Gender Recognition of Speech based on Decision Tree Model

Biliang Zhong [a], Yunxin Liang [b], Jiyu Wu [c], Bisen Quan [d], Chunxiang Li [e], Wei Wang [f], Jincun Zhang [g] and Zhenzhang Li [h, *]

Guangzhou Maritime University, Guangzhou 510725, China

[a]gdzhbl@126.com, [b]953971062@qq.com, [c]1040558191@qq.com, [d]1769128867@qq.com, [e]1085454204@qq.com, [f]1131294393@qq.com, [g]1776990374@qq.com, [h, *] zhenzhangli@126.com

Abstract. This paper introduces the decision tree binary classification algorithm to classify the gender of speech data. In speech this unstructured data, in order to recognize the gender of the input voice data by computer, it is necessary to extract the features of unstructured voice data into structured data that can be recognized by computer, and then use structured data and decision tree for binary classification to train the binary classification decision tree model. Finally, the binary classification model of decision tree is used to predict and identify the gender of structured speech data that need classification and recognition.

Keywords: Feature extraction, Decision Trees for Binary Classification, Identify gender.

1. Introduction

With the development of information technology, people's operation of computer is not limited to the input of keyboard and mouse and other basic equipment, but gradually turn to intelligent voice operation. Smart phone voice assistant can easily help people make phone calls, and PC voice assistant can help play music, which cannot be realized without voice recognition, voice understanding and other voice signal processing technology.

In the experiment of speech recognition, it is found that the accuracy of speech recognition will be improved if the gender of the speaker is known, compared with the gender of the speaker. In order to improve the accuracy of speech recognition, gender recognition of speech data is an important issue. In recent years, speech recognition has attracted extensive attention. Hunt [1] and MERON [2] use the weight space traversal and linear regression method to train the weights. Alias [3]and others screened the randomly generated weight samples based on genetic algorithm. Park [4] and others regard the selection of primitives as a classification problem in pattern recognition, and adopt the method of distinguishing training in speech recognition to train weight.

There are only two possible genders for identifying voice data: "male" and "female". It's a binary classification problem. At present, the gender recognition of speech data by classification algorithm has been relatively mature, and the accuracy rate of gender recognition by the classification model obtained from stable training data can reach more than 99%. Based on the binary classification algorithm of decision tree [5,6], this paper classifies and identifies the gender of the speech data with feature extraction.

2. Feature Extraction of Speech Data

The gender recognition of voice data is firstly to digitize the collected voice data and convert it into a digital signal sequence that is convenient for computer storage and processing. This step is called feature extraction of voice data. The data in this paper are from Primary objects [7], in which the feature extraction of speech data is carried out with the help of functions provided by warbleR package of R language. The warbleR package provides an open online repository for acoustic signal detection and adaptive control optimization measurements. It ueses similarity methods for cross correlation, dynamic time bending, measurement of acoustic parameters and analysis of interactive sound signals. Feature extraction steps of speech data are shown in figure 1. Table 1 shows some results of voice data extraction.

Table 1. Part of the data extracted from the voice data feature, the label field is the male female logo, 1 indicates male, 0 indicates female, and is used as the identification of the training model

maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.202531646	0.656250000	0.0781250	3.5937500	3.5156250	0.299191919	1
0.228571429	0.737079327	0.0781250	3.6484375	3.5703125	0.285776805	1
0.258064516	0.700846354	0.0937500	3.4062500	3.3125000	0.219134537	1
0.202531646	0.438878676	0.0390625	3.6796875	3.6406250	0.161212446	1
0.168421053	0.15234375	0.0937500	0.2109375	0.1171875	0.318518519	1
0.115942029	0.402698864	0	3.5781250	3.5781250	0.128509046	1
0.271186441	0.910714286	0.1640625	6.2421875	6.0781250	0.257069409	0
0.202531646	0.575892857	0.15625	5.3437500	5.1875000	0.165778499	0
0.250000000	0.627155172	0	6.2031250	6.2031250	0.148434689	0
0.225352113	0.363467262	0.1562500	3.5078125	3.3515625	0.118531469	0
0.190476190	1.1328125	0.1640625	6.1796875	6.0156250	0.252272727	0
0.210526316	0.677201705	0.1640625	6.9687500	6.8046875	0.125854245	0
0.258064516	0.741477273	0	6.2187500	6.2187500	0.149497487	0
0.242424242	0.383101852	0.1640625	4.6171875	4.4531250	0.096896086	0

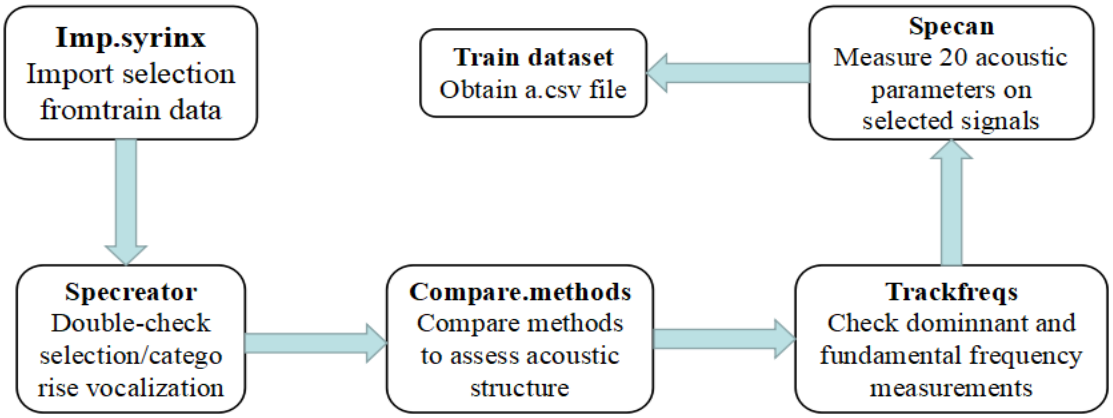


Fig. 1 Flow chart of the data extraction steps

3. Establishment of Decision Tree Model

In the classification problem, the regression tree (CART) [8], which is a binary tree, has been widely used in speech feature extraction. In the whole classification tree, each node is assigned a "Yes/No" problem, and according to such assignment, candidate primitives entering the root node are screened. Finally, each candidate primitive into the root node is entered into a leaf node according to the answer to a series of node questions. Candidate elements entering the same leaf node are considered to have similar contextual and acoustic characteristics. Decision tree is a combination of data-driven method and knowledge-based method for classification. Compared with data-driven method, it can give appropriate parameter estimation for primitives with sparse training data. And compared with the knowledge-based method, it can make up for the deficiency of expert knowledge.

Create a decision tree model with feature fields and labels. There are two ways for decision tree to split nodes. First is to split by Gini coefficient. Second, the second is to split in the way of information entropy. Gini coefficient (D) reflects the probability that two samples are randomly selected from data set D with inconsistent category markers. The smaller Gini(D) is, the higher purity of data set D is. $\text{Gini(D)}=1-\sum p_j^2\ 1-\sum p_j^2$. The gini-split decision tree steps are as follows:

(1) if all instances in the training data set belong to the same class of male or female, then T is a single-node tree, and this same class is taken as the class mark of this node, and T is returned;

(2) If feature A = bah, then T is a single node tree, and the class marker of the node is the one with the largest number of instances in the training data set, the decision tree T return;

(3) Otherwise, the Gini coefficients of train data set D for existing features are calculated. For each possible value a for each feature A, the train data set D is divided into D1 and D2 parts according to the test of A=a for "yes" or "no" at sample points, and then the Gini coefficients for A=a are calculated.

(4) Among all possible features A and all possible segmentation, points a, the feature with the smallest Gini coefficient and its corresponding segmentation points are selected as the optimal feature and the optimal segmentation point. According to the optimal feature and the optimal segmentation point, two sub-nodes are generated from the existing node, and train data set D is allocated to two sub-nodes according to the feature.

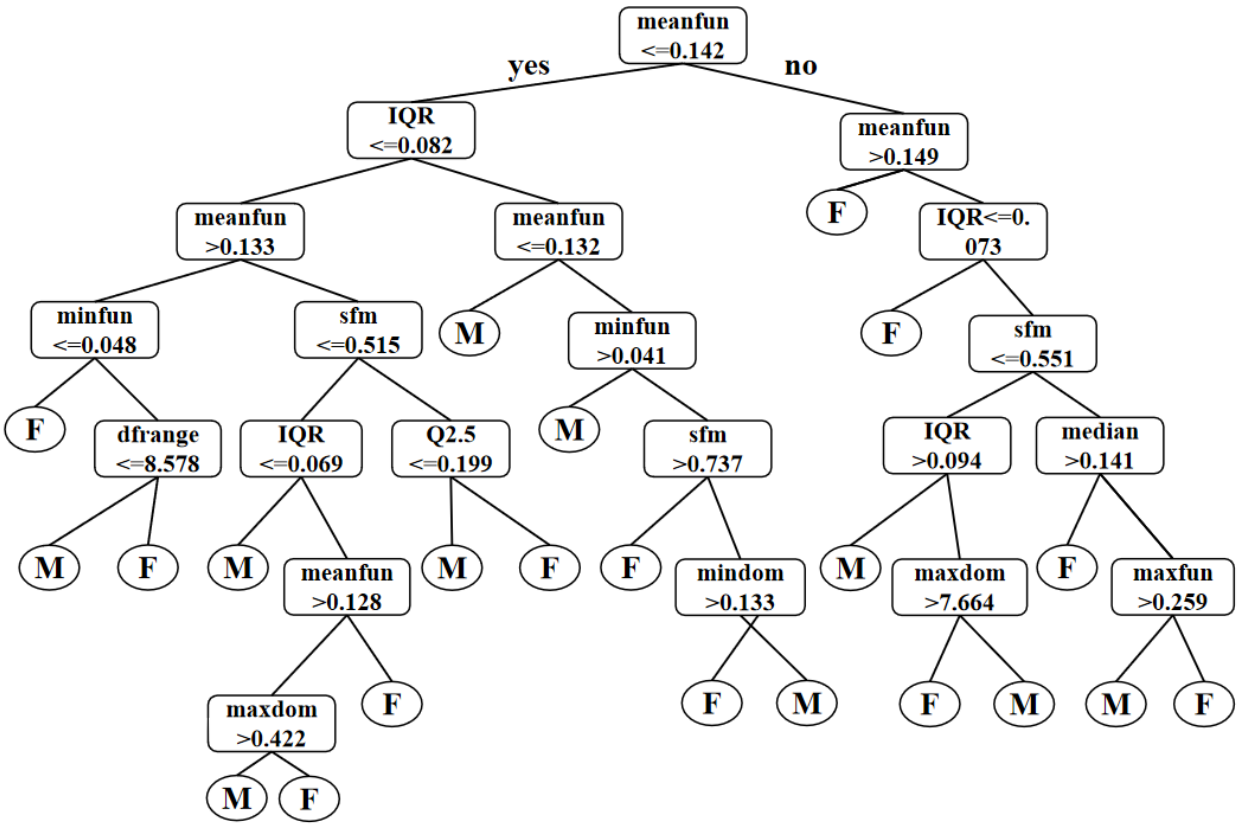


Fig. 2 Decision tree diagram, F for female, M for male

(5) Step (3) ~ (4) is called recursively for two sub-nodes until the stop condition is satisfied and the tree T is returned.

For the information entropy splitting method, the information entropy, H (X) is used to describe the uncertainty $H(X)=\sum_{x\in X} p(x)\log\frac{1}{p(x)}$; Conditional entropy H (X | Y) is used to describe the information entropy H (X | Y) = H (X, Y) - H (Y) of X under the premise that Y occurs. The information gain G (D, A) is used to represent the information of feature X and to reduce the uncertainty of information of class Y by G (D, A) = H (D) - H (D | A). The implementation steps are as follows:

(1) if all instances in the train data set belong to the same class of male or female, then T is a single-node tree, and this same class is taken as the class mark of this node, and T is returned;

(2) If feature A = bah, then T is a single-node tree, and the class label of the largest number of instances in the training data set is used as the class label of the node, and the decision tree T is returned.

(3) Otherwise, the information gain of each feature in feature A to train data set is calculated, and the feature Ag with the greatest information gain is selected.

(4) If the information gain of Ag is less than the threshold value, then T is set as a single node tree, and the class Ck with the largest number of instances in D is used as the class marker of the node, and T is returned.

(5) Otherwise, for each possible value AI of Ag, train data set D is divided into several non-empty subsets Di according to Ag=ai. The classes with the maximum number of instances in Di are used as markers to construct sub-nodes. The tree T is composed of nodes and their sub-nodes, and T is returned.

(6) for the ith child node, Di is used as training set and A-{Ag} is used as feature set. Recursively call steps (1) ~ (5) to get the subtree Ti, return Ti. According to the data in this paper, the structure of the decision tree model is shown in Figure 2.

4. Evaluation Method of Decision Tree Model

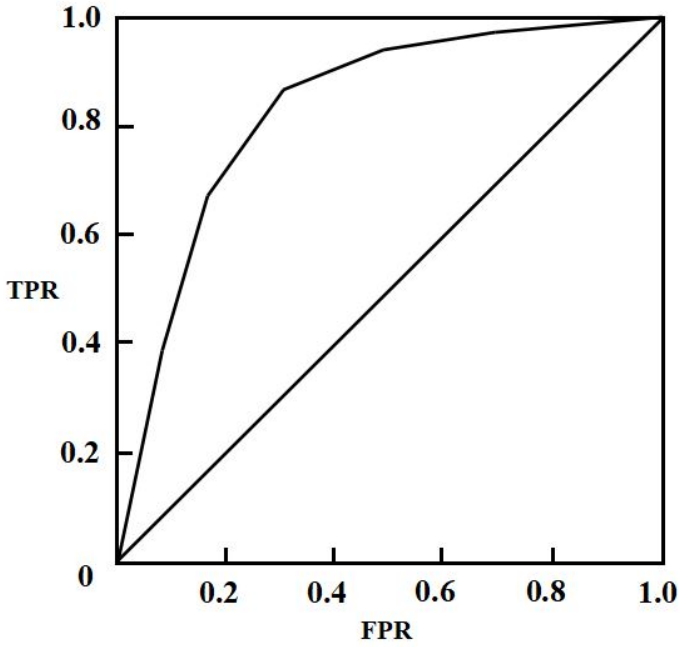


Fig. 3 ROC curve

The ROC [9] observation model correctly identifies the trade-off between the proportion of the positive case and the proportion of the model that incorrectly identifies the negative case data as a positive case. Here, we represent 0 for women; 1 for men. Thus, True Positives [10] (TP) indicates a prediction of 1, which is actually 1; Flase Positives (FP) indicates a prediction of 1, which is actually 0; True Negatives (TN) indicates a prediction of 0, which is actually 1; Flase Nagatives (FN): indicates that the prediction is 0, which is actually 0. Let TPR=TP/(TP+FN) denote the ratio that is correctly judged to be 1 in all samples that are actually 1; FPR=FP/(FP+TN), indicating that in all samples that are actually 0, The ratio that is incorrectly judged to be 1. The increase in TPR comes at the expense of an increase in FPR. AUC (Area under roccurve), the area under the ROC curve, is a measure of the accuracy of the model. According to the definition and training data of TPR and FPR, ROC curves can be drawn as shown in figure 3.

Calculate the value of AUC (Area Under the ROC Curve) using the trapezoidal integral method. and the formula is as follows:

$$P\left(F_P\right)=\alpha \tag{1}$$

$$P(T_p)=1-\beta \tag{2}$$

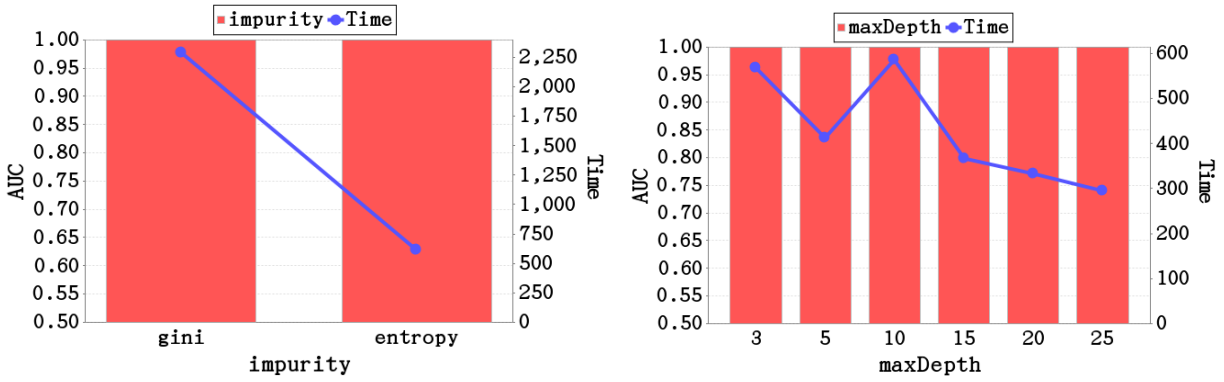
$$AUC=\sum_{i}\left((1-\beta_i \bullet \Delta\alpha) + \frac{1}{2}[\Delta(1-\beta)\bullet\Delta\alpha]\right) \quad (3)$$

Where, the probability of true positive is the probability of false positive, $\Delta\alpha=\alpha_i-\alpha_{i-1}$, $\Delta(1-\beta)=(1-\beta_i)-(1-\beta_{i-1})$, and the evaluation criteria of AUC (area under the ROC curve) are shown in table 2. $P(T_p)$ $P(F_p)$ $\Delta(1-\beta)=(1-\beta_i)-(1-\beta_{i-1})$ $\Delta\alpha=\alpha_i-\alpha_{i-1}$.

Table 2. AUC value setting

condition	explanation
AUC=1	Is the perfect situation
0.5<AUC<1	Better than random guess, with predictive value
AUC=0.5	As with random guessing, there is no predictive value
AUC<0.5	Better than random guess, but better than random guess if reverse prediction

5. Parameter Tuning of the Decision Tree Model



(a) (b)

Fig.4 Impurity parameter adjustment chart

For the Impurity parameter, first fix maxBins=10, maxDepth=10, and compare the parameters of the two parameters gini and entropy for impureness as shown in Fig. 4(a). It can be seen from Fig. 4(a) that the AUC values obtained by the two splitting modes are all 1, but the decision tree of the entropy splitting mode requires only 1/4 of the time of the gini, so the parameter selects the entropy mode. For the maxDepth parameter, the fixed impurity is gini, maxBins=10, and the parameters are compared for maxDepth=3, 5, 10, 15, 20, 25, as shown in Figure 4(b). As can be seen from Fig. 4(b), the obtained AUC values are all 1, but when maxDepth = 20, the required time is the smallest, so the parameter selection maxDepth = 20. Similarly, for the maxBins parameter, the fixed impurityis gini, maxDepth=10, and the parameters are compared for maxBins=3, 5, 10, 50, 100, 200. When maxBins=3, the obtained AUC=0.991, the time cost is also the most, so the parameter maxBins=3 is not selected; when maxBins=10, the time is the least, and the obtained AUC=1, so the parameter selects maxBins=10. Finally, use the program to do the loop, define three arrays, impurityArray, maxdepthArray, maxBinsArray, three Array combinations, a total of 2 * 6 * 6 = 72 sorting combinations, each combination is calculated to save the AUC value, and finally choose the largest AUC A combination is used as a parameter to the model. Its optimal parameters are as follows: impurity use entropy, MaxDepth=20, Maxbins=100 and AUC=1.0. According to the best parameters, AUC=0.9958, and the AUC of the model obtained by tuning is not much different, and it proves that the model does not have the problem of overfitting.

6. Gender Data Identification Test

(1) Recognition results of ordinary voices :(6 male voices, 5 female voices), as shown in table 3. After testing, the accuracy of the ordinary voice recognition can reach 99.9%.

Table 3. Common vocal recognition result

Identification(meanfreq):0.186206843	female
Identification(meanfreq):0.190558485	female
Identification(meanfreq):0.168424394	male
Identification(meanfreq):0.172459445	male
Identification(meanfreq):0.164848568	male
Identification(meanfreq):0.165658876	male
Identification(meanfreq):0.197437534	female
Identification(meanfreq):0.175527574	male
Identification(meanfreq):0.165454626	male
Identification(meanfreq):0.196459219	female
Identification(meanfreq):0.193559744	female

(2) Recognition results of daily speech :(three male voices and one female voice), as shown in table 4. After testing, the accuracy of daily speech recognition can be 90%. Due to the difference in training, there may be recognition errors.

Table 4. Recognition results of Daily speech

Identification(meanfreq):0.154661351	forecast:male
Identification(meanfreq):0.171214516	forecast:male
Identification(meanfreq):0.122191429	forecast:male
Identification(meanfreq):0.181569362	forecast:female

(3) The recognition result of the song (one song is female, one song is male), as shown in table 5. After testing, the accuracy of song recognition is poor.

Table 5. The recognition result of the song

Identification(meanfreq):0.183105907	forecast:female
Identification(meanfreq):0.204811209	forecast:female

(4) Summary of identification results:

The voice data of ordinary vocals is the closest to the 2250 voice samples trained. They are all from the Festvox website, so the recognition accuracy is the highest. The daily voice data is my own recorded audio data, which is different from the trained model. The accuracy rate will be reduced; the song data is popular music downloaded from the Internet, the length of time is generally greater than three minutes, the training data samples are 3 to 5 seconds, the difference is very large, which leads to lower recognition accuracy.

7. Summary

This paper proposes a method for predicting the gender of speech data using a decision tree algorithm. The test decision tree model uses only a small number of single samples as the training set. If you need to apply to a real recognition scenario, the trained dataset samples should be large enough and diverse. The key to using computer to analyze unstructured data is to unstructured data, the

decision tree model proposed in this paper can also be used in other data fields. For example, the meteorological bureau predicts whether it will rain based on environmental condition data.

Acknowledgments

The authors are grateful for The Scientific and Technological Projects of Guangdong Provincial (Grant No. 2017B010126002) and Science and Technology Planning Project of Guangdong Province (Grant No. 2017B010118002).

References

[1]. AJ Hunt, AW Blackl. Unit selection in a concatenative speech synthesis system using a large speech database. Atlanta: IEEE Press. 1996, p. 373-376.

[2]. Meron Y, Hirose K. Efficient weight training for selection-based synthesis. Budapest: ISCA Press, 1999, p. 2319-2322.

[3]. Francesc A, Xavier I. Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis. Euro speech 2003. Geneva: ISCA Press, 2003, p.1333-1336.

[4]. Papk Seung-seop, Kim Chong-kyu, Kim Nam-soo. Discriminative weight training for unit-based selection-based speech synthesis. Euro speech 2003. Geneva: ISCA Press, 2003, p.281-284.

[5]. Breiman. Classification and Regression Trees. Pacific Grove, CA: Wadsworth, 1984.

[6]. Black A, Taylor P. Automatically clustering similar units for unit selection in speech synthesis. Euro speech 97. Rhodes: ISCA Press, 2(1997), p.601-604.

[7]. Kory Becker, http://www.primaryobjects.com/.

[8]. R. K. Zimmerman, G. K. Balasubramani, Mary Patricia Nowalk, et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. BMC Infectious Diseases, 16(2016), p.503-514.

[9]. Takaya Saito, Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Plos one, 2015, p.1-21.

[10]. Hemang Parikh, Marghoob Mohiyuddin, Hugo Y. K. Lam. svclassify: a method to establish benchmark structural variant calls. Parikh et al. BMC Genomics, 17(2016), p.64-80.