Prediction of the Number of Speakers of World Languages by SVR

Shengqi Yuan *
School of North China Electric Power University, Baoding, China
* iridium_77@163.com

Abstract. The number of speakers of World Languages is an important figure to analysis the language. Support Vector Machine Regression Model can find a function to minimize the distance between each sample and the hyperplane, this function can be used to predicted the number of speakers precisely. To predict the number of speakers of top 10 language in future 50 years, Chinese will become the language of the largest number of people in the world and Telugu and Malay will replace Portuguese and Russian into the top 10 list.

Keywords: Prediction; Support Vector Regression; Language.

1. Background

The distribution of languages in the world is uneven. Among about 6,900 languages,10 languages [2] represented by English, Chinese and Spanish are the mother tongues of half of the world's population. There have also been many changes in the number of speakers in languages, which is caused by many factors, such as natural growth migration, transfer of refugees and so on.

Harris Drucker [5] compared support vector regression with a committee regression technique based on regression trees and ridge regression done in feature space. And expected that SVR will have advantages in high dimensionality space and doesn't need much training data.

2. Factors of Language Development

To predict the number of common language speakers, there are many factors. Natural growth rate, migration, transfer of refugees is the main factors. The political changes, economic development in the region will also affect the distribution of the speakers.

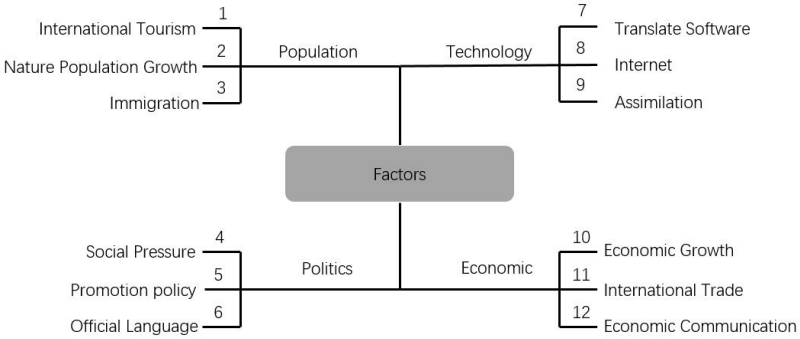

Fig.1 Influential factors classification

3. Support Vector Machine Regression Model

For the prediction of the number of people in the future, the simple use of multiple regression cannot get more accurate prediction results, meanwhile, the known data of the speakers is little. In order to obtain more accurate prediction results, it can be referred to SVR Model [1] to predict the data in the future.

The core idea of SVM is to find a hyperplane far away from the two types of sample points. When predicting the population growth, a hyperplane can be considered, which can minimize the distance between each sample and the hyperplane.

Assuming the data collection $T=\{(x_i,y_i),i=1,2,\cdots,l\}$ of annual number of speakers, among them, x_i is the column vector of the i sample, $\mathbf{X_i}=[x_i^1,x_i^2,\cdots,x_i^d]^T$, y_i is the corresponding output.

Assuming that a linear regression function is established in the high-dimensional feature space. $F(x)=w\Phi(x)+b$ Where $\Phi(x)$ is a non-linear mapping function.ϵ,defining the linear insensitive loss function ϵ.

$$L(f(x),y,\varepsilon)=\begin{cases}0,|y-f(x)|\leq\varepsilon\\|y-f(x)|-\varepsilon,|y-f(x)|>\varepsilon\end{cases}$$

Where $f(x)$ is a predicted value returned from the regression function, y is the true value of the corresponding, if the difference between $f(x)$ and y is less than or equal to ϵ, then the loss is zero, then we define the slack variables $\mathrm{x}_i,\mathrm{x}_i^*$, the problem of looking for w,b is expressed as the following expression which C is the penalty factor. The greater the error, the greater the penalty.

$$\begin{cases}\min\dfrac{1}{2}\|w\|^2+C\sum\limits_{i=1}^{i}(\xi_i)+\xi_i^*\\ s.t.\begin{cases}y_i-w\Phi(\mathbf{x}_i)-b\leqslant+\xi_i & ,i=1,2,\cdots,l\\ -y_i+w\Phi(\mathbf{x}_i)+b\leqslant\epsilon+\xi_i^*\\ \xi_i\geqslant0,\xi_i^*\geqslant0\end{cases}\end{cases}$$

Considering the Lagrange function and converting it to the dual form:

$$\max_{a,a^*}\left[-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j)-\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\epsilon+\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i\right]$$

$$s.t.\begin{cases}\sum\limits_{i=1}^{l}(\alpha_i-\alpha_i^*)=0\\ 0\leqslant\alpha_i-\alpha_i\leqslant C\\ 0\leqslant\alpha_i-\alpha_i^*\leqslant C\end{cases}$$

among them, the kernel function $K(x_i,y_i)=\Phi(x_i)\Phi(y_i)$,so We can get the regression function is

$$f(x)=b\Phi(x)+b^*=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,x)+b^*$$

In order to test the accuracy of the predicting results, defining the mean square error of the test set E and the decision coefficient R^2 as follows:

$$E=\frac{1}{l}\sum_{i=1}^{l}(\hat{y}_i-y_i)^2$$

$$R^2=\frac{\left(l\sum\limits_{i=1}^{l}\hat{y}_iy_i-\sum\limits_{i=1}^{l}\hat{y}_i\sum\limits_{i=1}^{l}y_i\right)^2}{\left(l\sum\limits_{i=1}^{l}\hat{y}^2-\left(\sum\limits_{i=1}^{l}\hat{y}_i\right)^2\right)\left(l\sum\limits_{i=1}^{l}y^2-\left(\sum\limits_{i=1}^{l}y_i\right)^2\right)}$$

4. Training and Testing of SVR Model

According to the UN [2], World Bank [3], and ethnographic data [4], the data of each native speakers in 6 years is easy to get.

The data of the previous three years are taken as input samples to train the model and the predictive value of the test sample is obtained. By testing the model, calculating that the MSE of SVR prediction model is 0.0125 and the coefficient of determination is 0.9511, which can be found that the model has a good generalization ability and the predicting error is small.

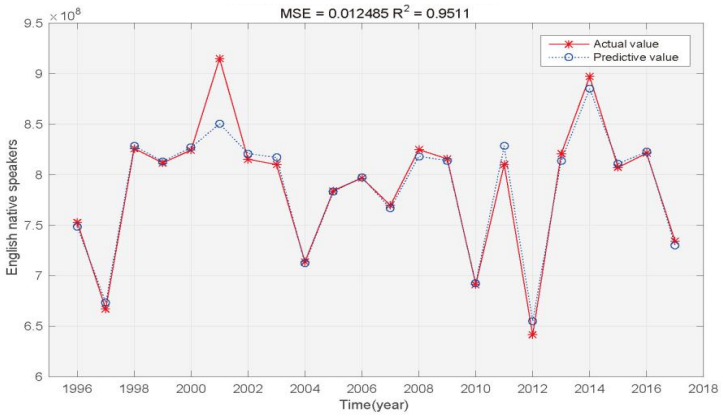

Fig. 2 SVR training test

When the availability of the predicting model is confirmed, the prediction of the number of speakers in the future and draw the below figure.

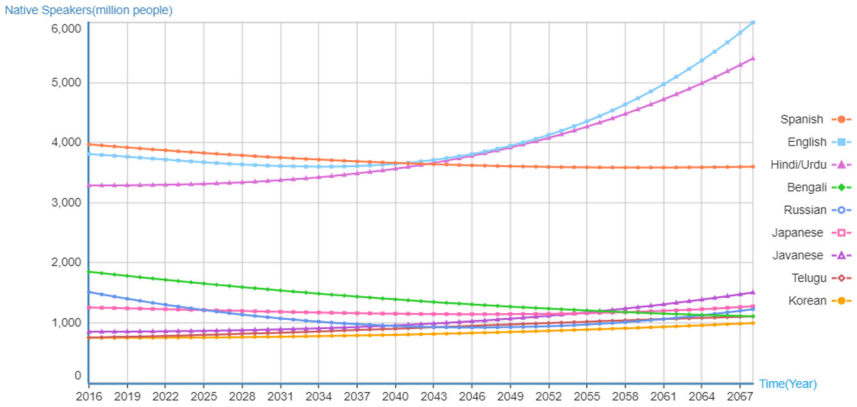

Fig. 3 Long-term prediction of native speakers

The population of Chinese is too large to observe the changing trends of the remaining languages well, so it is not drawn here.

Table 1. Numbers of speakers of top 10 languages

Area	Language	Toltal Speakers(10^3 people)	Percentage
African	Arabic	300759.98	25.96
East Asia	Chinese	1787343.09	86.45
East Europe and Russia	Russian	173536.98	90.35
India and Central Asia	Hindustani	581914.76	24.5
North America except Mexico	English	450604.37	69.61
Oceania	English	29227.95	83.54
South America and Mexico	Spanish	371435.7	48.87
South-East Asia	Vietnamese	120445.56	10.02
West Asia	Arabic	162739.42	18.78
West Europe	German	159246.36	18

5. Summary

Predict the number of speakers of world languages is an important thing. to obtain more accurate predicting results, using SVR model to simulate the native speakers and the total speakers in the next 50 years. Chinese will become the language of the largest number of people in the world. Based on the predicted results, Russian will into the top ten list.

Among these 12 influential factors, the regional natural population growth rate and immigrant occupy a greater weight respectively. After remove these two factors, SVR predict result change slightly, it can show that the SVR model has a great robustness.

References

[1]. Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression.Statistics and computing, 14(3):199–222, 2004.

[2]. Information on: http://data.un.org/Data.aspx?d=POP&f=tableCode%3a406.

[3]. Information on: https://data.worldbank.org.cn/.

[4]. Information on: https://www.ethnologue.com/statistics/status.

[5]. Utta Von Gleich, Language spread policy: the case of Quechua in the Andean republics of Bolivia, Ecuador, and Peru,2009.