A Comparison of H1N1 and H3N2 Viruses Using Decision Tree and Apriori Algorithm

Seongpil Jang, Hunseok Choi, Yeonje Jung, Eoram Moon, and Taeseon Yoon

***Abstract*—H1N1 virus and H3N2 virus was once a widespread epidemic in South Korea. The former was spread through human while the latter was spread via dogs. Those two viruses were widely spread in Korea in the same time. Also, those two viruses gave Koreans significant shock. As there wasn't suitable vaccine or medication, Koreans have to be in fear about epidemic viruses. As both of two viruses was spread in Korea same period, and had many patients, we thought that H1N1 virus and H3N2 virus might have lots of common traits, so we started to analyze the viruses. To compare characteristics of two viruses, we concerned several algorithms. To find general rules from characteristic-unknown viruses, we think apriori algorithm is proper rather than Support Vector Machine. After experiment with apriori algorithm, we decided to find proper rules using tree structure. So, we used Decision Tree to extract specific rules of two viruses. Using a decision tree and apriori algorithm in our experiment, it was possible for us to compare the characteristics of H1N1 and H3N2 viruses.**

***Index Terms*—H1N1 virus, H3N2 virus, the apriori algorithm, decision tree.**

I. Introduction

In South Korea, there was an outbreak of influenza, the virus H1N1, 6 years ago. It is told that the year 2009 is the year of new epidemic as well as the period of no-school for kids. This indicates that the Korean government found about the seriousness of the disease late, leading to a hazardous effect among kids. After 2 years of that epidemic, when people were relieved to know that they don't have to be scared by H1N1 for now, a new virus started to threaten the lives of, not humans, but dogs. Even though it is not a lot, some animal hospitals and dog domesticating facilities showed that the dog influenza virus, H3N2 in this case, has made its arrival and did spread much more quicker in those hospitals and facilities, specifically. These two viruses(H1N1, H3N2) have many interesting facts when it comes to South Korea infections, since there is noaccurate suggestion or explanation that can explain the infectious route of its 'Korea Visit'. In this study, we will examine and research about the origin of these two viruses in Korea, then compare two things in order to find things in common and explain the unknown subject.(Some of the research results were passed down to us by former research team of our school, and we were glad to continue on research and be delegated.)We used decision tree and apriori, in order to compare the similarities among amino acid structures of respective Hemagglutinin and Neuraminidase of H1N1 and H3N2.

II. Procedure and Method

A. H1N1 Virus

H1N1, the subtype virus of Influenza A virus, is well known as the virus that caused Spain flu and 2009's human influenza. It is one of the most common types of virus that cause Influenza to human, and has mutated little between the epidemic of Spain flu and that of human influenza in 2009. As it has many various types of mutations, epidemics are coming every year. As it is very epidemic, people in 2009 was in great fear with the various mutations of H1N1.

B. H3N2 Virus

H3N2, the subtype virus of Influenza A virus, is a type of virus that causes season influenza and swine flu. This virus is also well known as the cause of dog influenza in Korea, and showed its maximum strength as a dog-threatener in 2011, shortly after the epidemic of human influenza by H1N1.

C. Decision Tree

Decision Tree is a data mining method which is commonly used for inductive inference, in order to create a model that predicts the value of a target variable based on several input variables. It poses a series of questions about the features associated with data items. Each questions contained in a node, and every internal node points to one child node for each possible answer to its question [1].

A tree is termed "learned" by dividing the source set into subsets which usually branches down recursively. This sort of learning is termed as recursive partitioning. This process of a top-down induction of decision trees is a part of "greedy algorithm", and is one of the most common strategies for learning decision trees [2].

In data mining, decision trees can be described as the combination of mathematical and computational form, using the process of categorization and generalization. Generally, the diagram gradually comes in records of the form: $(X, Y)=(x_1, x_2, x_3,\ldots x_k, u)$[3].

The dependent variable, Y, is the target variable that we aretrying to understand, classify or generalize. The vector x is composed of the input variables, x_1, x_2, x_3 *et al*, which are used for testing [4], [5].

D. Apriori Algorithm

Apriori algorithm is a classic algorithm that points out the frequently appearing data and repeating rule in the given database [6]. The Apriori Algorithm works in two steps: in a first step the *frequent itemsets*(often misleadingly called

Manuscript received February 3, 2015; revised January 14, 2016.

Seongpil Jang, Hunseok Choi, Yeonje Jung, Eoram Moon, and Taeseon Yoon are with the Department of Nature Science, Hankuk Academy of Foreign Studies, Gyeonggi, South Korea (e-mail: kiml11@daum.net, juyoung274@naver.com, tsyoon@hafs.hs.kr).

doi: 10.18178/ijmlc.2016.6.1.576

large itemsets) are determined. These are sets of items that have at least the given minimum support. In the second step association rules are generated from the frequent itemsets found in the first step [7]. This algorithm uses a "bottom up" approach where frequent subsets are prolonged one at a time, which is also generally named the candidate generation. Each group of candidates is tested repeatedly until no more extensions are found. Utilizing the breadth-first search and a Hash tree structure, this algorithm counts candidate items efficiently by proceeding a few steps. First, it generates candidate item sets of length k from item sets of length k-1, which make candidates contain an infrequent sub patterns. As a result, according to the downward closure lemma, the candidate set will contain all frequent k-length item sets. In the final step, it scans the transaction database to determine frequent item sets among the candidates [8].

E. Procedure

First, we investigated the genome sequences of the H1N1 virus and the H3N2 virus from the NCBI which is also called "National Center for Biotechnology Information". We extracted a large amount of data. For the decision tree, we experimented with sequence for a 10-fold cross validation, classifying proteins into 4classes (class1-H1 of H1N1, class2-H3 of H3N2, class3-N1of H1N1, class4-N2 of H3N2). Then, we extracted data that has frequency value over 0.75. For the apriori algorithm, we extracted experimental data, and made graphs using experimental data to compare the characteristics of 4 classes.

III. Results

A. Decision Tree

According to Table I, several rules with frequencies above 0.75 were found. Also, it can be told that position 4 is critical factor that differs Classes because position 4 has appeared most frequently. Position 1 and position 5 can also be important factor in that the positions differs H1N1 from H3N2.

TABLE I: Rules under 5 Window

Class	Rule	Frequency
Class 1	Pos 1=V Pos 4=G	0.8
Class 2	Pos 4=P Pos 5=N	0.8
Class 3	Pos 1=I Pos 4=W	0.75
Class 4	Pos 4=R Pos 5=T	0.75

TABLE II: Rules under 7 Window

Class	Rule	Frequency
Class 1	Pos 3=D Pos 6=T	0.8
Class 2	Pos 3=M	0.75
	Pos 3=E Pos 9=E	0.75
Class 3	Pos 1=N Pos 6=Y	0.75
Class 4	Pos 3=V Pos 6=T	0.75

According to Table II numeral rules with frequencies above 0.75 were found. That is, it can be told that position 4 is critical factor that differs classes because position 4 has appeared most frequently. Position 2 can also be important factor because H1N1 and H3N2 have differences with position 2.

According to Table III, we could find some rules with frequencies above 0.75. That is, it can be told that position 3 is critical factor that differs classes because position 3 has appeared most frequently. Position 6 can also be important factor because position 6 has been observed quite frequently.

TABLE III: Rules under 9 Window

Class	Rule	Frequency
Class 1	Pos3=DPos 4=K	0.8
Class 2	Pos2=T Pos 4=D	0.8
Class 3	Pos 4=EPos7=S	0.75
Class 4	Pos 2=FPos4=Q	0.75

B. Apriori Algorithm

Using Apriori algorithm, we compared frequencies of amino appearances for both H and N. We used 5, 7, and 9 Window for deeper analysis, and used graphs to compare and contrast some notable differences between two. For N, graphs of Window 5, 7 for H1N1 and H3N2 have relatively similar appearance in terms of numbers and types of Amino acids. There are, however, some notable differences we can find on window 9 model. Closely examined, H3N2 structure is a wider version of H1N1 structure, which means that while H3N2 has the basic structure of H1N1 in itself, it has some extras on its own. (See Fig. 1, you can see that Amino acids like P, T, and Y are making the differences.)

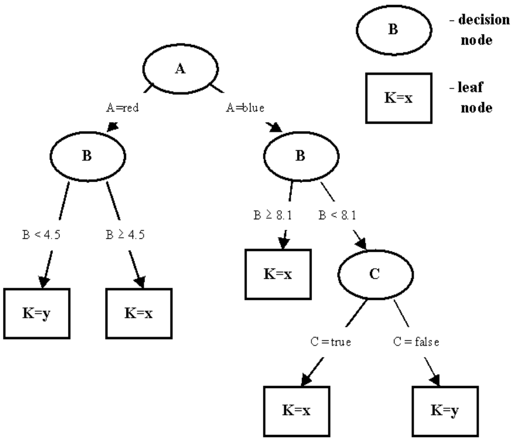

Fig. 1. Example of Decision tree Algorithm (http://dms.irb.hr/tutorial/tut_dtrees.php).

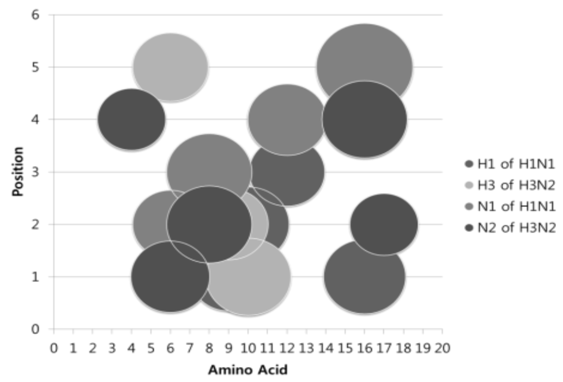

Fig. 2. Apriori under 5 Window.

For H, graphs of Window 5, 7, and 9 all showed great similarity. The numbers and the types of Amino acid of both H1N1 and H3N2 were very similar, with very few exceptions. Closely examined, opposed to the results of comparing N, H1N1 usually had more types of Amino acid(which is the only exceptions we found), but in terms of number, two viruses appeared almost the same. Overall, no clear distinctions were found for H.

We gave numbers to amino acids to make graphs more simply. (A-1, C-2, D-3, E-4, F-5, G-6, H-7, I-8, K-9, L-10, M-11, N-12, P-13, Q-14, R-15, S-16, T-17, V-18, W-19, Y-20).

In Fig. 2, we can see high similarity with those N proteins of two viruses. However, notifying that the area of H protein of H3N2 is more wider than that of H1N1, we can find H1N1 has more simple structure than that of H3N2.

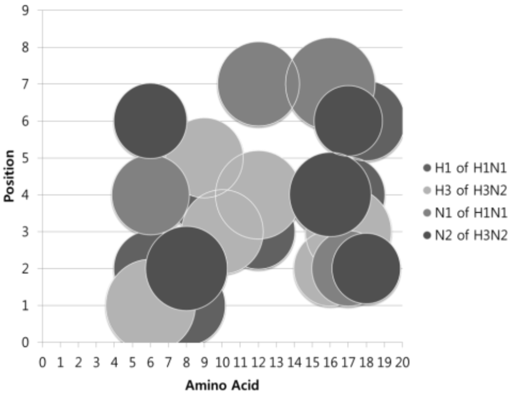

Fig. 3. Apriori under 7 Window.

In Fig 3, we can also see high similarity with those N proteins of two viruses. However, notifying that the area of H protein of H3N2 is more wider than that of H1N1, we can find H3N2 has more complex structure than that of H1N1.

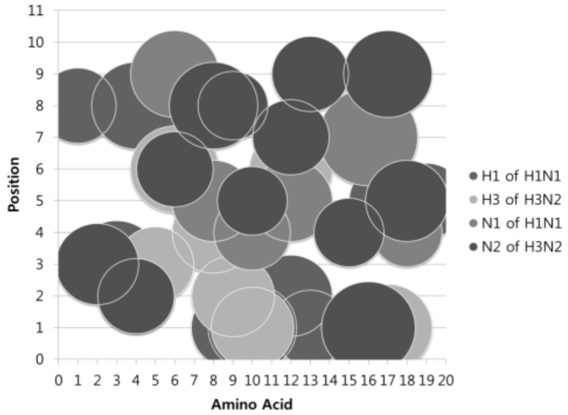

Fig. 4. Apriori under 9 Window.

In Fig. 4, we can see high similarity with those N proteins of two viruses. Also, in 9 window experiment, we can find some difference between H proteins of H3N2 and H1N1. This implies that there are quite differences among similar viruses such as H3N2 and H1N1.

These all figures suggest that it is hard to make vaccine of influenza viruses because influenza viruses have many mutations according to this result.

IV. Analysis and Discussion

With the decision tree algorithm, we could find several rules, choosing frequency values over 0.75. Through these results, we were able to catch numerous rules. This shows that H1N1 virus and H3N2 virus are able to be distinguished easily. Also it tells that vaccine for these flu viruses can't be developed easily. Our data supports the fact that it is hard for us to develop vaccine for flu because there are various differences. However, in this research, with the decision tree data, we found that position 4 in H1N1 virus and H3N2 virus may be a crucial factor in distinguishing flu viruses. Further experiments are expected to force scientists and doctors to find specific rules between H1N1 virus and H3N2 virus.

With the apriori algorithm, we found that Serine was most common amino acid between H1N1 and H3N2 viruses. It might imply that Serine is the general amino acid that is crucial to how influenza virus works and it's fatality. However, how Serine affects to human within flu viruses has not been experimentally discussed. Further experiments are needed to find mechanisms how Serine affects to human within flu viruses

V. Conclusion

Influenza viruses are mutate every day and have probability to cause widespread epidemic. Finding a vaccine for influenza virus, and creating a treatment for influenza virus has been a vital task. The data we extracted from the decision tree and Apriori algorithm further substantiated this phenomenon. We were able to find some similarities in the viruses, and also found that certain positions were significant factors for viruses. We think that the data we extracted form viruses will provide help for further investigations and experiments on influenza virus. Based on this research method, we think further research about comparing other epidemic influenza virus to find general amino acid sequence that make critical effect on human respiratory organism. Also, by using other analysis algorithm, such as Support Vector Machine, and SOM algorithm, we might find further rules that make difference between H1N1 and H3N2 clearer.

References

[1] J. J. Lee and T. Yoon, "The new approach on Fuzzy decision trees," *International Journal of Fuzzy Logic systems*, vol. 4, no. 3, July 2014.

[2] C. Heo and T. Yoon, "Deeper understanding about attributes of HIV employing support vector machine," *International Journal of Bioscience, Biochemistry and Bioinformatics*, vol. 4, no. 5, pp. 336-339, 2014.

[3] J. Yun, J. W. Seo, and T. Yoon, "The new approach on Fuzzzy decision trees," *International Journal of Fuzzy Logic Systems*, vol. 4, no. 3, July 2014.

[4] E. Go, S. Lee, and T. Yoon, "Analysis of ebolavirus with decision tree and Apriori algorithm," *International Journal of Machine Learning and Computing*, vol. 4, no. 6, December 2014.

[5] S. J. Lim and C. Heo, Y. Hwang, and T. Yoon, "Analyzing patterns of various avian influenza virus by decision tree," *International Journal of Computer Theory and Engineering*, vol. 7, no. 4, August 2015.

[6] J. H. Lee, S. H. Ahna, S. M. Pyuna, E. J. Janga, and T. Yoon, "Analysis of malaria inducing P. Falciparum P. ovale, and P. vivax through Apriori algorithm and decision trees".

[7] B. Christian and R. Kruse, "Induction of association rules: Apriori implementation," *Compstat. Physica-Verlag HD*, 2002.

[8] D. Y. Kim, H.-J. Kim, J. Bae, and T. Yoon, "Examining the probability of the critical mutation of H5N8 by comparing with H7N9 and H5N1 using Apriori algorithm and support vector machine," *International Journal of Computer Theory and Engineering*, vol. 7, no. 2, April 2015.

Seongpil Jang was born in Busan, Korea, in 1998. He is studying in Hankuk Academy of Foreign Studies, Natural Science. He studied statistics and bioinformatics, and wrote papers about his research results. He is still interested in statistics and many other algorithms that are used in analyzing genetical data sets. He is currently working on analyzing ML tree with Statistical method. Also, as he is interested in computational method, he is now doing his works about binal computing system and converting DNA sequence into binal data. Moreover, he is working with statistical method that can predict protein structure based on 1-dimensional amino acid sequence,

Hunseok Choi was born in Republic of Korea. He is currently a student in a science major of Hankuk Academy of Foreign Studies, Korea. He is interested in immunology, biotechnology and currently studying bioinformatics. He tries to learn more about various types of algorithm to operate the research well.

Yeonje Jung was born in South Korea, in 1998. He is now a student of Hankuk Academy of Foreign study natural science course. He lives in Seoul. He is interested in computer science and algorithm. He is trying hard to study computer language such as JAVA and C.

Eoram Moon was born in Korea, in 1998. He is an 11th grade student in the natUral Science Program at Hankuk Academy of Foreign Studies. He is now continuing his research in viruses and informatics via program with Taeseon, Yoon and his co-reseachers.

Taeseon Yoon was born in Seoul, Korea, in 1972. He got 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 in University of Korea, as a lecturer and as a adjunct professor in Ansan University. 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.