Deeper Understanding about Attributes of HIV Employing Support Vector Machine

Cheolho Heo and Taeseon Yoon

***Abstract*—Unlike direct treatment in the past, nowadays, data mining of information of diseases is very useful to cure patients. Also, with prediction of DNA sequence of specific illnesses, lots of people can avoid them. Bioinformatics, study of union of life science, biology and informatics, becomes one of the most important subjects to the future medical industry. A number of scientists and engineers have developed this area and as a result, various methodologies in aligning DNA sequences such as hidden markov model, artificial neural networks and support vector machines were developed during the last few decades. Especially, Support Vector Machine(SVM) is used in Supervised Learning, finding the furthermost hyperplane that separates given data. Unlike other methods, we can get more sophisticated and accurate results with learning method. Because of using SVM that has little parameters, we can also simplify the complex pattern and it is so effective in data analysis that we can easily investigate elements which have an effect on results. Moreover, to improve exactitude our study, we search and use DNA sequence data about HIV from NCBI(National Center for Biotechnology Information), which have reliable and numerous data.**

***Index Terms*—Human immunodeficiency virus (HIV), support vector machine (SVM), DNA sequence.**

I. Introduction

Bioinformatics is a study of amalgamative field involving both computational and biological sciences. As a science newly becomes evident, bioinformatics opens up new possibilities for molecular sciences and medicine, as it concerns not only sequence analysis but also genome annotation, gene expression, and prediction of biological activity [1]. With recent completion of human genome project and increasing various hereditary information, people have to manage more various data and the significance of bioinformatics becomes much higher than before. For one example of utilizing of this study, bio-companies and pharmaceutical companies can make and improve new drugs with very short time to investigate new medical substances because bioinformatics is a fundamental component of developing bio-industry. Anyway, in order to achieve more correct and reliable outcome with effect, we process our research by using Support Vector Machine (SVM). Comparing MDA, Logit, CBR with our methods, SVM shows us the most excellent accuracy of forecast [2]. This method not only has similar level of accuracy with Artificial Neural Network (ANN), other kind of bioinformatics methods, but also surmounts weak points of ANN such as local optimization. To use SVM, we can get lots of benefits to progress our research. First, it is very easy to interpret results because SVM is based on obvious theoretical grounds [3]. Furthermore, we can accomplish distinguish-learning quickly with small amount of learning data. Fig. 1 is one example of linear-SVM-scatterplot.

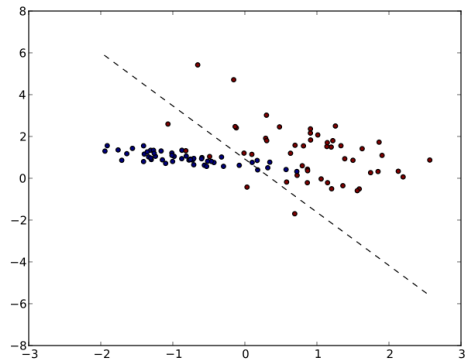


Fig. 1. Linear-SVM-scatterplot.

There are four functions of SVM used into our research, Poly1, Poly2, Normal and RBF. In our research, we skip the explanation of the normal kernel. In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, which represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models [4]. Poly1 is a function of SVM related with linear function and Poly2 is a function related with non-linear function [5]. Next, the (Gaussian) radial basis function kernel, or RBF kernel, is a popular kernel function used in support vector machine classification. Because support vector machines and other models employing the kernel trick do not scale well to large numbers of training samples or large numbers of features in the input space, several approximations to the RBF kernel (and similar kernels) have been devised. Fig. 2 is a picture that means an aspect of Polynomial function of SVM and Fig. 3 that means an aspect of RBF function of SVM.

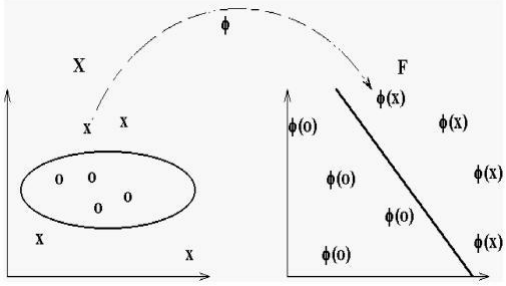


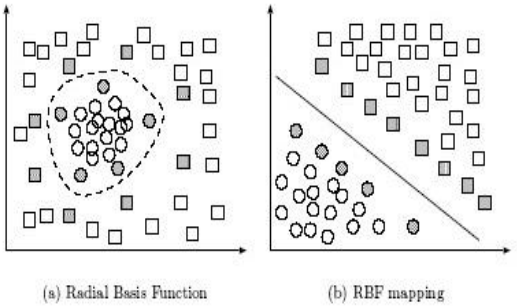
Fig. 2. Polynomial function of SVM.

The target of our experiment is Human Immunodeficiency

Manscript received February 28, 2014; revised April 28, 2014.

The authors are with the Hankuk Academic of Foreign Studies, Yongin, Korea (e-mail: dydakchry@naver.com, tsyoon@hafs.hs.kr).

Virus. Human Immunodeficiency Virus (HIV) is the source virus of Acquired Immune Deficiency Syndrome (AIDS) [6], [7]. After infecting the HIV, T-lymphocytes will be destroyed. This T-lymphocytes is a cell having a role on immunity, so if HIV came into our body, level of immunity of us will decrease quickly [8], [9]. Moreover, this virus arouses various infectious diseases and tumors and serious cases, causes people to die. There are HIV-1 and HIV-2 in types of HIV. The most dangerous and widespread HIV is HIV-1 [10]. Otherwise, HIV-2 is spread in some areas of Africa. HIV-1 has a number of subtypes [11]. The diversity of these subtypes is increasing because of the continuous hereditary transform [12]. This is the key point why we process our research with HIV. Because of various hereditary transform of HIV, people have difficulty in making antiviral agents of HIV (such as ribavirin, interferon-alpha and so on). Ref. [13] also, anti-HIV drugs have many side effects. If these drugs can affect to our body to cure, infectees will have to take drugs forever (If they stop, HIV will proliferate again and it can arouses severe infectious diseases and tumors) [14], [15]. To avoid this hard process and predict correct cause, our research result can give you significant and crucial information. Fig. 4 is a picture of HIV containing its structure and Genome.



Separable classification with Radial Basis kernel functions in different space. Left: original space. Right: feature space.

Fig. 3. RBF function of SVM.

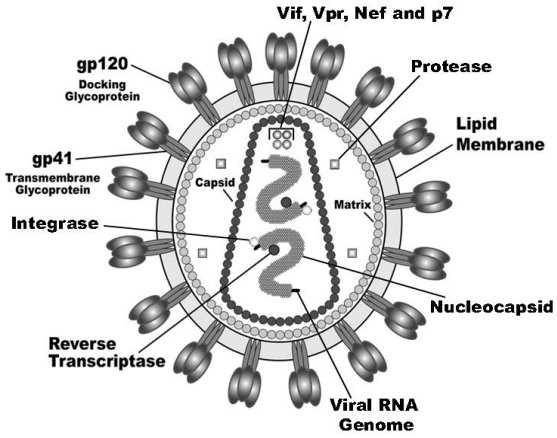


Fig. 4. Structure and genome of HIV.

II. Materials

To get precise result, we have to use proven data about HIV. So we use data from an article about HIV infection: Impact of HIV Infection on the Recurrence of Tuberculosis in South India (by Sujatha Narayanan, Soumya Swaminathan, Philip Supply and etc.). In this article, there are total of 239 data. Using these data, we adjusted them and made 60 data. Underlying table is some examples of these data.

TABLE I: Examples of Adjusted HIV Data

1	T	Q	I	M	F	E	T	F	1
2	G	Q	V	N	Y	E	E	F	1
3	P	F	I	F	E	E	E	P	1
4	S	F	N	F	P	Q	I	T	1
5	D	T	V	L	E	E	M	S	1
6	A	R	V	L	A	E	A	M	1
7	A	E	E	L	A	E	I	F	1
8	S	L	N	L	R	E	T	N	1
9	A	T	I	M	M	Q	R	G	1
10	A	E	C	F	R	I	F	D	1
11	D	Q	I	L	I	E	I	C	1
12	D	D	L	F	F	E	A	D	1
13	Y	E	E	F	V	Q	M	M	1
14	P	I	V	G	A	E	T	F	1
15	T	L	N	F	P	I	S	P	1
16	R	E	A	F	R	V	F	D	1
17	A	E	T	F	Y	V	D	K	1
L18	A	Q	T	F	Y	V	N	L	1
19	P	T	L	L	T	E	A	P	1
20	S	F	I	G	M	E	S	A	1
21	D	A	I	N	T	E	F	K	1
22	Q	I	T	L	W	Q	R	P	1
23	E	L	E	F	P	E	G	G	1
24	S	Q	N	Y	P	I	V	Q	1
					…				

III. Experiments

A. Methods

We progress our research applying 10-fold cross validation and balance of the data. Afterward, we make and arrange tables that represent outcomes of experiment and compare each function (Normal, Poly^1, Poly^2 and RBF) respectively. Finally, to get some general values, we get average of ten experiments.

```
Optimization finished (3540 misclassified, maxdiff=0.00097).
Runtime in cpu-seconds: 1.07
Number of SV: 7845 (including 7832 at upper bound)
L1 loss: loss=7832.93035
Norm of weight vector: |w|=0.09988
Norm of longest example vector: |x|=39.91240
Estimated VCdim of classifier: VCdim<=16.60297
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=78.41% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>21.59% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>100.00% (rho=1.00,depth=0)
Number of kernel evaluations: 398340
Writing model file...done
[root@eng SVM]# ./svm_classify hiv/hiv_test1.txt hiv/normal1 hiv/normal1_prediction
Reading model...OK. (200 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 60.00% (18 correct, 12 incorrect, 30 total)
Precision/recall on test set: 57.14%/80.00%
```

Fig. 5. Result of test 1 on normal function.

We set 114 data to hiv_0, 248 data to hiv_0, 100 data to train and 15 data to test. With repetition of experiments, we are able to get more accurate and precise outcomes. Underlying captured pictures are the situation results of SVM experiment test 1 on editplus3. We can notice here a kind of functions (Normal, Poly1, Poly2 and RBF) and process of test about those functions. Ultimately, what we have to identify is accuracy of test set, implying properties of HIV virus. We will check these and find novel features of HIV. Underlying figures are result of test 1 on each function.

```
Optimization finished (72 misclassified, maxdiff=0.00058).
Runtime in cpu-seconds: 0.02
Number of SV: 174 (including 163 at upper bound)
L1 loss: loss=166.75621
Norm of weight vector: |w|=0.13387
Norm of longest example vector: |x|=47.50789
Estimated VCdim of classifier: VCdim<=41.43277
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=85.50% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>14.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>14.14% (rho=1.00,depth=0)
Number of kernel evaluations: 34053
Writing model file...done
[root@eng SVM]# ./svm_classify hiv/hiv_test1.txt hiv/pol1 hiv/poly1_prediction
Reading model...hiv/pol1: No such file or directory
[root@eng SVM]# ./svm_classify hiv/hiv_test1.txt hiv/poly1 hiv/poly1_prediction
Reading model...OK. (174 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 66.67% (20 correct, 10 incorrect, 30 total)
Precision/recall on test set: 64.71%/73.33%
```

Fig. 6. Result of test 1 on Poly1 function.

```
Optimization finished (49 misclassified, maxdiff=0.00000).
Runtime in cpu-seconds: 0.01
Number of SV: 200 (including 200 at upper bound)
L1 loss: loss=196.80913
Norm of weight vector: |w|=0.17863
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=1.06382
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=47.50% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>5.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>100.00% (rho=1.00,depth=0)
Number of kernel evaluations: 24591
Writing model file...done
[root@eng SVM]# ./svm_classify hiv/hiv_test1.txt hiv/RBF1 hiv/RBF1_prediction
Reading model...OK. (200 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 66.67% (20 correct, 10 incorrect, 30 total)
Precision/recall on test set: 100.00%/33.33%
```

Fig. 7. Result of test 1 on RBF function.

```
poly2

Checking optimality of inactive variables...done.
 Number of inactive variables = 100
done. (983944 iterations)
Optimization finished (35 misclassified, maxdiff=0.00099).
Runtime in cpu-seconds: 24.75
Number of SV: 125 (including 75 at upper bound)
L1 loss: loss=98.81086
Norm of weight vector: |w|=0.18481
Norm of longest example vector: |x|=2257.00000
Estimated VCdim of classifier: VCdim<=173995.13631
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=62.50% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>38.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>37.62% (rho=1.00,depth=0)
Number of kernel evaluations: 35444442
Writing model file...done
[root@eng SVM]# ./svm_classify hiv/hiv_test1.txt hiv/poly1_2 hiv/poly1_2_prediction
Reading model...OK. (125 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 73.33% (22 correct, 8 incorrect, 30 total)
Precision/recall on test set: 73.33%/73.33%
```

Fig. 8. Result of test 1 on Poly2 function.

B. Results

In each experiment, there are some gaps among each set. In one example, the gap of experiment between Set7 and Set 8 of RBF is 20%. Including this, gaps can make lots of errors to the result. However, thanks to our repetitive test, we can minimize such errors.

From these tables, we can find that accuracy of RBF (Accuracy: 75.33%) is the top of four functions, following Poly2 (63.67%), Poly1 (53.33%) and Normal (51.67%). Specially, noticing the outcomes of RBF and Poly2, we can predict that HIV has non-linear attributes.

TABLE II: THE ACCURACY ON TEST SET (1~5)

Function	Set1	Set 2	Set 3	Set 4	Set 5
Normal(%)	60.00	56.67	50.00	60.00	46.67
Poly1(%)	66.67	56.67	63.33	63.33	56.67
Poly2(%)	73.33	60.00	60.00	70.00	70.00
RBF(%)	66.67	86.67	83.33	80.00	66.67

TABLE III: THE ACCURACY ON TEST SET (6~7)

Function	Set 6	Set 7	Set 8	Set 9	Set 10
Normal(%)	50.00	50.00	60.00	36.67	46.67
Poly1(%)	43.33	30.00	50.00	43.33	60.00
Poly2(%)	66.67	53.33	60.00	70.00	53.33
RBF(%)	83.33	63.33	83.33	66.67	73.33

TABLE IV: THE ACCURACY ON TEST SET (AVERAGE)

Function	Average
Normal(%)	51.67
Poly1(%)	53.33
Poly2(%)	63.67
RBF(%)	75.33

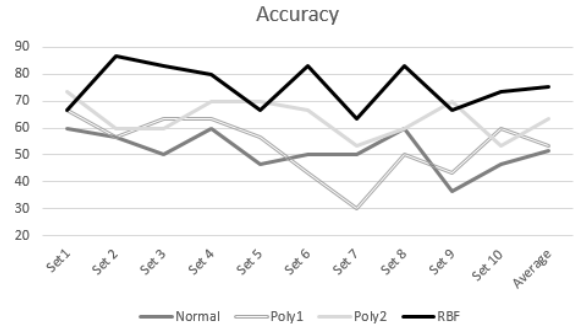


Fig. 9. Graph of experiment outcomes.

IV. CONCLUSION

In the past, lots of studies reported that the features of cleavage and non-cleavage of HIV are linear. However, through our experiment and outcome, we find that they are also having non-linear property by tendency of RBF and Poly2 functions. This conclusion from our research, unearthing new property of HIV virus, will be able to contribute for human health.

REFERENCES

[1] A. Cunningham, H. Donaghy, A. Harman, M. Kim, and S. Turville, "Manipulation of dendritic cell function by viruses," *Current Opinion in Microbiology*, vol. 13, no. 4, pp. 524–529, 2010.

[2] C. Corinna and V. N. Vladimir, "Support-vector networks," *Machine Learning*, vol. 20, 1995.

[3] C. M. Ferris and T. S. Munson, "Interior-point methods for massive support vector machines," *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 783–804, 2002.

[4] L. S. Algebra, *Graduate Texts in Mathematics 211*, New York: Springer-Verlag, ISBN 978-0-387-95385-4, MR 1878556, 2002.

[5] S. Sirayanone and R. L. Hardy, "The Multiquadric-biharmonic method as used for mineral resources, meteorological, and other applications," *Journal of Applied Sciences and Computations*, vol. 1, pp. 437–475, 1995.

[6] S. Narayanan, S. Swaminathan, P. Supply, S. Shanmugam *et al*., "Impact of HIV infection on the recurrence of tuberculosis in South India," *Glynn*, pp. 704–711, March 2010.

[7] D. M. Buhmann, *Radial basis functions: theory and implementations*, Cambridge University Press, ISBN 978-0-521-63338-3, 2003.
[8] D. T. Jamison, J. G. Breman, A. R. Measham *et al.*, *Disease Control Priorities in Developing Countries*, 2nd ed., Washington (DC): World Bank, 2006.
[9] J. M. Coffin, S. H. Hughes, H. E. Varmus, *Retroviruses*, Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 1997.
[10] O. Öztürk, A. Aksaç, A. Elsheikh, T. Özyer, and R. Alhajj, "A consistency-based feature selection method allied with linear SVMs for HIV-1 protease cleavage site prediction," *PLoS One*, August 2013.
[11] C. R. Cohen, J. R. Lingappa, J. M. Baeten, M. O. Ngayo *et al.*, "Bacterial vaginosis associated with increased risk of female-to-male HIV-1 transmission: A prospective cohort analysis among african couples," *PLoS Med.*, June 2012.
[12] R. A. Weiss, "How does HIV cause AIDS?" *Science* 260 (5112): 1273–9, Bibcode: 1993Sci...260.1273W, May 1993.
[13] S. Hempel, K. D. Shetty, P. G. Shekelle *et al.*, *Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations*, Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Sep.
[14] D. C. Douek, M. Roederer, and R. A. Koup, "Emerging Concepts in the Immunopathogenesis of AIDS," *Annu. Rev. Med.*, vol. 60, pp. 471–84, 2009.
[15] B. Manuel *et al*., *Solving Polynomial Equations: Foundations, Algorithms, and Applications*, Springer, ISBN 9783540273578, 2006.

Cheolho Heo was born in 1997. He is currently a student in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in chemistry and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.

Taeseon Yoon was born in 1972. He is currently a teacher in Science Major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in engineering and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.