ORIGINAL ARTICLE

Influenza Early Warning Model Based on Yunqi Theory*

HU Xue-qin (胡雪琴)¹, Gerald Quirchmayr², Werner Winiwarter³, and CUI Meng (崔蒙)¹

ABSTRACT Objective: To establish an early warning model to simulate the outbreak of influenza based on weather conditions and Yunqi theory, an ancient calendar theory of Chinese medicine (CM). Methods: Tianjin, a northeastern city in China, was chosen as the region of research and applied the influenza-like illness attack rate (ILI) as the baseline and warning line to determine the severity of influenza epidemic. Then, an influenza early warning model was constructed based on the theory of rough set and support vector machines (RS-SVM), and the relationship between influenza and meteorology was explored through analyzing the monitoring data. Results: The predictive performance of the model was good, which had achieved 81.8% accuracy when grouping the obtained data into three levels that represent no danger, danger of a light epidemic, and danger of a severe epidemic. The test results showed that if the host qi and guest qi were not balanced, this kind of situation was more likely to cause influenza outbreaks. Conclusions: The outbreak of influenza closely relates to temperature, humidity, visibility, and wind speed and is consistent with some part of CM doctrine. The result also indicates that there is some reasonable evidence in the Yunqi theory.

KEYWORDS Yunqi theory, influenza, meteorological factors, early warning model

Yunqi theory is a unique theory that ancients studied the relationship between climate change and disease. Guided by the thought of correspondence between human and nature, it associates the meteorological changes of nature with the state of the human body to predict the occurrence of disease in the human body and proposes a cyclical model for the prevention and treatment of diseases.¹¹

Since the severe acute respiratory syndrome (SARS) epidemic in 2002, some scholars used this theory to test and verify the trend of SARS. Gu¹² summarized that the drought situation and the high temperature in 2000 were the two main climate factors that caused the epidemic outbreak. This phenomenon is consistent with the prediction of "an epidemic will develop in three years" in Yunqi theory. Since then, this theory has once again aroused people's interest and attention. In this paper, in order to build the meteorological prediction model of influenza based on Yunqi theory, we will briefly introduce two related research background facts about our work.

Influenza is a highly contagious infectious disease, with the characteristics of sudden onset, rapid spread, serious infection, and regional group-occurring. In this paper, influenza was diagnosed based on the World Health Organization (WHO) definition of influenza-like illness (ILI), which states that "fever ≥100 °F (37.8 °C), oral or equivalent, and cough or sore throat, and in the absence of other laboratory confirmed evidence." Previous research on the health effects of climate change has mainly focused on the impact of meteorological indicators such as temperature and humidity. However, in this paper, we added a new indicator to our early warning model, i.e., the relationship between host qi and guest qi based on Yunqi theory.

METHODS

Data Sources
So far there exist 441 network laboratories and 556 national influenza surveillance sentinel hospitals in China, and the surveillance network has covered all provinces. Physician-diagnosed ILI is considered to be a widely applicable alternative to influenza virus isolation. However, techniques for the detection of
influenza viruses have not been widely used at clinical hospitals in China. Therefore, this paper applied the ILI attack rate (ILI%) as the baseline and warning line to determine the severity of influenza epidemic. We hoped to do the early influenza forecast by observing the warning line.

In this paper, the data on ILI% were obtained from the Chinese Center for Disease Control and Prevention (CDC), and the duration of monitoring was from October 3, 2005 to September 20, 2009 (a total of 207 weeks).

\[
\text{ILI\%} = \frac{\text{Number of cases of ILI}}{\text{Number of cases of outpatient and emergency}}
\]

When cases that met the case definition of ILI were found, designated staff within China CDC’s medical institutions should fill the case report forms for infectious disease within the specified time limit. Therefore, the data had better accuracy and representativeness. The epidemic curve of weekly reported ILI% is shown in Figure 1.

Meteorological data were downloaded from the National Meteorological Information Center (http://cdc.cma.gov.cn/index.jsp), which were the daily monitoring data in the same period, including mean temperature, dew point, mean humidity, mean sea level pressure, mean visibility, and mean wind speed. A total of 14 meteorological factors have been selected in this paper. In order to correspond with the weekly report of ILI%, meteorological elements were processed into weekly averages as well.

Data Preprocessing
Rough set (RS) theory could not handle the continuous attributes directly; therefore, we had to carry out the classification of ILI% first. We set the baseline value and warning line value as the dividing line between groups and classified the whole influenza data into three groups: light, moderate, and serious. The baseline and warning line settings were as follows. We did normality tests on the ILI% data during the epidemic periods of 2005–2009 (November to April of next year) by PASW statistics 17.02; if they met features of the normal distribution, we chose \( \bar{x} \pm 2s \) as the warning line; if not, we used the 75th percentile value as the warning line. Because the distribution of ILI% was abnormal, we took 11.8% (P75) as the baseline. The same principle was applied to the establishment of baseline values. First, the ILI% data in the non-epidemic period (May to October) was tested for normality; if they met the normal distribution, \( \bar{x} \pm 2s \) was chosen as a baseline value; if not, we took the 75th percentile as the baseline value. In this paper, we took 17.7% (P75) as the warning line value through statistical analysis. A non-epidemic period was under the baseline, a serious epidemic was beyond the warning line, and between these two lines was a light epidemic period. The groups of influenza data are shown in Table 1.

Table 1. Groups of Influenza Data

| ILI%    | Level      | Group            |
|---------|------------|------------------|
| <11.8   | 1          | Non-epidemic period |
| [11.8, 17.7) | 2          | Light epidemic period |
| >17.7   | 3          | Serious epidemic period |

According to the Yunqi theory, one year can be divided into six seasons, and each season has two qi names: host qi and guest qi. There are three types of relationships between these two kinds of qi, which are listed in Table 2. We took the Yunqi situation of the year 2006 as an example that is listed in Table 3.

Table 2. Types of Relationship between Host Qi and Guest Qi

| Relationship between host qi and guest qi | Type |
|------------------------------------------|------|
| The host qi and guest qi were balanced   | 0    |
| The host qi and guest qi were not balanced and antagonistic | -1   |
| The host qi and guest qi were not balanced but still in well-being | 1    |

Building RS and Support Vector Machines Model
RS theory, introduced by Pawlak in the
early 1980s\cite{10} is a mathematical tool to deal with vagueness and uncertainty. One of the main advantages of RS theory is that it does not need any preliminary or additional information about data. Attribute reduction is a key problem in RS theory and its applications. A reduct is a minimal attribute subset of the original data that has minimal redundancy but still has the same discernibility power as all the attributes in the RS\cite{11}. In this paper, we had 14 conditional attributes, and some of them were redundant attributes. Therefore, we used attribute reduction based on RS theory to acquire the minimal attribute subset that included six attributes. It improved the classification efficiency of support vector machine (SVM) when we used the minimal attribute subset as the input nodes of our model.

Meteorological factors are continuously changing values within a certain range; therefore, these data have to be discretized first. The equal frequency binning technique was used to do the discretization of the raw data. The interval number 3 was used to get the break point set of the decision table.

A genetic algorithm was implemented to do the attribute reducts. The algorithm described by Øhrn, et al\cite{3} was used, which supported for both cost information and approximate solutions. The algorithm's fitness function \( f(B) \) is

\[
    f(B) = (1 - \alpha) \times \frac{cost(A) - cost(B)}{cost(A)} + \alpha \times \min \left( \epsilon, \frac{|S \cap B - \delta|}{S} \right)
\]

where \( S \) is the set of sets corresponding to the discernibility function. The subsets \( B \) of \( A \) were found through the evolutionary search driven by the fitness function. The parameter \( \alpha \) defines weighting between subset cost and hitting fraction, while \( \epsilon \) was relevant in the case of approximate solutions. The result of attribute reduction is listed in Figure 2.

\[\text{Table 3. Yunqi Situation of the Year 2006 (Month-Day-Year)}\]

| Time interval | First qi | Second qi | Third qi | Fourth qi | Fifth qi | Sixth qi |
|---------------|----------|-----------|----------|-----------|----------|----------|
| Type          | 0        | –1        | 1        | 1         | 1        | 1        |

\[\text{Table 4. Example of Decision-Making Rules}\]

| Mean temperature (°C) | Mean humidity | Min visibility (km) | Max wind speed (km/h) | Mean wind speed (km/h) | CM | Group |
|------------------------|---------------|---------------------|-----------------------|------------------------|----|-------|
| [20.07, *)             | [87.42, *)    | [2.84, *)           | [\*, 18.71]          | [\*, 7.21]             | (0, 1) | 1     |
| [4.64, 20.07]          | [* , 64.00]   | [2.84, *)           | [22.64, *)            | [10.07, *)             | (0, 1) | 2     |
| [* , 4.64]             | [64.00, 87.42] | [* , 1.77]          | [* , 18.71]          | [* , 7.21]             | (1, *) | 3     |

\[\text{Figure 2. Attribute Reductinos}\]

After consulting experts and combining the results with the expertise, we selected the first reduction as the set of attributes that included six attributes. It is shown in Figure 2 that Chinese medicine (CM, Yunqi theory) was included in each set of attribute reduction, from which could be conjectured that these two events might have a certain correlation.

\[\text{RESULTS}\]

Based on the result of attribute reduction, 104 decision rules generated by the genetic algorithm were acquired. By the method of basic filtering, the rule set was reduced to 27 rules. Three examples of rules for decision-making are listed in Table 4.

As seen in Table 4, influenza epidemics had certain characteristics, namely low temperature, low visibility, proper humidity, low wind speed, and the host qi and guest qi not being balanced but still in well-being. The partial results of our studies were consistent with other studies. For example, a study conducted by Lowen, et al\cite{12} indicated that influenza virus transmission was dependent on relative humidity and temperature, especially the relative humidity. Based on our test, we summarized that the visibility and wind speed played important roles in the influenza epidemic. Recently, as the Chinese economic has been booming, the air pollution of the city got worse. The low level of visibility became gradually one of the accelerators to the respiratory system diseases. For example, Brunekreef, et al\cite{13} found that exposure to pollutants...
such as airborne particulate matter and ozone has been associated with increases in mortality and hospital admissions due to respiratory and cardiovascular diseases. Furthermore, we also found that the low wind speed would accelerate the flu epidemic. Probably, the low wind speed slows down the air flow and leads to a higher concentration of virus.

**Influenza Early Warning Model**

In this paper, RS was used for information preprocessing to get the attribute reduction set, and then SVM was applied to do influenza classification and prediction. During network training, the attribute set obtained by RS reduction constituted the input layer of the network. The individual steps of influenza prediction were described in the following five sections.

We had a new set \( D_{\text{SVM}} \) according to the new attribute set and the corresponding raw data \( D_{\text{RST}} \). Then, we normalized the new attribute set with \( x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \), where \( x'_i \) represented the normalized attribute value and \( x_i \) represented the attribute value before normalization. \( \max(x_i) \) and \( \min(x_i) \) were the maximum and minimum values of the \( i \)th attribute. Sixty-three percent data were randomly selected from \( D_{\text{SVM}} \) as training set \( D_{\text{SVM,Tr}} \). Cross-validation was used to train the data, the penalty factor was set \( C=5 \), and radial basis function was chosen as the kernel function. The remaining 37.0% data formed the test set \( D_{\text{SVM,Ts}} \). It was classified by the SVM classifier to output the classification results. The classification prediction test was done on the ROSETTA platform based on RS theory.

**Prediction**

According to the process described above, the classification prediction on the Matlab7.8.0 platform was completed. Finally, an accuracy of 81.8% (63/77) was achieved. This was better than the result of RS if we did the prediction on the ROSETTA platform directly. The result of RS-based prediction only reached 61.3%. Specific classifications of the groups are shown in Figure 3. The prediction accuracy for each group is listed in Table 5.

**DISCUSSION**

Our experimental results suggest that the prediction accuracy of RS-SVM was not very high even if it was higher than RS. When analyzing the reasons, we found that scarce (and possibly noisy) data might be the reason. Here, we focused on the second and third groups of data to analyze the cause of classification errors. The analysis results suggest that there were some objective reasons that led to such classification errors.

For example, erroneous data item No. 72 belonged to the third group actually; however, our model forecasted it as the second group. We checked the date of this week, which was from May 14, 2007 to May 20, 2007. We also checked the meteorological records of this week to find that the maximum wind speed was higher than usual for this period, and the instantaneous wind speed reached 50 km/h or more in 4 consecutive days. We checked the weather records at that time and found that a number of dust storms have occurred during this week in Tianjin. Therefore, we speculate that this particular error has a certain relation with the dust storms. In addition, by looking at the data for the erroneously classified items 22 and 26, we suspect that these errors may be caused by the clinician's diagnosis errors since the diagnosis of influenza is easily confused with the common cold or other upper respiratory tract infectious diseases. The low reported influenza data will lead the RS-SVM model to misjudge the second group as the third.

In summary, the main work of this paper collected the continuous ILI% data and related meteorological data from Tianjin city for 5 years. Then, the relationship between influenza and weather was analyzed based on RS theory. Finally,
a minimal attribute subset was acquired and an RS-SVM prediction model for influenza was constructed. Our results are summarized as follows. (1) From the results of RS attribute reduction, we draw the conclusion that ILI% had a close relationship with mean temperature, mean humidity, minimal visibility, maximal wind speed, and mean wind speed. It was also found that influenza spread and Yunqi theory also have a very close relationship, which will help to develop the theory of CM. (2) Some decision rules were established for influenza epidemics according to the meteorological factors. The results showed, for example, that when mean temperature was below 4.6 °C, mean humidity was between 64 and 87, maximal wind speed was below 18.7 km/h, mean wind speed was below 7.2 km/h, and the host qi and guest qi were not balanced but still in well-being, there was a high possibility that a serious influenza epidemic season would occur. (3) An RS-SVM was an established influenza early warning model, which had better performance when compared with the RS prediction model. It illustrated that an RS-SVM model for influenza prediction was entirely feasible and would produce a satisfactory result.

In future work, we will improve our study in the following areas: (1) based on ILI% data processing, we will use the average ILI% of 3 weeks (the same, before, and after week) as the ILI% value of 1 week. This method will make up for the shortcomings of data bias (such as holidays) to some extent and is conducive to data stability of monitoring. (2) So far, we have just applied a linear warning line. In the future, we will try a non-linear warning line to observe whether it can produce a more objective analysis of the influenza trend and increase the sensitivity of warning. (3) Since ILI% is not entirely consistent with influenza pandemics, we have to find a better measure when ILI% data are inconsistent with the rate of influenza virus isolation and influenza outbreaks. (4) In this paper, we had only used one Yunqi factor in our predictive model; in the future, we hope we can apply more Yunqi factors, such as pestilence every 3 years, qi going too far, and qi not going far enough, in disease forecasting and prevention. Through such work, we hope to open a new door for this ancient theory and to promote and develop it.

In short, we hope our research can help the public health officials and professional medical staff to better respond to regional, seasonal influenza. Our study results can be further tested through future, additional laboratory, epidemiological, and other modeling studies.

Acknowledgment
We are grateful for the guidance and help from the Chinese Center for Disease Control and Prevention.

REFERENCES
1. Huang TX, Liu HT. Practical doctrine of Yunqi theory: climate and disease prevention. Beijing: Academy Press;2005:16-17.
2. Gu ZS. Yunqi theory in Canon of medicine and disease prediction. Clin J Tradit Chin Med (Chin) 2004;16:93-95.
3. Yang L. Yunqi theory of traditional Chinese medicine. Beijing: Beijing Science and Technology Press;2003:64-65.
4. Pawlak, Z. Rough set theory and its applications. J Telecommun Inc Technol 2002;3:7-10.
5. Yao YY, Zhao Y. Attribute reduction in decision-theoretic rough set models. Inf Sci 2008;178:3356-3373.
6. Øhrn A, Komorowski J, Skowron A, Synak P. Rough sets in knowledge discovery 2: applications, case studies and software systems. In: Polkowski L and Skowron A, eds. Studies in fuzziness and soft computing. Berlin: Physica-Verlag,1998;19:572-576.
7. Lowen, AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathogens 2007;3:1470-1476.
8. Brunekreef B, Holgate ST. Air pollution and health. Lancet 2002;360: 1233-1242.

(Received April 25, 2011)
Edited by WANG Wei-xia