Analysis of Hantaviruses Glycoprotein Sequence Using SVM Algorithm

Yusin Kim, Youngmin Ko, Daniel P. Jeong, and Taeseon Yoon

***Abstract*—Hantaviruses are single-stranded, enveloped, negative sense RNA viruses of the Bunyaviridae family. They induce deadly hemorrhagic fever with fatality up to 40%. Currently, there is no specific cure for Hantaviruses, so more researches about this mortal virus should be done. In order to effectively analyze a variety of different Hantaviruses, we utilize a model called the support vector machine (also known as SVM) which is generally used for analyzing and classifying binary data.The basic mechanism of the SVM is to find the most optimal hyperplane, or the maximum-margin hyperplane, which can separate different types of data with the least error bound. Out of all of the hyperplanes that may be used to classify the data points, the most optimal hyperplane is the one that has the largest margin, or separation among different types of data. In other words, the optimal hyperplane is chosen in the case where the distance between the nearest points of each group of data is maximized. Ultimately, using the selected hyperplane, SVM classifies the data points and computes values such as accuracy and sensitivity. At the end of its operation, the SVM algorithm prints out the computed values. Such SVM algorithms can be used to learn the characteristics of each Hantavirus such as sequence patterns and abundance of amino acids. Since we are the first time to scientifically investigate the Hantavirus with SVM, it is expected that the results of this research will be greatly helpful for further in-depth researching and the development of the cure for the virus.**

***Index Terms*—Accuracy, glycoprotein sequence, Hantavirus, SVM.**

I. Introduction

A. Hantavirus

Since the first Hantavirus, Hantaan virus (HTNV), 49 Hantaviruses have been identified and at least 22 of them are pathogenic. Hantavirusescan induce Hemorrhagic Fever with Renal Syndrome (HFRS) and Hantavirus Pulmonary Syndrome (HPS).

Hemorrhagic Fever with Renal Syndrome (HFRS) can be mild, moderate, or severe depending on the causative virus. HFRS caused by Hantaan virus (HTNV) and Seoul virus (SEOV) is characterized by its clinical features in various systems, such as neurological, gastrointestinal, and cardiovascular systems with fatality rate up to 12 %. NE is a mild form of HFRS, caused by the Puumala virus (PUUV) with fatality rate of 0.1% up to 0.4% [1].

HPS is more lethal than HFRS and it is also the main cause of death of those resulting from Hantavirus infection. It induces myocardial dysfunction and hypoperfusion with average fatality rate of 40%. Some of the main symptoms of HPS include tachypnea, tachycardia, postural hypotension, and visceral hemorrhage [2]. Meanwhile, visceral hemorrhage, which severely damages internal organs, is the deadliest out of the four listed above. Unfortunately, there are no existing cures for HFRS and HPS.

Various rodents are carriers of Hantaviruses. Each rodent subfamily carries diverse viruses; some of them are pathogenic to humans, while others are not. Some Hantaviruses were detected in pigs, bats, cats, birds, and dogs, and it is not clear whether these species are parts of the infected population or spillover hosts [3]. Hantaviruses can be transmitted by contacting with excreta, saliva, urine, and feces of infected animals. The only case of interpersonal infection of Hantaviruses was detected during the Andes virus HPS outbreak occurred in Argentina.

Recently, the Hantavirus received global attention when it was identified to be the etiologic agent of HPS outbreak in the Four Corners region of the United States in 1993 [4]. Every year, approximately 150,000 to 200,000 patients are hospitalized around the world due to Hantavirus related diseases [5].

In this research, we analyzed the features of 4 different Hantaviruses: Seoul virus (SEOV), Sin Nombrevirus (SNV), Puumalavirus (PUUV), and Prospect Hill virus (nonpathogenic). The Seoul virus is an urban type virus while HTNV is a rural type virus; regardless of the differences, both are deadly to human. The hosts of SNV are New World rats and mice (Sigmodontinae) and those of SEOV are Old World rodents group (Murinae) [3]. PUUV is less severe than the SEOV, SNV, and HTNV, while the Prospect Hill virus is nonpathogenic.

B. SVM

SVM, a very powerful algorithm used for prediction, classification, and regression problems, is a kernel-based margin classifier, which was utilizes both for statistics and optimization. As previously mentioned, SVM draws an optimal hyper plane in a high dimensional feature space that defines a boundary which maximizes the margin between the hyperplane and the nearest data samples from each of the two classes [6]. Following this procedure, SVM gives a better generalization property. The SVM solves the following optimization problems:

$$arg\ min\{\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i\}$$

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where **w** is a normal vector perpendicular to the hyperplane and $\boldsymbol{\xi_i}$ stands for slake variables which allow misclassifications. **C** is a penalty parameter which balances the trade-off between the margin and the training error.

Manuscript received January 31, 2014; revised April 17, 2014.

Yusin Kim is with the Hankuk Academy of Foreign Studies, Korea (e-mail: ushin612@naver.com).

C. SVM-Multi Class

Multiclass SVM is used to evaluate instances which are drawn from various sets of several elements by using support vector machines. Similar to SVM-light, SVM-multiclass also consists of a learning module (svm_multiclass_learn) and a classification module (svm_multiclass_classify). Classification module can be used for the application of an already-learned model.

The usage of the SVM-multiclass is as simple as SVM-light: Using the command prompt, call "svm_multiclass_learn -c 1.0 example_filemodel_file" to make an output "model_file" with the training set "example_file" using the regularization parameter 'c' that is set to 1.0.

SVM-multiclass solves the following optimization problems:

$$\text{min} \quad 1/2 \quad \Sigma_{i=1..k} w_i \times w_i + C/n\Sigma_i = {}_{1..n}\xi_i$$

$$s.t. \text{ for all } y \text{ in } [1..k]: [\, x_1 \bullet w_{yi} \,] >= [\, x_1 \bullet w_y \,] + 100 \times \Delta(y_1, y) - \xi_1$$

...

$$\text{For all } y \text{ in } [1...k]: [\, x_n \bullet w_{yn} \,] >= [\, x_n \bullet w_y \,] + 100 \times \Delta(y_n, y) - \xi_n$$

where C is the regularization parameter that trades off margin size and training error, and $\Delta(yn, y)$ is the loss function that returns 0 if yn equals y, and 1 otherwise [7].

II. Materials and Methods

A. Hantavirus Glycoprotein Sequence

Hantavirus glycoprotein sequence data were obtained from national center for biotechnology information (ncbi) protein database.

Seoulvirus (1133aa)http://www.ncbi.nlm.nih.gov/protein/AAD31904.1
Sin Nombre virus (1140aa)
http://www.ncbi.nlm.nih.gov/protein/AFV71283.1
Puumala Virus (1148aa)
http://www.ncbi.nlm.nih.gov/protein/AFQ60653.1
Prospect Hill Virus (1142aa)
http://www.ncbi.nlm.nih.gov/protein/CAA38922.

B. Evaluating

Hantaan River virus, Sin Nombre Virus, Seoul Virus, Prospect Hill Virus were the data sets used for evaluating. We classified each of the viruses with dividing them into 5 window, 7 window and 9 window (Each of them indicates 5 sequences, 7 sequences and 9 sequences in one piece). We evaluate 1) Number of Support Vectors, 2)Final Epsilon of KKT conditions, 3) Zero/one-error on test set, and 4) Accuracy for SVM-(single) evaluation. Number of Support Vectors determines the shape of data, as well as the parameter value.

KKT(Karush-Kuhn-Tucker) conditions are the first order necessary conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied. We ran the experiment by using 10 fold cross validation.

III. Result

We used numbers to express the viruses. Each number indicates a different type of viruses: 1. Sin Nombre virus 2. Seoul virus 3. Puumala virus 4. Prospect Hill virus. Each figure below shows the values of analogy which includes accuracy and precision value. We used 5 criterion for evaluation: 1. Normal 2. Polynomial 3. RBF 4. Sigmoid 5. Polynomial with 2 as its index value (A polynomial with exponent degree of 2).

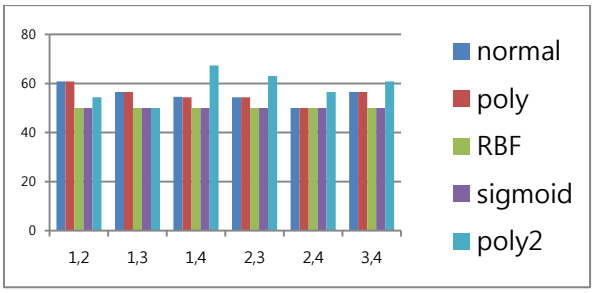

Fig. 1. Accuracy values for the analogy: 5 window.

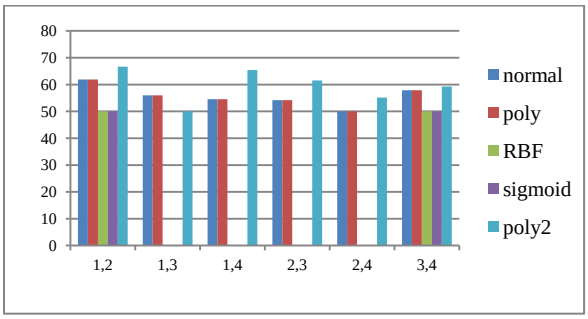

Fig. 2. Precision values for the analogy: 5 window.

Fig. 1 shows the accuracy for the analogies between each virus in 5 window experiment. Accuracy values for every analogy are slightly low, which means that 5 window is not appropriate for making accurate organization

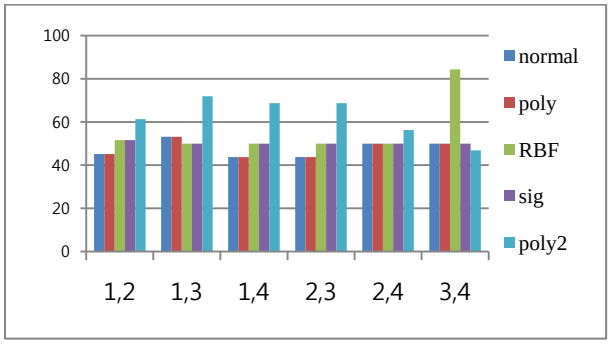

Fig. 3. Accuracy values for the anlalogy: 7 window.

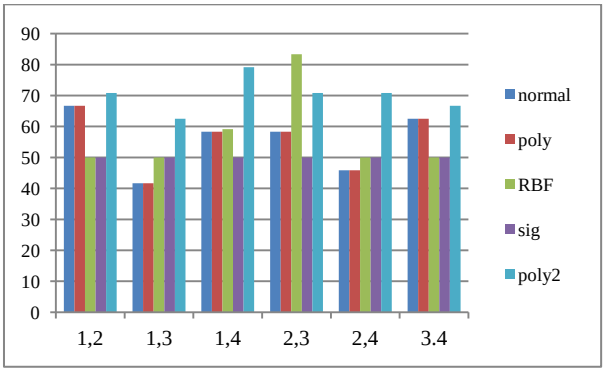

Fig. 4. Precision values for the analogy: 7 window.

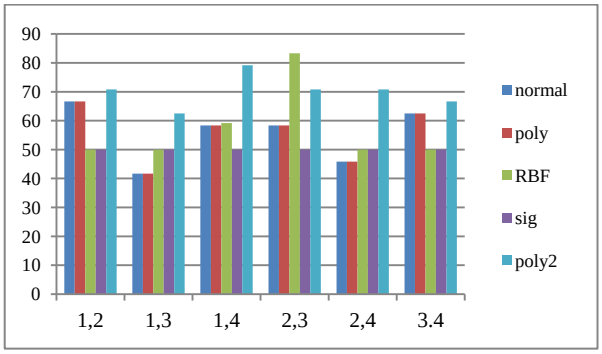

Fig. 5. Accuracy values for the analogy: 9 window.

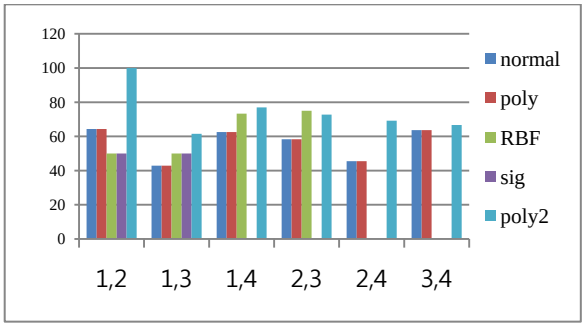

Fig. 6. Precision values for the analogy: 9 window.

Fig. 3 - Fig. 6 indicate the accuracy and precision values for the comparison between each virus in 7 window and 9 window experiment. Accuracy value was high in 9 window experiment because the quantity of information of sequence was adequate in 9 window experiment.

Accuracy value up to 80% shows that the samples can be organized. The comparison between Seoul virus and Puumala virus (2, 3) in the 9 window experiment shows high accuracy in RBF and polynomial 2, which indicates that the two viruses can be separated using nonlinear equation.

TABLE I: 5 WINDOW, NORMAL 10 FOLD CROSS VALIDATION

	1	2	3	4	5
SV	14	14	15`	14	13
KKT	0.0579	0.09203	0.07238	0.0375	0.07713
Zero/one error	74.21%	80.43%	72.83%	69.57%	73.91%

6	7	8	9	10
13	16	14	12	12
0.03279	0.05584	0.02986	0.091	0.02941
65.22%	80.43%	68.48%	77.17%	77.17%

TABLE II: 5 WINDOW, POLYNOMIAL 10 FOLD CROSS VALIDATION

	1	2	3	4	5
SV	5	5	5	5	5
KKT	0.03251	0.04752	0.05196	0.06054	0.03839
Zero/one error	78.12%	75%	79.69%	71.88%	79.69%

6	7	8	9	10
5	4	5	5	5
0.09217	0.11848	0.05404	0.05831	0.0685
78.12%	75.00%	71.88%	70.31%	75.00%

Table I and Table II indicate the number of SV and KKT value and Zero/one error on test sets in 5window experiment.

TABLE III: 9 WINDOW, NORMAL 10 FOLD CROSS VALIDATION

	1	2	3	4	5
SV	16	18	17	17	22
KKT	0.09963	0.06581	0.08703	0.09874	0.03908
Zero/one error	83.33%	70.83%	70.83%	75.00%	77.08%

6	7	8	9	10
17	21	25	18	26
0.08487	0.07264	0.05506	0.04666	0.0795
81.25%	75.00%	70.83%	75.00%	77.08%

TABLE IV: 9 WINDOW, POLYNOMIAL10 FOLD CROSS VALIDATION

	1	2	3	4	5
SV	7	4	5	4	3
KKT	0.05797	0.06642	0.04593	0.08893	0.25491
Zero/one error	0.25491	79.17%	75.00%	75.00%	75.00%

6	7	8	9	10
6	4	3	5	
0.02215	0.03071	0.34913	0.06082	
72.92%	83.33%	75.00%	79.17%	

Table III and Table IV are the number of SV and KKT value and Zero/one error on test sets in 9 window experiment.

IV. DISCUSSION

A high accuracy value obtained from the SVM signifies that the data is relatively easy to classify (the higher, the easier). In our case, the accuracy value was relatively low when we used linear classification to separate the data; in other words, it showed that it is relatively difficult to classify the given data. Based on such a result, we derived two different hypotheses: 1) Data with linear characteristics are not well-classifiable. 2) Data that is not well-classifiable by the linear classification method has nonlinear characteristics.

To check the second hypothesis, we increased the exponent to 2 and attempted classifying the data using nonlinear classification. With an average accuracy level close to 70%, the result showed that the data does in fact have nonlinear characteristics, explaining why they were not well-classifiable. However, the relatively high average does not mean that all parts of the data are well-classifiable with nonlinear classification; in fact, there were some parts of the data that had accuracy values of percentages in the mid 60's.

Moreover, we divided the each type of virus into three different sequence lengths of 5 window, 7 window, and 9window for analyses. We mainly focused on the values obtained through the classification of 9 window sequences because they resulted in higher resultant values. 5 window and 7 window sequences have shorter lengths; thus, they contain less information in comparison with the 9window sequences, which explains why using 9 window sequences return more accurate results.

The support vector machine is one of the most effective and efficient data analyzing algorithms. When comparing two different classes using the SVM, the fact that two different classes of data are not classified. Using this algorithm allows us to conclude that the two classes are identical. However, as more effective data classification methods are developed, it may become possible to classify the two.

When we analyzed the Seoul virus and Prospect Hill virus in 5 window and 7 window sequences, both linear and nonlinear classifications resulted in a very low accuracy value of about 50%, whereas it was easy to classify them when we used 9 window sequences for both of them. We suppose that there is a certain common pattern between the Seoul virus and the Prospect Hill virus, and we are planning to conduct further in-depth researches to find it in the future.

REFERENCES

[1] B. Settergren, "Clinical aspects of nephropathiaepidemica (Puumala virus infection) in Europe: A review," *Scand J Infect Dis*., vol. 32, pp. 125-132, 2000.

[2] S. M. Raboni, G. Rubio *et al*., "Clinical survey of hantavirus in southern Brazil and the development of specific molecular diagnosis tools," *Am J Trop Med Hyg*., vol. 72, pp. 800-804, 2005.

[3] B. Zhenqiang, B. H. F. Pierre, and C. E. Roth, "Hantavirus infection: A review and global update," *J Infect Developing Countries*, vol. 2, no. 1, pp. 3-23, 2008.

[4] S. T. Nichol, C. F. Spiropoulou, S. Morzunov *et al*., "Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness," *Science*, vol. 262, pp. 914-917, 1993.

[5] J. A. Lednicky, "Hantavirus: A short review," *Arch Pathol Lab Med*., vol. 127, pp. 30-35, 2003.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*, Cambridge University Press, 2000.

[7] T. Joachims, "SVM-Multiclass: Multi-class support vector machine," Cornell University, Department of Computer Science, 2008.

Yusin Kim was born in Daegu, Korea. He now lives in Yongin, Korea. Since March 2013, he has been with the Hankuk Academy of Foreign Studies majoring in natural science.

Youngmin Ko was born in Seoul, Korea. He now lives in Yongin-si, Korea. He has been attending HAFS, Hankuk Academy of Foreign Studies, from 2013. He belongs to an international course in his academy.

Daniel P. Jeong is with Hankuk Academy of Foreign Studies, Korea and has been involved in the work related to computer science. His current research interests include computer algorithms, multimedia applications, real-time operating systems, mobile communications, and network security.

Taeseon Yoon was born in Seoul, Korea, in 1972. He was a Ph.D. candidate degree in computer education from the Korea University, Seoul, Korea, in 2003. From 1998 to 2003, he was with EJB analyst and SCJP. From 2003 to 2004, he joined the Department of Computer Education, University of Korea, as a lecturer in Ansan University, and as an adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a computer science and statistics teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.