 58 respectively. Finally, running SVM algorithm with Gaussian Kernel provided a training and test accuracy of 98.7 and 98.55 respectively.

We were able to come up with a good technique to answer our research question. SVM outperformed all the algorithms under comparison and also the existing techniques (mentioned in the related work section) proposed by our competitors. The reason we implemented four algorithms is that we wanted to figure out which Machine Learning Algorithms was the best suited algorithm for predicting HIV progression. The question that we are trying to answer is very critical and inaccurate information while trying to predict the severity of the infection can be misleading. We have proved in our project that SVM algorithm provides an accuracy of approximately 99%.

**5 Summary and Conclusions**

Through a series of hypothesis formation and experimentation, we were able to identify key elements that determined the progression of the HIV virus. Of all the four algorithms that we have implemented, it is significant to note that SVM outperforms all the algorithms implemented with better training and test accuracy. Hence, we conclude that SVM algorithm with Gaussian Kernel is the best suited algorithm to predict the likelihood that the patient’s infection will become less severe.

One shortcoming that we encountered was the limited amount of data. The training data was skewed so that only around one-fifth of the training examples had positive classifications. In the future, we would like to gather more data to strengthen our predictions. Also, we would like to use PR and RT sequence as features for the algorithms after sequencing those features using Smith Waterman algorithm to predict the likelihood that the patient’s infection will become less severe.

**6 References**

[1] [Online] Available: http://en.wikipedia.org/wiki/HIV

[2] <http://www.kaggle.com/>

[3] D. Chou, R. Iu, R. Krishna, and A. Liang, “An Analysis on the Prediction of HIV Progression,” pp. 1–9, 2012.

[4] Y. Singh, N. Narsai, and M. Mars, “Applying machine learning to predict patient-specific current CD 4 cell count in order to determine the progression of human immunodeficiency virus ( HIV ) infection,” vol. 12, no. 23, pp. 3724–3730, 2013.

[5] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer, “Multi-task learning for HIV therapy screening,” *Proc. 25th Int. Conf. Mach. Learn. - ICML  ’08*, vol. 1, pp. 56–63, 2008.

[6] Y. Singh and M. Mars, “Support vector machines to forecast changes in CD 4 count of HIV-1 positive patients,” vol. 5, no. 17, pp. 2384–2390, 2010.