

Comparison between SARS CoV and MERS CoV Using Apriori Algorithm, Decision Tree, SVM

Seongpil Jang^{1,a}, Seunghwan Lee¹, Seong-Min Choi¹, Junwon Seo¹, Hunseok Choi¹, Taeseon Yoon²

¹ Natural Science, Hankuk Academy of Foreign Studies, HAFS, Yongin-si, Republic of Korea

² Department of Computer Science and Engineering, Korea University, KU, Seoul, Republic of Korea

Abstract. MERS (Middle East Respiratory Syndrome) is a worldwide disease these days. The number of infected people is 1038(08/03/2015) in Saudi Arabia and 186(08/03/2015) in South Korea. MERS is all over the world including Europe and the fatality rate is 38.8%, East Asia and the Middle East. The MERS is also known as a cousin of SARS (Severe Acute Respiratory Syndrome) because both diseases show similar symptoms such as high fever and difficulty in breathing. This is why we compared MERS with SARS. We used data of the spike glycoprotein from NCBI. As a way of analyzing the protein, apriori algorithm, decision tree, SVM were used, and particularly SVM was iterated by normal, polynomial, and sigmoid. The result came out that the MERS and the SARS are alike but also different in some way.

1 Introduction

Middle East respiratory syndrome (MERS) was first identified in Saudi Arabia since an outbreak of new flu in 2012. The respiratory infection is caused by MERS coronavirus (MERS-CoV), case of which have been reported in 25 countries including Saudi Arabia, Egypt, United Kingdom, United States, and South Korea. 1,211 infections and 492 deaths were reported so far, which means about 40% of those who infected die of the disease. Since no specific vaccine and treatment of the virus were known as of 2015, the comparison of the virus with SARS coronavirus (SARS-CoV) has largely increased in the importance

1.1 SARS CoV

Severe Acute Respiratory Syndrome (SARS), a respiratory disease first appeared in China and spread out over the countries such as Singapore, Vietnam since 2002. Large number of both human and animals were affected severely by SARS for more than a decade. According to WHO, 8,096 infections and 774 deaths were reported in 2003[1]. Fever above 38°C is the initial symptom of SARS, which is followed by shortness of breath. SARS is spread out by SARS-coronavirus (SARS-CoV), which is a positive and single-stranded RNA virus. It has diverse structure of proteins, which include RNA replicase (1A, 1B), spike glycoprotein(S), small envelope glycoprotein(E), membrane glycoprotein(M), nucleocapsid phosphoprotein(N). Each protein has

different roles, all of which are essential in various steps of virus life cycle [2].

1.2 MERS CoV

MERS (Middle East respiratory syndrome) causes lethal respiratory diseases. MERS is occurred by MERS-CoV, the single stranded RNA virus[3]. MERS is known to have come from camels, the host of the virus, but the exact route is not unfolded yet. Although MERS-CoV is coming up in endemic low level of threat, the possibility of MERS-CoV evolving to raise infectivity between human beings should not be overlooked. Corona virus is single stranded RNA virus. To enter host cells, MERS-CoV attaches to dipeptidyl peptidase 4, the receptor. And there is a process of fusion of the virus and the cell, using Protease cleavage of the S protein. Then the virus inserts the RNA to cytoplasm. The RNA of the virus then transcript and replicate on double membrane vesicles and other membranous structures, which are from the endoplasmic reticulum. Transcription of the seven subgenomic mRNA(messenger RNA)s happens through negative-strand subgenomic RNA intermediates. Subgenomic RNAs are 3' co-terminally nested and they are joined to a common leader encoded at the 5' end of the genome. The RNA of the virus is encapsidated in the N protein and is transported to the endoplasmic reticulum Golgi intermediate compartment (ERGIC), the site where they assemble. Viral RNA encapsidated in the N protein then buds into vesicles lined with the S, M, and E proteins. Vesicles are then transported to the cell surface before releasing [4].

^a heresphill@gmail.com

1.3 Spike glycoprotein

Spike glycoprotein(S protein) is one of common structural proteins of coronavirus, belonging to class I fusion protein [5]. It is arranged on the surface of the virus forming a unique “corona”, or crown-like appearance. The ectodomain of all coronavirus includes common two types of subunits, which are S1 domain and S2 domain. S1 domain, which is responsible for receptor binding, contains two subdomains, N-terminal subdomain and C-terminal subdomain. Both subdomains can function as receptor binding domains (RBDs) for host cells. 2 heptad repeats (HR1, HR2) can be found in S2 domain, which is responsible for virus-cell membrane fusion. The coronavirus infection begins with binding between RBD in S1 domain and receptor of the host cell, followed by viral fusion and entry resulted by the formation of fusogenic core between the HR1 and HR2 regions in S2 domain [6].

2 Method and experiment

2.1 Decision tree

Decision Tree is one of many data mining methods that are popular in many fields, including machine learning and statistical analysis. The algorithm can only classify category type structure inputs, which means in order to classify numerical value, it must first go through a discretization process to categorize the values [7]. By asking questions associated with data items, it uses a process of greedy algorithm, which branches down the original source into subsets, and shows the result of classification in a form of divided source. One of the pros of using decision tree algorithm is that it shows the result in a form of tree, which helps researches to analyze the result more easily [8].

2.2 Apriori algorithm

Apriori algorithm is famous for its simplicity though it is very effective in finding association rules between items. The algorithm takes two-step progress, which is finding large itemset and extracting all association rules which have confidence going over the minimum confidence [9]. In order to find large itemset, Apriori algorithm first calculates the support of each and every item sets that can be made, which is calculated in a formula of $P(A)$. (Where A is an item set.) After calculation, it extracts every item set that has more support than the minimum support that we first set. It repeats the progress until it finds no item set that has higher support than the minimum support, and eventually creates a large itemset [10]. Then it calculates the confidence of each and every association rules that can be made in a large itemset, which is calculated in a formula of $P(A | B)$. (Where the association rule is $B \Rightarrow A$.) After calculation, it extracts every association rule that has more confidence than the minimum confidence that we first set. Thanks to its simplicity and effectiveness, this algorithm is widely used even to these days [11].

2.3 SVM (Support Vector Machine)

SVM algorithm is an initial of “Support vector machines”. Also it is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Now let us introduce the system of this algorithm [12]. First this algorithm makes a “Maximum margin hyper plane” to separate group into two parts, and let’s say group A’s vector is A vector(point) and group B’s vector is B vector. And, let’s connect these vector(points) and when a polygon is made, we call this convex hull. The important thing is the vector(point) which is outside the polygon. We say the line and the side that can separate the group “hyperplane”. To test SVM’s validity, we compare Accuracy rate. If Accuracy rate is high, two parts are similar. If not, two parts are very different[13].

2.4 Experiment

In our research, we used SARS and MERS Spike glycol protein data. We obtained SARS (DQ412574.1) and MERS(KP236092.1) Spike glycol protein sequence data in NCBI. Next, we performed experiment with decision tree algorithm first. In this experiment, we deterred MERS as class 1, SARS as class 2 and performed 10-Fold validation test in 7, 9, 13 window. In 10-Fold validation test, 9 data is used to make algorithm proper, and 1data is used to classify. Then, we conducted another experiment with apriori algorithm. We conducted experiment in 7, 9, 13 window. After experiment, we extracted rules from results. Finally, we conducted other experiment with SVM. We conducted this experiment in Normal, Polynomial, and sigmoid method. With 10-fold validation test, we used 8 data to make algorithm proper, and we used 2 data to classify

3 Results

3.1 Decision tree

Table 1. Results under 7 window

Class	Rule	frequency
Class1	Pos9=H	.75
Class2	Pos9=H	.923

Table 2. Results under 9 window

Class	Rule	frequency
Class1	Pos5=H	.75
Class2	Pos1=H	.75

Table 3. Results under 13 window

Class	Rule	frequency
Class1	Pos7=N	.8
Class2	Pos7=I	.818

In results under 7window, rules that frequency is higher than .75, rule-designating amino acids are quite similar. In this result, we could find that SARS and MERS are very similar viruses.

In 9window and 13window results, there are quite difference between two viruses. Comparing these differences, there is possibility that amino acids like isoleucine, asparagine, arginine can be key factor of making differences between two viruses

3.2 Apriori algorithm

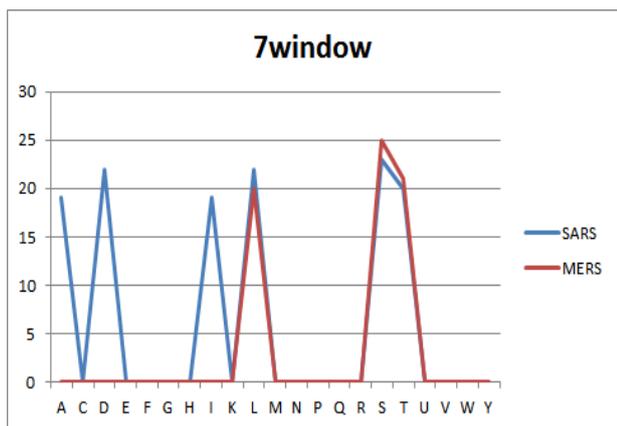


Figure 1. Results under 7window

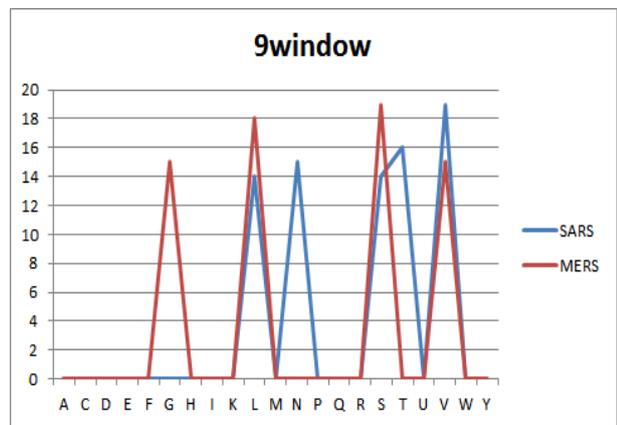


Figure 2. Results under 9window

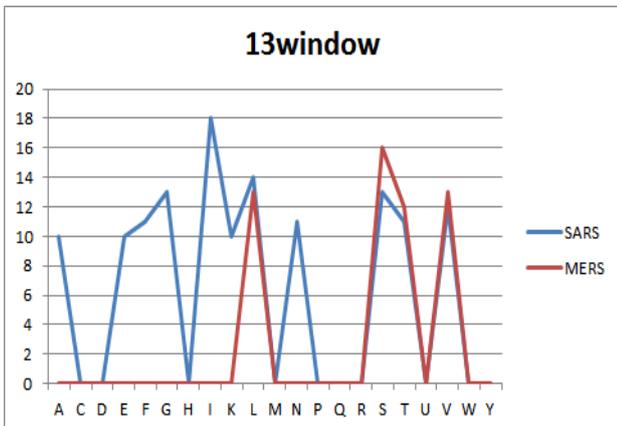


Figure 3. Results under 13window

As we can see in results by apriori algorithm, there is high similarity between SARS and MERS. However, there are some differences. First, as we can see in Figure 1 and Figure 3, amino acid sequence is differ from isoleucine between two viruses.

Next, in Figure 2 and Figure 3, we can find that asparagine is key factor that differ MERS from SARS. Comparing decision tree results under 13 window to apriori results under 13 window, we could find high similarity.

That is, isoleucine and asparagine are main factors that differentiate MERS from SARS

3.3 SVM

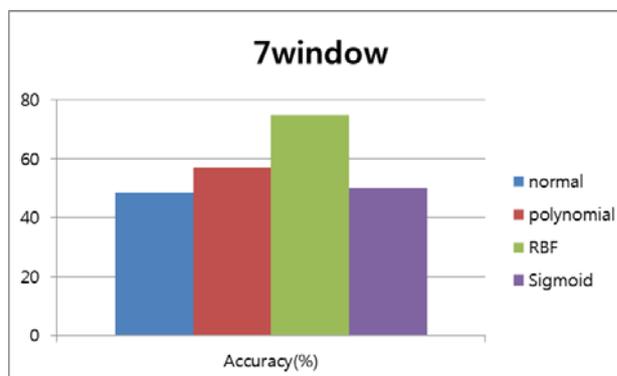


Figure 4. Results under 7window

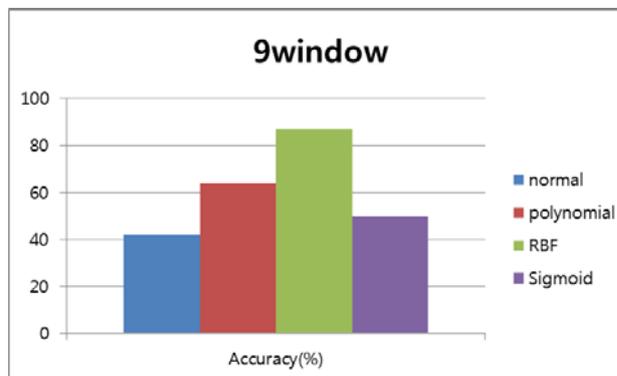


Figure 5. Results under 9window

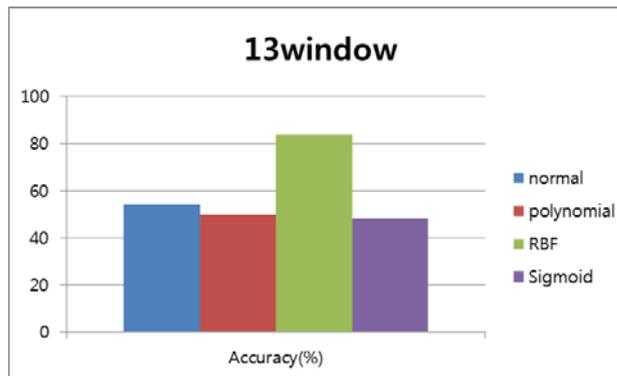


Figure 6. Results under 13window

In Normal SVM experiment, we can find that Accuracy rate is quite low. This means that there is obvious difference between MERS and SARS. Similarly, as we can see in Polynomial SVM experiment and sigmoid SVM experiment, we can find that other methods are showing low Accuracy rate. These results are showing that MERS and SARS is quite difficult to be distinguished by linear classifying process. However, different with those linear classifying process, RBF algorithm, which is non-linear classifying process, showed some positive results.

As we can look in 7 window, 9 window, 13 window results, the accuracy of RBF algorithm results are quite high. They are even higher than accuracy 75%. These results are showing that MERS and SARS viruses are very similar viruses. However, by non-linear classifying process, MERS and SARS viruses can be distinguished easily. In conclusion, MERS and SARS viruses are quite different viruses at all.

4 Conclusion

This paper applied the computational program, apriori algorithm, decision tree, and SVM. To conduct our experiment, we used Spike glycoprotein of MERS and SARS. To analyze this, we used three different algorithms. Apriori algorithm is algorithm that extracts frequent rules from given dataset. With decision tree, we can classify two viruses. With SVM, we found that MERS and SARS are truly different viruses.

MERS virus is very critical and epidemic disease. However just like our experiment, MERS is different from SARS. In order to block the virus from spreading out and killing people, we performed experiment to analyze key factor of MERS virus.

As we can find, rule extracted from apriori algorithm showed that amino acid Isoleucine and Asparagine are the main factors of MERS. For further research we are going to conduct experiment with other proteins except Spike glycoprotein.

References

1. Peiris, J. S. M., *et al.* Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study, *The Lancet*, **361**, 9371, 1767-1772 (2003)
2. Seto, W. H., *et al.* Effectiveness of precautions against droplets and contact in prevention of nosocomial transmission of severe acute respiratory syndrome (SARS), *The Lancet*, **361**, 9368, 1519-1520 (2003)
3. de Groot, Raoul J., *et al.* Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group, *Journal of virology*, **87**, 14, 7790-7792 (2013)
4. Assiri, Abdullah, *et al.* Hospital outbreak of Middle East respiratory syndrome coronavirus. *New England Journal of Medicine*, **369**, 5, 407-416 (2013)
5. Simmons, Graham, *et al.* Characterization of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) spike glycoprotein-mediated viral entry. *Proceedings of the National Academy of Sciences*, **101**, 12, 4240-4245 (2004)
6. Qian, Zhaohui, Samuel R. Dominguez, and Kathryn V. Holmes. Role of the spike glycoprotein of human Middle East respiratory syndrome coronavirus (MERS-CoV) in virus entry and syncytia formation. *PloS one*, **8**, 10, e76469 (2013)
7. Borgelt, Christian, and Rudolf Kruse. Induction of association rules: Apriori implementation, *Compstat. Physica-Verlag HD* (2002)
8. Borgelt, Christian, Recursion Pruning for the Apriori Algorithm, *FIMI* (2004)
9. Friedl, Mark A., and Carla E. Brodley, Decision tree classification of land cover from remotely sensed data, *Remote sensing of environment*, **61**, 3, 399-409 (1997)
10. Magerman, David M. Statistical decision-tree models for parsing. *Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics* (1995)
11. Safavian, S. Rasoul, and David Landgrebe. A survey of decision tree classifier methodology, *IEEE transactions on systems, man, and cybernetics* **21**, 3, 660-674 (1991).
12. Schüldt, Christian, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. *IEEE* (2004)
13. Furey, Terrence S., *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**, 10, 906-914 (2000)