The Prediction Model of Language Users around the World Based on Neural Networks

Xiaojiang Yuan [a], Yuzhou Gao [b] and Yidong Chen [c]

Jimei University Xiamen, Fujian, China,

Changshu Institute of Technology Suzhou, Jiangsu, China,

Broward College Fort Lauderdale, Florida, America,

[a]498300260@qq.com, [b]28092900@qq.com, [c]378730973@qq.com

Abstract

Language is the key to communication. This document carries out a research on the developing trend of world languages. First we built a quantity model of language users (NLS). This model includes factors of economy, technology, culture, policy and immigration. Then we employed principal component analysis to screen the main factors of each language. After that, we built a neural network to predict the changes of the number of language users in the next fifty years. Eventually, we came to the conclusion that the number of language users will not be influenced by slow growth of world population and the pattern of immigration. Moreover, the geographical distribution of these languages will stay relatively steady in the meantime.

Keywords

Principal Component Analysis; Neural Network; Prediction; Language Use.

1. Introduction

There are around 6,900 kinds of languages in the world, ten of which are spoken by about half population of the world. The level of native language, however, is different from that of second language. For instance, Mandarin ranks the first in native speakers because of the large population of China, while English is the second popular language around the world. The number of users of a language may fluctuate as time passes, the reasons of which vary. What is more, in a world that shares technologies, languages separated by distance are able to influence and interact with each other.

2. Principal component analysis

Coordinate rotation in N-dimensional space is the essence of principal component analysis. It will not change the digital structure, and primitive variables of linear combination are uncorrelated, thus it can reflect the information carried by primitive variables to the maximum extent. This document employs principal component analysis to analyze the relativity between all factors. We assume the index of p consists of three dimensional random variables.

$$X^0 = (x_1, x_2, x_3, \cdots, x_p)$$

Comprehensive index:

$$\begin{cases} y_1 = a_{11}x_1 + a_{21}x_2 + a_{31}x_3 + \cdots + a_{p1}x_p \\ \vdots \\ y_p = a_{1p}x_1 + a_{2p}x_2 + a_{3p}x_3 + \cdots + a_{pp}x_p \end{cases}$$

In the above equation:

$$\sum_{j=1}^{p} a_{jk}^2 = 1$$

y_1 is maximum variance, and each variable is uncorrelated.

Assume that $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p$ is the eigenvalue of covariance matrix, thus the contribution rate of principal component k is:

$\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$ The higher the total contribution rate of the first m principal components is, the less the loss of information is.

This document relies on Rstudio in order to acquire the chief components of each language, and the conclusion is as follows:

The chief component which influences UK is mainly immigration.

The chief component which influences Malaysia and Bengal is economy.

The two chief components which influence Russia is culture and technology, the contribution rate of which is 59% and 88%.

The chief component of each language is different.

3. Building the model of neural networks

Self-learning and self-adaption are the two characteristics of neural networks, which will not be influenced by many human factors under the circumstance of not building complicated models. Thus, we employ neural networks to predict the number of language users.

This document builds the following prediction system of the number of language users based on the above analysis. We arrange the date of the population of ten languages from 1966 to 2016, with the help of international integrated data set. After interpolation and data standardization, we choose 56 sets of training data and 50 sets of data for testing.

This document builds neural networks on the basis of data standardization, using trained neural networks to predict the number of people who speak Mandarin in 2067, as shown in Figure 1.

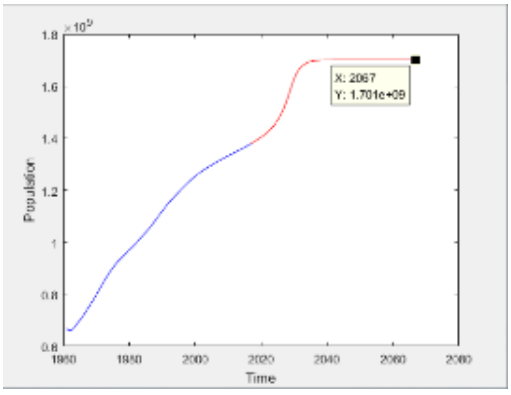

Figure 1 The prediction result of NAR neural networks

As shown in Figure 1, the number of Mandarin speakers is increasing. The trend becomes mild without number increasing in 2040. Thus, we predicate that there will be 1.701 billion Mandarin speakers in 2067.

We continue the fitting and prediction of other countries, which almost share the same trend except the countries speaking Portuguese and Russian. The numbers are all increasing and then maintain steady till 2020. As we can see, the number of Portuguese speakers increases a little over the past 60 years, but will fluctuate in the following 50 years. And the number of Russian speakers first increases

a little then decreases a little over the past 60 years. In the following 50 years, that number will slightly increases and then slightly decreases.

Last but not least, we employ neural networks to predict world population, as shown in Figure 2.

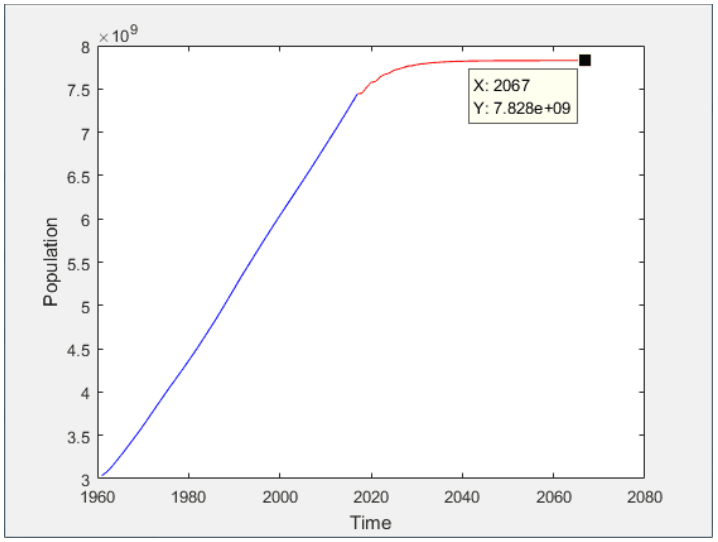

Figure 2 The prediction of world population

Like most countries being analyzed, world population first increases, then maintain steady. We predict that world population will reach 7.828 billion in 2067.

4. Summary

The changing of language users is like the immigration of human beings. Although massive immigration of British population is the major trend, the prediction ratio of English over last fifty years is 5.71%, 0.55% less than that of 2016. So we are able to come to the conclusion that the number of language users will not be influenced by slow growth of world population and the pattern of immigration. Moreover, the geographical distribution of these languages will stay relatively steady in the meantime.

References

[1] Yang Jie, Zhan Jun, Zhang Jichuan. 30 examples of neural network in Matlab[M]. Publishing house of electronics industry, 2014:01-01