An Ensemble Learning Method for Dialect Classification

Shuai Ye¹*, Ruoyan Zhao¹, Xinru Fang¹

¹ Software and Applied Science and Technology Institute of Zhengzhou University, Zhengzhou, China

*Corresponding author e-mail: zzuyeh@163.com

Abstract. Dialect Classification Task is the first step of the Multilingual Automatic Speech Recognition System. Because of the difference of accent between dialects in different regions, the problem of Dialect Classification is a very challenging one. Dialect classification is widely used in information processing, military information retrieval and other fields. Therefore, the study of dialect classification is of great significance. This paper proposes an ensemble learning method for dialect classification. Firstly, the low accuracy of dialect data sets is processed and amplified. Then, three models, GRU, CNN and DNN, are used to classify dialects respectively, and the final dialect types are determined by voting. The accuracy of dialect classification by this method is higher than that of the single model with the best performance, the validity of the model is verified.

1. Introduction

With the continuous development of artificial intelligence application field, the recognition and classification of dialects have attracted more and more attention, but the recognition of dialects is much more complex than that of mandarin [1-2]. Although great progress has been made in the research and development of Mandarin speech recognition, there are many differences between dialects in the vast area, especially those in the south and standard mandarin. It is of great significance to study dialect recognition and classification and improve its classification accuracy [3].

The main task of dialect classification is to get the corresponding dialect types from a given speech data set. The structure of a speech recognition system is shown in Figure. 1 below. It consists of four modules: frontend processing, acoustic model, language model and decoder [4]. In this paper, the method of improving classification accuracy is mainly aimed at the acoustic model in the system. The task of the acoustic model is to calculate $P(O | W)$, which is to generate the probability of speech waveform for the model. The acoustic model is an important part of the speech recognition system, which occupies most of the cost of speech recognition. Traditional speech recognition systems generally use GMM-HMM based acoustic model. But DNN-HMM-based acoustic model can achieve significant performance improvement compared with the traditional GMM-HMM acoustic model. CNN model is generally used in image processing[5]. The application of the spectrum map in the task of dialect recognition and classification to CNN model can overcome the instability caused by sampling time and frequency in traditional dialect recognition[6]. DNN is good at mapping features to independent space. However, DNN and CNN models do not consider the association information between voices. Therefore, we consider using DNN + CNN + RNN model fusion method for dialect recognition and classification tasks.
Ensemble Learning [7] is widely used in classification and regression tasks. Its idea is very simple, that is, to construct several different classifiers and combine them linearly to get a stronger classifier to make the final decision, that is, the idea of "three smelly cobbler top one huge Liang". Given that GRU, CNN and DNN have their own advantages for dialect recognition and classification, and their advantages are complementary, this paper adopts the method of Blending Model method for dialect recognition. For modeling ability, CNN is good at reducing frequency domain changes, GRU can provide long-term memory [8], DNN is good at mapping features to independent space [9]. In this paper, the model first trains and recognizes dialects independently through GRU, CNN and DNN, and then classifies dialects by voting based on the recognition results to further improve the classification accuracy.

In the second part, we will introduce the Ensemble Learning and the structure of our network model in detail; in the third part, we will introduce our experimental process and analyze the experimental results; in the fourth part, we will evaluate the model and outlook the application scenarios of the model based on the experimental results.

2. Related Works

2.1. Ensemble Learning and Integration Strategy

Ensemble Learning is a machine learning method that uses a series of learners to learn, and uses Ensemble Learning method to integrate the learning results, so as to achieve better learning effect than a single learner. If a single classifier is assimilate to a decision maker, the method of ensemble learning is that multiple decision makers making a decision together. There are two main problems to be solved in ensemble learning. The first is how to get several individual learners. The second is how to choose a combination strategy to assemble these individual learners into a strong learner.

There are two choices for the first problem. The first is that all individual learners are of one kind, or homogeneous. The second is that all individual learners are not of one kind, or heterogeneous. At present, homogeneous individual learners are the most widely used. Generally speaking, the method of ensemble learning refers to homogeneous individual learners. Homogeneous individual learners can be roughly divided into two categories according to whether there are dependencies. The first is the strong dependencies between individual learners. That is, a series of individual learners need to be
generated serially. The representative algorithm is boosting series. The second is individual and random forest. There is no strong dependence between the learners in Random Forest. A series of individual learners can be generated in parallel. The representative algorithm is bagging.

Considering that the three basic models selected in the dialect recognition and classification problem are homogeneous, that is, the algorithm we adopt is bagging. Bagging's weak learners have no dependencies and can be generated in parallel. The training set of Bagging's individual weak classifier is obtained by random sampling. Through T times of random sampling, we can get T sampling sets. For this T sampling set, we can train T weak classifiers independently, and then get the final strong classifier through the set strategy of T weak classifier. For the dialect classification problem in this paper, due to the small number of data sets, the method of random sampling is not used, but all data sets are used to train each basic classifier. The structure of the Ensemble Learning in the dialect classification problem is shown in Figure. 2 below.

![Ensemble Learning Structure](image)

**Figure 2. Ensemble Learning Structure**

### 2.2. Our Network Structure

CLDNN (Convolutional, long short-term memory, fully connected deep neural networks) [10] has achieved good results in speech recognition. There are two layers of CNN application in CLDNN, which is the representative of shallow CNN application. CNN and LSTM have better performance than DNN in speech recognition task. For modeling ability, CNN is good at reducing frequency domain changes, LSTM can provide long-term memory, so it has a wide range of applications in time domain, and DNN is suitable for mapping features to independent space. In CLDNN, CNN, LSTM and DNN are combined in a network to achieve better performance than a single network. Referring to CLDNN, this paper chooses CNN, GRU and DNN as the basic classification model for dialect classification. There are many similarities between CNN and GRU, such as the expansion of traditional neural network, the updating of model by forwarding calculation and backward calculation, the coexistence of multiple neurons in each layer of neural network horizontally and the connection of multilayer neural network vertically. But there are also many differences, such as CNN for spatial expansion, RNN for time expansion, neurons and multiple time output calculations. RNN can be used to describe the output of continuous state in time with memory function; CNN can be used for static output; CNN can reach a depth of more than 100 layers under certain conditions, while RNN has limited depth. It is of great significance to fuse these basic models with a large amount of information and have temporal and spatial characteristics. Our network structure is shown in Figure. 3 below.
Feature extraction methods in voice datasets include spectral analysis, FBank feature [11], MFCC feature [12] and so on. For CNN network, we use input spectrogram; FBank feature lacks the DCT cepstrum link of MFCC feature extraction. The other steps are the same. FBank feature is close to the response characteristics of human ears, and the extraction steps are relatively simple compared with MFCC. For GRU and DNN network, FBank feature is input. Firstly, the speech data set is preprocessed, then FBank features are extracted frame by frame, and FBank feature matrix is input into the neural network for training. Converting the original sound signals into spectrograms is usually not enough for dialect recognition and classification. Since the spectrogram is a two-dimensional image and suitable for CNN input, we continue to use CNN to extract high-level features based on the spectrogram, after extracting the feature vectors of the spectrogram, we continue to use CNN to extract high-level features. Because in the process of cutting the spectrum, the number of the spectrum can be cut is related to the length of speech signal, but it is impossible to recognize and classify the speech features of different lengths, so the problem of how to get the uniform length features of speech signals with different lengths is solved by pooling the features after convolution operation. Therefore, for the speech spectrogram extracted from the audio signal, it is overlapped in the time domain and cut into image blocks. Then it is sent into CNN to extract the feature vector $X$. For the vector set of the same audio sample, the pooling method is used to get the fixed length feature vectors describing the whole audio sample, and the full connection layer is used to classify the dialects.

3. Experiment
The dialect classification speech data set includes five hundred dialect phonetic data of Changsha, Hebei, Hakka, Minnan, Nanchang and Shanghai. It is recognized and classified by GRU algorithm. The recognition accuracy of each dialect is shown in the table 1 below.

| Dialect Type | Correct Number | Accuracy Rate |
|--------------|----------------|---------------|
| Changsha     | 490            | 98.02%        |
| Heibei       | 488            | 97.61%        |
| Hakka        | 181            | 36.24%        |
| Minnan       | 246            | 49.22%        |
| Nanchang     | 460            | 92.33%        |
| Shanghai     | 447            | 89.42%        |
We classify different dialects based on different deep learning acoustic models, and measure the accuracy of dialect recognition using LSTM, GRU, CNN and DNN algorithms, the result is shown in Table 2 below.

Table 2. Dialect Classification Accuracy under Different Algorithms

| Number | NOAA-LSTM | GRU | CNN | DNN | Fusion Model |
|--------|-----------|-----|-----|-----|--------------|
| 2      | 99.37%    | 99.37% | 99.04% | 98.33% | 99.78% |
| 3      | 95.43%    | 95.88% | 90.46% | 88.34% | 97.44% |
| 4      | 91.40%    | 90.76% | 86.33% | 84.23% | 95.33% |
| 5      | 88.74%    | 86.34% | 80.42% | 77.42% | 91.44% |
| 6      | 83.42%    | 78.30% | 74.02% | 72.08% | 85.29% |

By analyzing the data in the table above, it can be seen that when the number of classification type is 6, the recognition accuracy of several algorithms is at the middle and low levels. The recognition and classification accuracy of GRU is slightly lower than LSTM, but the structure of GRU algorithm is simpler and the generality is better than LSTM. So we choose GRU as the cyclic neural network algorithm in the Ensemble Learning. On the six classification problem, the recognition accuracy of GRU, CNN and DNN algorithm is quite different. The analysis shows that the three algorithms have different ideas and complementary advantages, so they are more suitable as the basic classifier algorithm in Ensemble Learning.

To explore the validity of the fusion model, the classification accuracy of DNN, CNN, GRU, NOAA-LSTM and fusion model on the two classification and only six classification problems were measured, and the experimental results were plotted as a broken line graph as shown in the following Figure 4.

Figure 4. Accuracy Comparison between Fusion Model and Other Model
4. Conclusion
In this paper, we describe the process of model fusion for dialect recognition and classification. Based on the classification accuracy of six dialects, we compare the accuracy of dialect classification using GRU, CNN, DNN, NOAA-LSTM and fusion model separately, and prove the validity of the fusion model.

Acknowledgments
Thanks very much for the funding and support of Henan Province's College Students Innovation and Entrepreneurship Training Plan (project number: S201810459069) and Zhengzhou University's College Students Innovation and Entrepreneurship Training Plan (project number: 201810459069).

References
[1] Ma, Bin, Donglai Zhu, and Rong Tong. "Chinese dialect identification using tone features based on pitch flux." 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Vol. 1. IEEE, 2006.
[2] Lei, Yun, and John HL Hansen. "Dialect classification via text-independent training and testing for arabic, spanish, and chinese." IEEE Transactions on Audio, Speech, and Language Processing 19.1 (2011): 85-96.
[3] Bahari, Mohamad Hasan, et al. "Non-negative factor analysis of gaussian mixture model weight adaptation for language and dialect recognition." IEEE/ACM transactions on audio, speech, and language processing 22.7 (2014): 1117-1129.
[4] Zhang, Shiliang, et al. "Compact Feedforward Sequential Memory Networks for Large Vocabulary Continuous Speech Recognition." Interspeech. 2016.
[5] Zhang, Mengmeng, Wei Li, and Qian Du. "Diverse region-based CNN for hyperspectral image classification." IEEE Transactions on Image Processing 27.6 (2018): 2623-2634.
[6] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).
[7] Zhou, Zhi-Hua. "Ensemble learning." Encyclopedia of biometrics (2015): 411-416.
[8] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).
[9] Al-Marri, Mubarak, et al. "Computer Aided Qur’an Pronunciation using DNN." Journal of Intelligent & Fuzzy Systems Preprint (2018): 1-15.
[10] Sainath, Tara N., et al. "Convolutional, long short-term memory, fully connected deep neural networks." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.
[11] NADEU, et al. "Time and frequency filtering of filter-bank energies for robust HMM speech recognition." Speech Communication 34.1-2(2001):93-114.
[12] Alam, Md. Jahangir, P. Kenny, and D. O'Shaughnessy. "Low-variance Multitaper Mel-frequency Cepstral Coefficient Features for Speech and Speaker Recognition Systems." Cognitive Computation 5.4(2013):533-544.