Wa Language Syllables Classification by Support Vector Machine Based on Genetic Algorithm

Mei-jun FU, Jie SU and Wen-lin PAN*
School of Mathematics & Computer Science, Yunnan Minzu University
*Corresponding author

Keywords: Wa language, Support Vector Machine (SVM), Genetic Algorithm (GA), K-Crossover Validation Method (K-CV).

Abstract. In this paper, propose the method to classify Wa syllables by support vector machine (SVM) based on genetic algorithm (GA). Firstly, select all the Wa syllables from the speech corpus recorded. Secondly, extract pitch and formant frequencies by vowel main body extension method and linear predictive coding method respectively. Thirdly, describe every syllable with a 12-dimensional feature vector composed of its mean values, minima and maxima, belonging to its pitch frequency sequence and its first three formant frequency sequences respectively. Fourthly, divide the obtained data set into training set and test set, the former use to train SVM, and the latter test SVM. Fifthly, normalize data set into [0,1], and select the radial basis function as kernel function. Finally, use the K crossover validation method (K-CV) based on genetic algorithm (GA) to get the best parameter values train SVM. As a result, the predictive accuracy of Wa syllables classification by test set arrives at 84.83%, which has shown that the proposed method is effective and feasible on Wa syllables classification.

Introduction

The Wa nationality[1, 2] is one of Chinese nations. They live mainly in the southwest border of Yunnan Province, China and northern Myanmar. And the native language they use to communicate with each other is Wa language. Wa language belongs to Va-De’ang branch, Mon-Khmer languages, Austro-Asiatic family. China Wa language can be divided into three kinds of dialects that are "Barao dialect", "Lawa dialect" and "Wa dialect". Currently, Wa language is based on the Barao dialect and the voice of YanShuai is standard pronunciation. The speech corpus used here was all recorded in YanShuai.

It is well known that pitch and formant frequencies[3-5] are the two most important parameters of speech signal. The pitch changing pattern is called pitch tone that plays a very important role in distinguishing the meaning of different syllables. Formant includes two important parameters: resonant frequency and its width of bandwidth (bandwidth), which are important parameters that can distinguish different vowels, including voiced consonants.

Syllable classification, in essence, still belongs to pattern recognition[6], which can be summarized as "feature extraction $\rightarrow$ syllable pattern recognition". For every Wa syllables, after speech signal preprocessing, we extract its each frame of pitch frequency and of the first three formant frequencies. Then, we take the mean values, minima and maxima belonging to pitch frequency sequence and the first three formant frequency sequences respectively, to form a 12-dimensional feature vector as its classification pattern. We will classify these obtained syllable classification patterns by support vector machine (SVM) based on genetic algorithm (GA).

It is well known that after mapping inseparable samples into high dimensional linear separable feature space by means of its kernel function, support vector machine (SVM) converts the high-dimensional space problem into a quadratic programming problem by the structural risk minimization principle, and then obtains the global optimal solution through the convex optimization method. Compared with the
traditional pattern recognition methods, SVM is more suitable for dealing with nonlinear and high dimensional pattern recognition problems with small samples. For the above reason[7-11], SVM is very suitable for Wa syllables pattern recognition under the small sample condition. However, in the process of recognizing Wa syllables by SVM, penalty parameters and kernel function parameters will directly affect the recognition ability to SVM. Not only are the traditional methods computationally complex but also difficult to obtain the global optimal solution, which makes it difficult to improve the adaptive identification ability to SVM. Genetic Algorithm (GA)[12] is a kind of random searches and optimization methods, possessing the global parallel search ability. Therefore, in this paper, in order to improve the adaptive identification ability to SVM, we will use the K-cross validation method (K-CV) based on Genetic Algorithm (GA) to search the global optimal solution more efficiently. And experiment to find out that the predictive accuracy on Wa syllables classification arrives at 85.23%, which has shown that proposed method is effective and feasible on Wa syllables classification.

Extract Wa-language Syllable Classification Patterns
The main steps to construct Wa syllable patterns are expressed in the following (Figure 1).

1) Selecting all the Wa syllables speech signal from the corpus.
2) Preprocessing Wa syllables, including eliminating direct current (DC) components[13], eliminating trend items[14], framing, windowing, and pre-emphasising.
3) Endpoint detection by improved spectral entropy method[15].
4) Pitch extraction by vowel main body extension method. Formant extraction by linear predictive coding (LPC)[15] method based on adaptive variational mode decomposition (AVMD)[16].
5) For any given Wa syllables, taking the mean values, minima and maxima belonging to pitch frequency sequence and the first three formant frequency sequences respectively to form a 12-dimensional feature vector as its syllables classification pattern.

C-Support Vector Machine (C-SVM)
Here, mainly introduce the C-support vector machine model (C-SVM) with binary classes. For two classes of sample set: \( \{ (x_i, y_i) \}_{i=1}^{l} \), \( x_i \in \mathbb{R}^n \), \( y_i \in \{-1,1\} \), where \( y_i \) is the category label corresponding to the \( i \)-th input eigenvector \( x_i (i = 1, \cdots, l) \), selecting appropriately the kernel function \( K(\bullet, \bullet) \) and parameter \( C \) for SVM, we can construct the following optimization model:

\[
\max \sum_{i=1}^{l} \alpha_i - (1/2) \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \quad s.t. \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1, \cdots, l;
\]

which is a convex quadratic programming problem. Hence, there exists the optimal solution. Solving the above, we get the optimal solution: \( \alpha^* = (\alpha_1^*, \cdots, \alpha_l^*)^T \) where \( T \) represents transpose of matrix. Take a positive component \( \alpha_j^* \) from \( \alpha^* \) satisfying \( 0 < \alpha_j^* < C \), then the corresponding optimal decision function is...
Similarly, we can obtain a multi-class support vector machine by modifying directly the above objective function. Certainly, on the other hand, we can construct a multi-class support vector machine by combining some binary class support vector machines. Here, we will use multi-class C-support vector machine (C-SVM).

It is well known that the necessary and sufficient condition as a kernel function of SVM is that the function satisfies Mercer theorem. The choice of kernel function directly affects the generalization ability and recognition precision. In a SVM, the kernel function used commonly includes mainly four types: linear function, polynomial function, gaussian radial basis function, sigmoid function.

**C-support Vector Machine (C-SVM) Based on Genetic Algorithm (GA)**

**Wa-language Syllable Classification Based on C-SVM**

**Data Preprocessing.** Based on three different normalization ways, the predicted classification accuracy of the same test set is shown respectively in Table 1.

**Selecting Kernel Functions.** For four types of different kernel functions, the predicted classification accuracies of the same test set are shown in Table 2, when the data set is normalized into [0, 1].

From Table 1 and Table 2, we found that when we normalize the data set into [0, 1] and use Gaussian kernel function, we get the highest accuracy with 81.3953%.

**Parameters Optimization**

It is necessary to adjust the relevant parameters (mainly the penalty parameter c and the radial basis function parameter g) so as to obtain the best prediction accuracy when predicting classification by SVM. In Model 1, the penalty parameter c and the radial basis function parameter g are given by means of test experience. Generally speaking, using the cross validation (CV) method to find the best parameters c and g under a certain sense, we can effectively avoid under-fitting and overfitting. In this paper, we use the K-CV method to optimize our C-SVM related parameters.

At first, we search roughly the best parameters \( \log_2 c \) and \( \log_2 g \) in a large grid such as \( \log_2 c \times \log_2 g \in [-8,8] \times [-8,8] \), according to specific step sizes \( \Delta(\log_2 c) = 0.5 \) and \( \Delta(\log_2 g) = 0.5 \). Here, \( \log_2 x \) denotes base-2 logarithm of \( x \), and \( \Delta(\log_2 x) \) the step size of \( \log_2 x \). In this case, the best \( c=27.8576 \) and \( g=0.57435 \), the classification average accuracy at this time is 83.6207%.
Support Vector Machine Based on Genetic Algorithm

In the K-CV sense, we can find the best parameters c and g by grid search method. Therefore, we are able to obtain the highest classification accuracy in the K-CV sense, that is, the global optimal solution. Sometimes, however, we find that our search for the optimal range of parameters c and g will be very time-consuming. Using the heuristic algorithm, it is possible to find the global optimal solution without traversing all the parameter points in the grid.

Genetic Algorithm (GA). Genetic algorithm (GA) is an evolutionary algorithm, using selection, crossover and mutation operators to search the optimal solution. Its global search ability is strong while its local search ability is weak. Generally speaking, we get only the suboptimal solution of our problem, rather than the optimal solution. The basic elements of genetic algorithm include chromosome coding method, fitness function, genetic operation and operating parameters. The basic steps of genetic algorithm include: coding, initial population generation, selection, crossover, mutation. The operation parameters of the genetic algorithm in this paper are shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3. Genetic Algorithm Operation Parameters Setting.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
</tr>
<tr>
<td>Maximum Generation</td>
</tr>
<tr>
<td>The number binary bits of the variables</td>
</tr>
<tr>
<td>Crossover probability</td>
</tr>
<tr>
<td>Mutation probability</td>
</tr>
<tr>
<td>Generation probability</td>
</tr>
</tbody>
</table>

![Figure 2. Algorithm of optimizing SVM parameters by GA.](image1)

![Figure 3. The fitness (accuracy) curve based on GA.](image2)

Search the Best Parameter Based on GA. Here, the accuracy of K-CV in training set is regarded as the fitness function value in GA. The overall algorithm process of optimizing SVM parameters by GA is shown in Figure 2. Use this method, we have the best penalty parameter c=23.2021 and the best kernel function parameter g=0.50926 (Figure 3), different from those by K-CV Method based on grid.

Conclusion

In this paper, we proposed the SVM method based on GA to classify Wa-language syllables. From the experimental results, the classification accuracy both reaches 84.83%, if data set is normalized into [0, 1] and SVM is equipped with Gaussian radial basis function, using the best parameters c and g obtained based on whether grid or GA. This same result shows that our method is effective on Wa-language syllable classification. Of course, as we pointed out that our SVM method based on GA is more efficient.
In the future, we will study more efficient methods for pitch and formant extraction to improve the accuracy of GA-SVM.

Acknowledgments
This work is financially supported by Yunnan Provincial Department of Education Science Research Fund Project (2016YJS078), Yunnan Minzu University Overseas Masters Program (3019901).

References