Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8872248, 14 pages
https://doi.org/10.1155/2021/8872248

Research Article

Feature Selection Using Maximum Feature Tree Embedded with Mutual Information and Coefficient of Variation for Bird Sound Classification

Haifeng Xu, Yan Zhang, Jiang Liu, and Danjv Lv

College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China

Correspondence should be addressed to Yan Zhang; zydyr@163.com

Received 27 September 2020; Revised 30 December 2020; Accepted 1 February 2021; Published 13 February 2021

Academic Editor: Paolo Spagnolo

The classification of bird sounds is important in ecological monitoring. Although extracting features from multiple perspectives helps to fully describe the target information, it is urgent to deal with the enormous dimension of features and the curse of dimensionality. Thus, feature selection is necessary. This paper proposes a scoring feature method named MICV (Mutual Information and Coefficient of Variation), which uses the coefficient of variation and mutual information to evaluate each feature's contribution to classification. And then, a method named ERMFT (Eliminating Redundancy Based on Maximum Feature Tree) based on two neighborhoods to eliminate redundancy to optimize features is explored. These two methods are combined as the MICV-ERMFT method to select the optimal features. Experiments are conducted to compare eight different feature selection methods with two sounds datasets of bird and crane. Results show that the MICV-ERMFT method outperforms other feature selection methods in the accuracy of the classification and is less time-consuming.

1. Introduction

Birds are sensitive to changes in habitats and surroundings, and they are a good indicator of biodiversity and the ecosystem [1]. Because birds generally have a wide range of movement and cannot be observed promptly, bird sounds are one of the important ways to identify them [2].

Bird sounds are a class of environmental sounds. Some famous feature extraction methods used in audio signal processing include Mel-Frequency Cepstral Coefficients (MFCC) [3] in the frequency domain and Short-Time Fourier Transform (STFT) [4] and Wavelet Transform (WT) in the time domain [5]. Furthermore, Tsau et al. [6] suggested a method that extracts features from Code Excited Linear Prediction (CELP) bit streams. Researchers have been extracting features from multiple aspects to retrieve enough information to describe the target. However, the curse of dimensionality occurs as the numbers of the features and samples grow. It also increases the time cost of analyzing data, affects the models' generalization, and reduces the effectiveness of solving problems [7]. To avoid the curse of dimensionality, selecting a subset of features from the feature pool is necessary.

The feature selection process in pattern recognition is composed of feature scoring and feature optimization. Feature scoring, the key to feature selection, finds the most distinguishable features in the classification space. Generally, feature scoring methods can be grouped into four classes: similarity-based, information-theory-based, statistics-based, and sparse-learning-based [8]. So far, researchers have proposed many different feature scoring methods [9]. For example, in unsupervised feature selection, Nonnegative Laplacian is used to estimate the feature contribution [10]. Constraint Score is applied in feature scoring in environmental sound classification [11]. The ReliefF-based feature selection algorithm is employed to select features in automatic bird species identification [12]. PCA is used as a feature reduction technology to realize bird sounds' automatic recognition [13].

Meanwhile, feature optimization, the second phase of feature selection, selects a subset of features, characterized by low redundancy and high contribution to the classification, from the feature sequence ordered by scores. Filter, Wrapper, and Embedded are three types of methods used to select a subset of features, and many studies have proposed various feature optimization algorithms based on these methods. Binary Dragonfly Optimization Algorithm, PSO (Particle Swarm Optimization), and Artificial Bee Colony are some examples. Specifically, S-shaped and V-shaped transfer function can be used to map continuous search space to discrete search space [14]. Mutual information can be combined with PSO to eliminate redundant features [15]. In some research, the gradient enhanced Decision Tree [16] is used to evaluate feature contribution, and Artificial Bee Colony is applied to optimize the features [17]. Pearson correlation coefficient is a common evaluation metric used in literature, which evaluates the correlation between features, and is followed by Artificial Ant Colony to select high-quality features [18].

Most feature scoring methods, such as Constraint Score and Laplacian, are based on the correlation and differences among spatial distances between features. Although these algorithms have low time complexity, the diversity of the features is neglected. Specifically, units of the features are usually different. Some algorithms calculate the mutual information between the feature sample and the label from a probabilistic and statistical perspective [15]. However, the label is generally a discrete variable, while features are continuous variables. In recent years, many studies regard feature selection as an optimization process and combine feature selection with intelligent searching methods [9, 19–22]. The multiobjective optimization problem of a large dataset has a high time and space complexity. A reduction in the features' dimensions usually decreases in the classification model's sensitivity and generalization.

Regarding the issues mentioned above, from an information theory perspective, this paper proposes a feature scoring method MICV (Mutual Information and Coefficient of Variation). MICV utilizes the characteristics of mutual information and coefficient of variation and aims to minimize intraclass distance and maximize interclass distance. A feature optimization method, ERMFT (Eliminating Redundancy Based on Maximum Feature Tree), is suggested based on a minimum spanning tree concept. Experiment results show that the MICV-ERMFT method can effectively reduce the data dimension and improve the classification model's performance. Compared with eight feature evaluation methods, the MICV-ERMFT method has significant improvement in the performance on the same dataset in this paper.

2. Materials and Methods

In bird sounds' recognition, there exists a variety of methods to extract features and classify the sounds. For example, Human Factor Cepstral Coefficients are used to extract bird sound features, and classification and recognition are performed by the maximum likelihood method [23]. Zottesso et al. [24] suggest a method that extracts bird song features based on the spectrogram and texture descriptors and uses the dissimilarity framework for classification and recognition. In this paper, the classification process of bird sounds is divided into three stages: feature extraction, feature selection, and classification recognition. Feature selection is selected as the research focus. The proposed classification process of bird sounds based on MICV-ERMFT is shown in Figure 1:

Stage 1. Preprocess the bird sounds' audio data (remove noises and converse the channel), and use MFCC and CELP to extract features from the preprocessed data and construct dataset $D_{\mathrm{M\&C}}$ (dataset formed by the merger of MFCC and CELP features).

Stage 2. Apply the MICV method on $D_{\mathrm{M\&C}}$, evaluate the contribution, and score each feature. Sort the feature sequence in ascending order, denote as F, and calculate the Pearson correlation coefficient for the features and build a maximum feature tree T. Then, apply the ERMFT method to eliminate redundant features and construct a new dataset $D_{\mathrm{M\&C}'}$.

Stage 3. Build a classification model on $D_{\mathrm{M\&C}'}$ and analyze the classification results.

2.1. Feature Extraction. Birds make sounds in the same way as humans do [25, 26]. The frequency of human language used for daily communication ranges from 180 Hz to 6 kHz, and the most used frequency range for bird calls is from 0.5 to 6 kHz [25, 27]. Under this assumption, we process the features of bird sounds in a way similar to processing the human language. MFCC (Mel-Frequency Cepstral Coefficient) and CELP (Code Excited Linear Prediction) are applied to the raw bird sounds data to extract features in this paper.

2.1.1. MFCC. MFCC [3] is a human-hearing-based, nonlinear feature extraction method. The process is shown in Figure 2.

Step 1. A single-frame, short-term signal $x_w(i,n)$ is obtained by separating frames and adding a window function to the original audio signal $x(n)$. Adding a window function reduces the frequency spectrum leakage. This paper selects the 20 s as a frame and uses the Hamming window.

Step 2. To observe the distribution of $x_w(i,n)$ in frequency domain, FFT (fast Fourier transform) is used to transform the signal from the time domain to frequency domain, named $X(i,k)$:

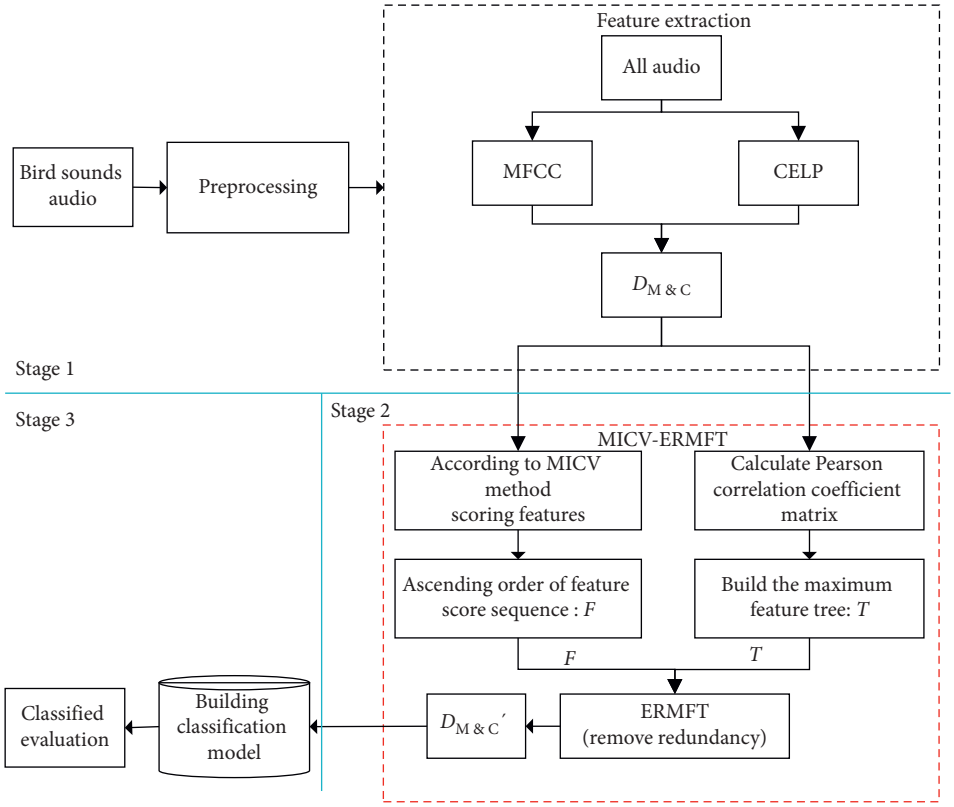

Figure 1: Bird sound classification model based on MICV-ERMFT feature selection.

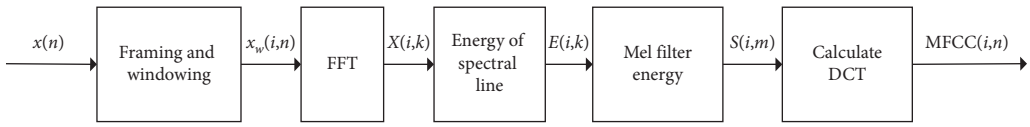

Figure 2: MFCC schematic.

$$X(i,k) = \mathrm{FFT}[x_i(m)]. \tag{1}$$

Step 3. Calculate the energy of the spectral line per frame:

$$E(i,k) = [X(i,k)]^2. \tag{2}$$

Step 4. Calculate the energy of $E(i,k)$ through the Mel filter:

$$S(i,m) = \sum_{k=0}^{n-1} E(i,k) H_m(k), \quad 0 \le m < M, \tag{3}$$

where i is the i-th frame, k is the k-th spectral line in the spectrum, and $H_m(k)$ is the analysis window with a sample length of k;

Step 5. Take the logarithm of the energy of the Mel filter and calculate the DCT (Discrete Cosine Transform):

$$\mathrm{mfcc}(i,m) = \sum_{m=0}^{M-1} \log[S(i,m)] \cos\left(\frac{\pi n(2m-1)}{2M}\right), \tag{4}$$

where m is the m-th Mel filter, i is the i-th frame, and n is the spectral line after the DCT.

In this paper, MFCC uses 13-dimensional static coefficients (1-dimensional log energy coefficient and 12-dimensional DCT coefficients) as extraction parameters [3, 28]. The resulting sample has 13 features.

2.1.2. CELP. The CELP feature extraction method is derived from LPC (Linear Predictive Coding) based on a compression coding tech G.723.1. The LPC is extracted from the 0th to 23rd bits from the bit coding in each frame, forming the 10-dimensional LPC. Another 2-dimensional feature, the lag of pitch, is extracted from the 24th to the 42nd bit stream in each frame. The extraction of CELP is shown in Figure 3.

Endpoint detection is performed after the original audio file is preprocessed. Then each audio is divided into several sound segments. Each sound segment is considered as a sample in the experiment. For each frame, features are extracted using MFCC (13 dimensions) and CELP (12 dimensions). The sampling rate is 16 kHz; audio is a single channel. Each sample contains several frames. For each

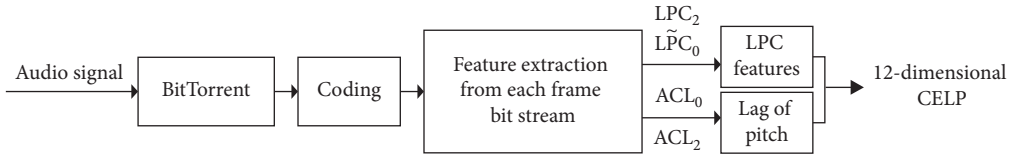

Figure 3: CELP feature extraction process.

detection segment (including many frames), the mean, median, and variance of each feature are calculated to obtain 75-dimensional data. The feature extraction process is shown in Figure 4.

2.2. Feature Scoring Method MICV. Based on the principle of small distance within classes and large distance between classes, features that are easy to distinguish are selected. To calculate the degree of feature differentiation, mutual information MIEC (Mutual Information for Interclass) is used to measure the interclass distance, and the coefficient of variation CVAC (Coefficient of Variation for Intraclass) is used to measure the intraclass distance.

The MIEC and CVAC methods are combined to calculate the classification contribution degree of features. The calculation equation is

$$\text{micv}_f = \lambda\,\text{miec}_f + (1-\lambda)\text{cvac}_f. \tag{5}$$

Because intraclass distance and interclass distance have different weights, the coefficient $\lambda\,(0<\lambda<1)$ is introduced to adjust the weights.

2.2.1. MIEC. Mutual information measures the correlation or the dependency between two variables. For two discrete random variables X and Y, mutual information $I(X;Y)$ is calculated as

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{6}$$

In equation (6), $p(x,y)$ is the joint probability density function of x and y, $p(x)$ and $p(y)$ are the marginal probability density functions of x and y.

Generally, when mutual information is used to select features, variables X and Y represent the feature vector and label vector. In this paper, X and Y represent two vectors of different classes under the same feature. Given feature space F and classification space C, the interclass mutual information of f-th feature, miec_f, is calculated as

$$\text{miec}_f = \sum_{i\in C}\sum_{j\in\{C\setminus i\}} I(i;j). \tag{7}$$

In equation (7), i and $j\,(i\neq j)$ are the samples of f-th feature in i-th class and j-th class. miec_f is the interclass mutual information of f-th feature in F. The interclass difference feature f is greater when the miec_f is smaller, and vice versa.

2.2.2. CVAC. In statistics, the variation (CV) coefficient measures the variation between two or more samples or the dispersion between them. The expression is

$$Cv = \frac{\sigma}{\mu}, \tag{8}$$

where μ and σ are the mean and standard deviation of the samples. Given feature space F and classification space C, the intraclass coefficient of variation of feature f, cvac_f is calculated as

$$\text{cvac}_f = \sum_{i=1}^{C} Cv_i. \tag{9}$$

In equation (9), Cv_i represents the CV of samples in class i. The feature f has a higher cohesion when cvac_f is smaller.

2.3. Feature Selection Method MICV-ERMFT. After scoring the features using the MICV method, high-quality features are selected. MICV-ERMFT is used to eliminate redundant features in the feature array sorted by scores. The process is shown in Algorithm 1.

2.3.1. Build Maximum Feature Tree. The maximum feature tree is derived from the minimum spanning tree. For an undirected graph $G(V,E)$, each edge has a weight w, a minimum spanning tree is a subset of edges E' that connect all the vertices V with no cycle, and the total weight of edges in E' is minimum. In a maximum feature tree, features are represented as vertices and weights of the edges are decided by Pearson correlation coefficient. $P_{(F_r,F_c)}$ represents the correlation coefficient between features F_r and F_c, which is calculated as

$$P_{(F_r,F_c)} = \frac{\sum_i (F_{ri}-F_r)(F_{ci}-F_c)}{\sqrt{\sum_i (F_{ri}-F_c)^2}\sqrt{\sum_i (F_{ci}-F_c)^2}}, \tag{10}$$

$$I_{(F_r,F_c)} = \frac{-\log_2\left(1-P^2_{(F_r,F_c)}\right)}{2}. \tag{11}$$

In equation (10), F_{ri} represents the i-th sample of feature r; F_r is the feature r's mean value of all samples. In equation (11), $I_{(F_r,F_c)}$ is the correlation coefficient between features r and c. Algorithm BMFT (building the max feature tree) uses equations (10) and (11) to calculate the correlation coefficient matrix and construct the maximum feature tree. Details are described in Algorithm 2.

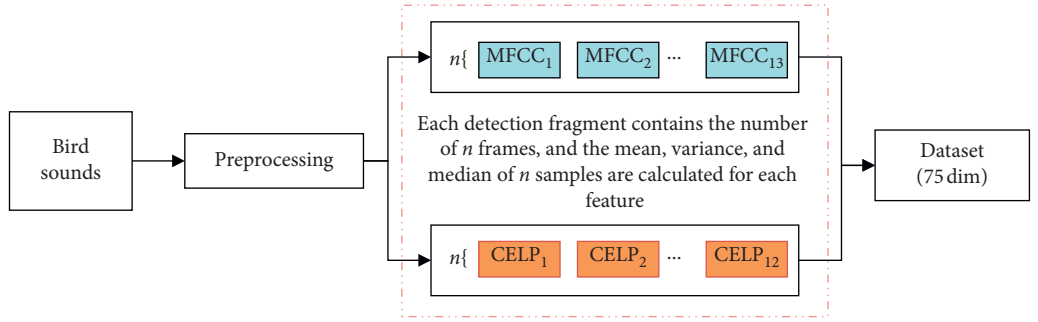

Figure 4: Extraction process of bird sounds feature.

2.3.2. Remove Redundant Features Based on Two Neighborhoods. ERFTN (Eliminate Redundant Features based on Two Neighborhoods) is based on eliminating redundancy using the concept of two neighborhoods. One example with a maximum feature tree T and feature sequence F sorted using the MICV method is demonstrated in Figure 5:

As shown in Figure 5, given max feature T, $F = \{f_2, f_1, f_3, f_4, f_5, f_7, f_9, f_8, f_6, f_{10}\}$ sorted with MICV method in ascending order, the steps of the ERFTN algorithm are listed as Algorithm 3. The final feature subset of F is $\{f_2, f_3, f_7, f_{10}\}$.

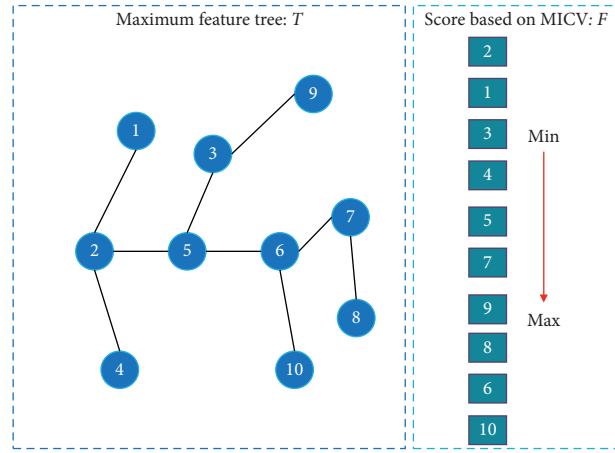

Figure 5: Schematic of the ERFTN.

3. Experiments and Results Analysis

3.1. Experimental Dataset. Currently, there are many websites dedicated to sharing bird sounds from around the world, such as Avibase [29] and Xeno-Canto [30]. Recordings of bird sounds are collected and annotated on these websites. The tapes include various types of voice expressions (multiple calls and songs) of various individuals recorded in their natural environment. The dataset used for this paper comes from the Avibase, which is a collection of MP3 or WAV audio files. These audio files are unified into the 16 kHz sampling rate and monochannel. Since the audio files are not all bird sounds, the bird sounds in the audio are separated through the voice activity detection (VAD) [25, 31], and then the MFCC and CELP features are extracted according to the process shown in Figure 4.

The experiments used two datasets including bird sounds and crane sounds. We have selected six different bird species from different genera in bird sounds, which contains 433 samples. The crane sound dataset includes 343 samples from seven species of *Grus*. The dataset information is shown in Tables 1 and 2.

3.2. The Experiment of MICV Scoring Method. To verify the proposed method's effectiveness, two separate experiments are conducted to test the MICV scoring method and MICV-ERMFT feature selection method. The classifiers used in the experiments include Decision Tree (J48), SVM, BayesNet (NB), and Random Forests (RFs). The feature scoring method is compared with ConstraintScore (CS) [11] and six other feature scoring methods provided by Weka [32] including Correlation (Cor), GainRatio (GR), InfoGain (IG), One-R (OR), ReliefF (RF), and SymmetricalUncert (SU) in experiments.

3.2.1. Classifier Performance Evaluation. Kappa, F_1 score, and accuracy rate were used as evaluation indicators.

(1) Kappa. Cohen's Kappa coefficient is a statistical measure that indicates the interrater reliability (and also intrarater reliability) for qualitative (categorical) items:

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e}, \tag{12}$$

where p_o is the overall classification accuracy, which is calculated by the number of correctly classified samples divided by the total number of samples. Based on the confusion matrix, assume the numbers of real samples in each class are $\{a_1, a_2, \ldots, a_n\}$, the numbers of predicted samples are $\{b_1, b_2, \ldots, b_n\}$, and p_e is calculated as

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \cdots + a_n \times b_n}{n * n}. \tag{13}$$

(2) F_1 Score. It is an index used to measure the accuracy of classification models in statistics, while taking into account the accuracy and recall of classification models. As shown in equation (14), precision represents the precision rate and recall represents the recall rate.

Name: MICV-ERMFT feature selection.
Input: Dataset D (m: number of samples; n: number of features).
Step:

(1) Calculate MICV using equation (5) for each feature in D.
(2) Sort the MICV feature sequence in ascending order to obtain F. According to F, select data in D and gradually add one, and use the base classifier to score. Delete the feature that led to the decline of the index, and obtain the feature sequence F^*, and map F^* to D to get dataset D^*.
(3) Calculate Pearson correlation coefficient matrix P for the feature vector by D^*.
(4) Apply algorithm **BMFT** (Algorithm 2) **t**o construct a maximum feature tree T for P.
(5) Apply two-neighborhood based redundancy eliminating algorithm **ERFTN** (Algorithm 3) on F^*; denote the result array as F^{**}.
(6) Map F^{**} to D^* to get dataset D^{**}.

Output: new dataset D^{**}.

Algorithm 1: MICV-ERMFT feature selection.

Name: BMFT (building max feature tree).
Input: Correlation coefficient matrix $P_{n\times n}$ #n is the number of features.
Step:

(1) Initialize root $T = \{1\}$.
(2) Set the elements on the main diagonal as −1. # set the value as −1 on the main diagonal to eliminate the influence from the feature itself.
(3) **while** $|T| \le 2n + 1$ **do** /*Since P is a Strongly Connected Graph, and T records the adjacency relationship of the elements in P, there are $2n$+1 elements in T including the initial node. */
(4) $D = P_{(1:\, n,T)}$ /*D records the correlation coefficients of the neighboring nodes of the nodes in T (D is a column vector mapped from T to P) */
(5) $D_{(T,1:\, n)} = -1$ /* The nodes that have been visited are recorded in T. This operation is equivalent to deleting the accessed data in D. */
(6) $\text{row}_{\text{id}} = \text{FindIndex}[\max(D)]_{\text{row}}$ #Find the maximum value of all nodes adjacent to visit_id and record the row index.
(7) $T = T \cup \{T_{\text{end}}, \text{row}_{\text{id}}\}$/*{$T$_end, row_id} records the adjacency relationship, for example, T_end and row_id are adjacent nodes */
(8) **end while**
(9) **return** T

Output: Maximum feature tree T.

Algorithm 2: BMFT.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{14}$$

(3) Accuracy. The accuracy is calculated based on the equation:

$$\text{accurary} = \frac{n}{M}. \tag{15}$$

In equation (15), n represents the correct number of classifications, and M represents the number of all samples.

Each dataset is divided into 70% training set and 30% test set. Each experiment is repeated 10 times to average some biased results.

3.2.2. MICV λ Parameter Setting. In equation (5), use λ to adjust the weight coefficients of MIEC and CVAC. The experiments set $\lambda \in \{0.1, 0.2, 0.3, 0.4, .5, 0.6, 0.7, 0.8, 0.9\}$ and calculate the MICV with J48 classifier. When the highest Kappa is reached, the ratio of the number of selected features to the total features is listed in Table 3. A lower ratio indicates a better performance. Table 3 shows that better results can be obtained when λ is set at 0.1 or 0.3 or 0.2. In the following experiments in this paper, λ is set to 0.1.

3.2.3. Compare MIEC, CVAC, and MICV. The selected feature set has a decisive effect on the classification model. Features with higher scores normally lead to more positive classification performance. The experiments sort the feature sequence in ascending order according to feature scores obtained from MIEC, CVAV, and MICV, respectively. In Figure 6, in most cases, the red curves are more stable to ascend, which shows that, with the increase of features gradually, the classification model's performance will be improved, especially in Figure 6(a). CVAC and MIEC methods have obvious fluctuations in Figures 6(a) and 6(b). To sum up, combining MIEC and CVAC works better than using them alone.

Name: ERFTN (Eliminate Redundant Features based on Two Neighborhoods).
Input: T: Max feature Tree by Algorithm BMFT, F: Features sorted with MICV method.
Step:

(1) Get the first element x in F.
(2) $V = \{y | y \in T, y \text{ is the adjacent vertices of } x\}$.
(3) Update F by deleting all vertices in V, that is $F = F \backslash V$.
(4) Choose the next unvisited element as x.
(5) Repeat (2) to (4), until all the elements in F are visited.
(6) Output F as the final feature subset.

Output: F.

Algorithm 3: ERFTN.

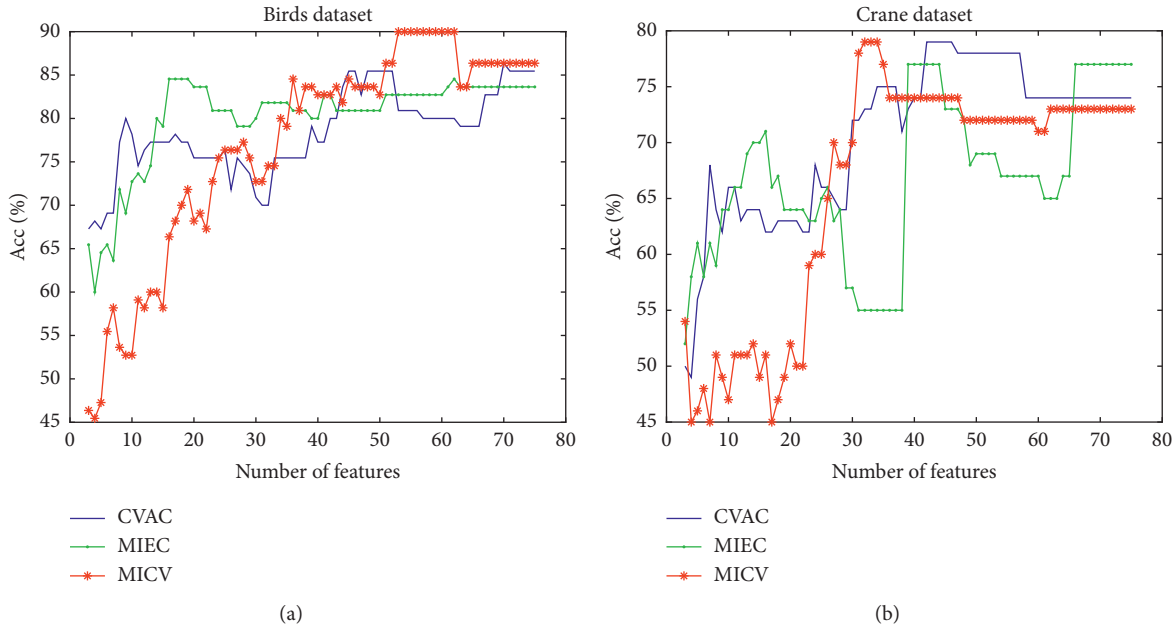

Figure 6: Experimental results of MIEC, CVAC, and MICV feature selection methods: (a) Birds dataset and (b) Crane dataset.

3.2.4. Experiment of MICV Results and Analysis. In this section, the proposed MICV is tested on the Birds dataset and the Crane dataset. The results of the experiment in Figures 7 and 8, show that, at the same number of selected features, the Kappa value of the MICV method is basically higher than that of other methods. As the number of features increases, the Kappa value of the MICV method can converge earlier and remains relatively stable compared with other methods. MICV is more effective compared with the results of other feature evaluation methods.

Tables 4 and 5 record the best classification results (Kappa, accuracy, and F_1 scores) for each feature scoring sequence, as well as the number of features used to obtain this value. The bold one on the left side of "|" in each row in the table indicates that the method has the least number of features than other methods, and the bold on the right indicates that the method has the highest evaluation indicator score. Table 4 shows that, in bird dataset, MICV methods had the highest Kappa value under four different classifiers. In J48, NB, and RFs classifiers, MICV methods had the lowest number of features and the highest score of evaluation indicators in most cases. As shown in Table 5, the performances of MICV in J48, NB, and RFs classifiers are significant.

In summary, the MICV method is more effective in selecting optimal features than the other seven methods. The method can also get a good modeling effect by using a lower dimension.

3.3. Experiment of MICV-ERMFT Feature Selection. In the second part of the experiment, features are evaluated using CS

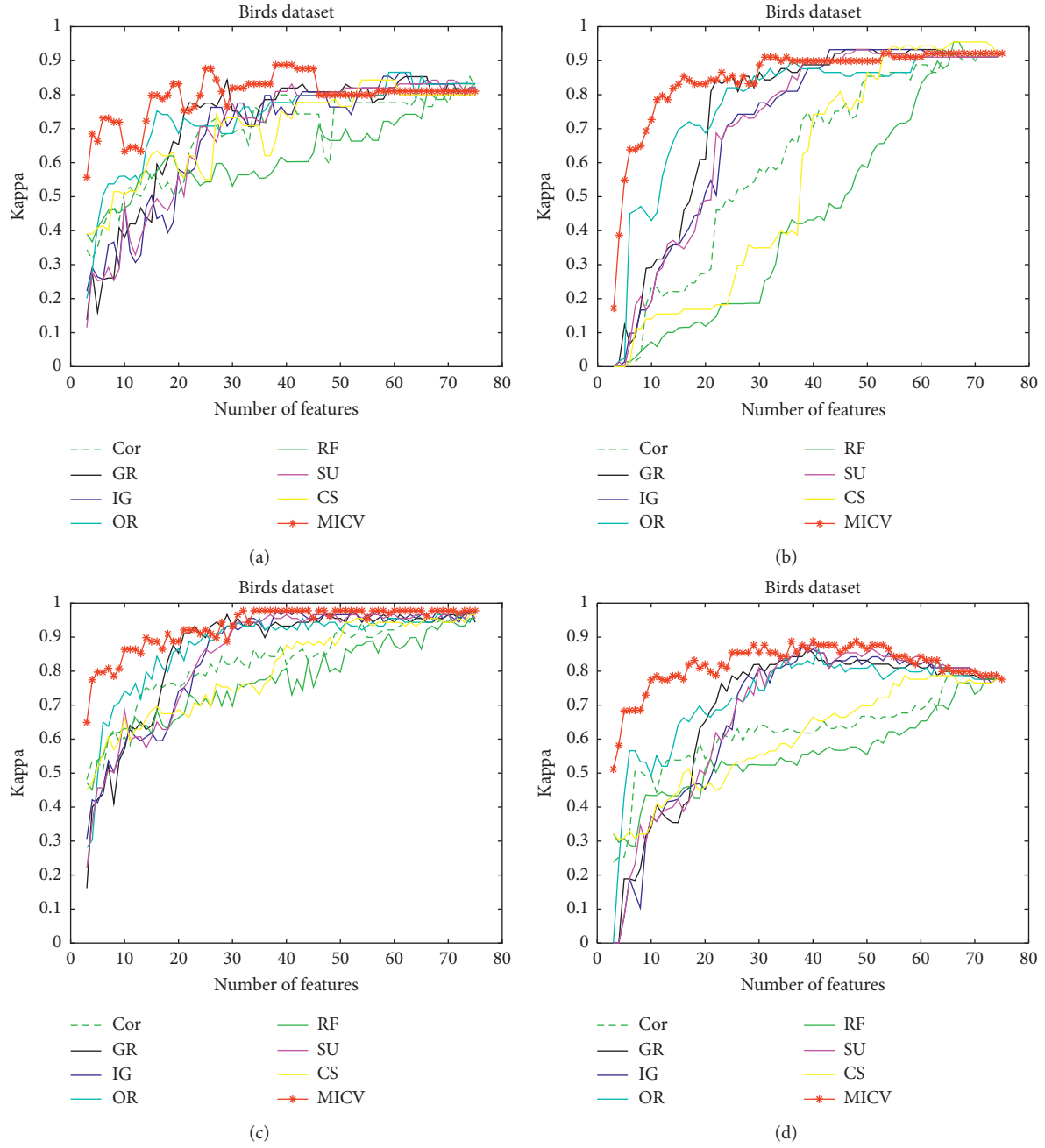

Figure 7: Experimental results of MICV and other 7 feature evaluation methods in different classifiers (Birds dataset). (a) J48. (b) SVM. (c) RFs. (d) NB.

and six other Weka methods, including Cor, GR, IG, OR, RF, and SU.

3.3.1. Procedure of Experiment. The procedure is demonstrated in Figure 9. Eight different methods (MICV and the seven other methods mentioned above) are used to evaluate each feature's classification contribution and score the features. After sorting the features in an ascending order based on the scores, the ERMFT method is then used to eliminate redundant features, resulting in a feature subset F′. F′ is then mapped to Dataset, resulting in Dataset'. J48, SVM, BayesNet (NB), and Random Forests (RFs) are the experiment's classifiers. For each independent dataset, it is divided into 70% training set and 30% test set. Each experiment is repeated ten times, and the average Kappa is calculated. Also, the DRR (Dimensionality Reduction Rate) as an evaluation indicator is introduced.

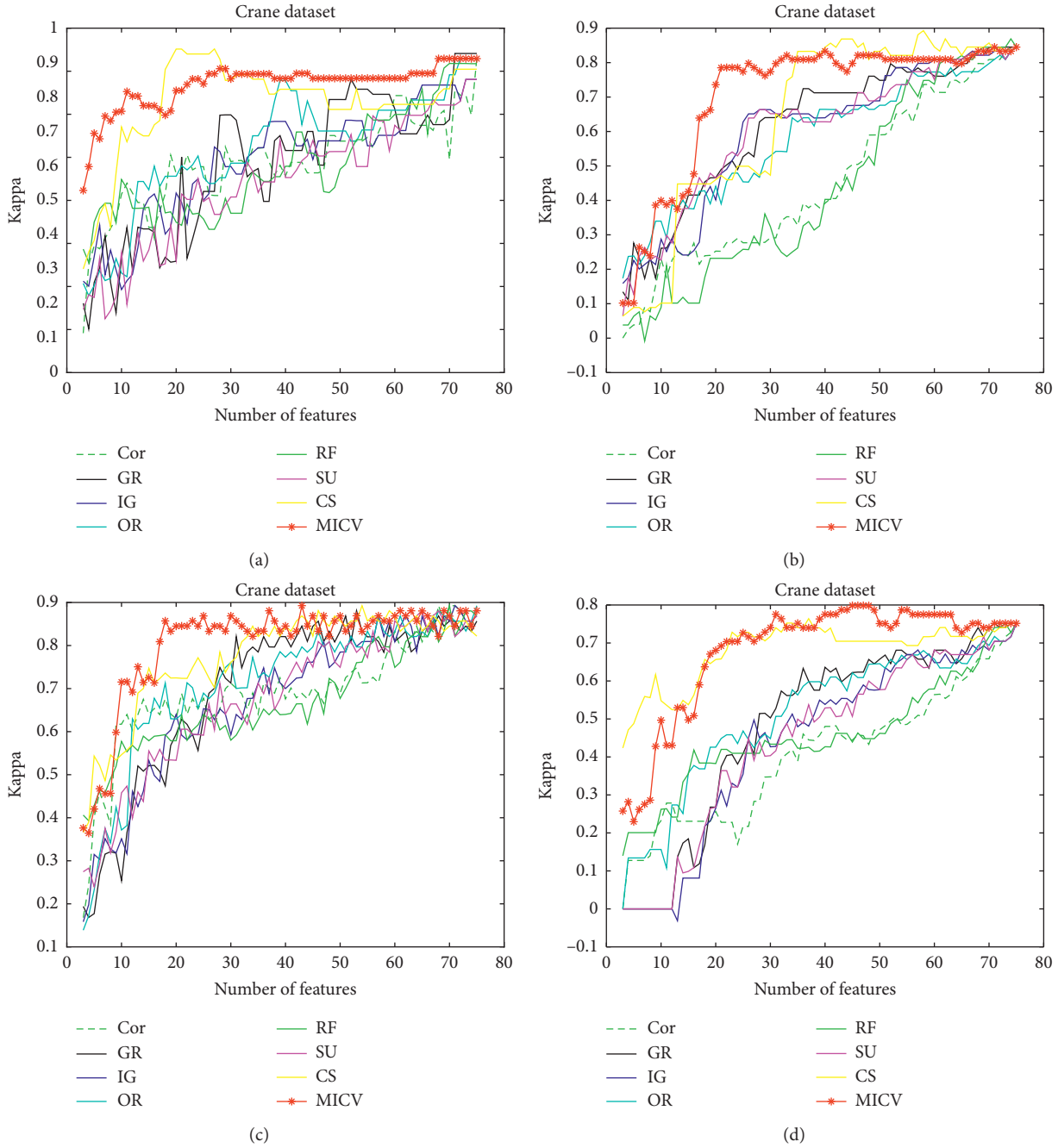

Figure 8: Experimental results of MICV and other 7 feature evaluation methods in different classifiers (Crane dataset). (a) J48. (b) SVM. (c) RFs. (d) NB.

Table 1: Bird sound dataset information.

Latin name	Eng. name	Genus	Number of samples	Rate
Phalacrocorax carbo	Great cormorant	*Phalacrocorax*	36	8.31
Numenius phaeopus	Whimbrel	*Numenius*	90	20.79
Aegithina nigrolutea	White-tailed iora	*Aegithina*	120	27.72
Chrysolophus amherstiae	Lady Amherst's	*Chrysolophus*	68	15.70
Falco tinnunculus	Common kestrel	*Falco*	61	14.09
Tadorna ferruginea	Ruddy shelduck	*Tadorna*	58	13.39

Table 2: Crane sounds dataset information.

Latin name	Eng. name	Genus	Number of samples	Rate (%)
Grus vipio	White-naped crane		24	7.00
Grus canadensis	Sandhill crane		39	11.37
Grus virgo	Demoiselle crane		60	17.49
Grus grus	Common crane	*Grus*	62	18.08
Grus monacha	Hooded crane		62	18.08
Grus japonensis	Red-crowned crane		29	8.45
Grus nigricollis	Tibetan crane		67	19.53

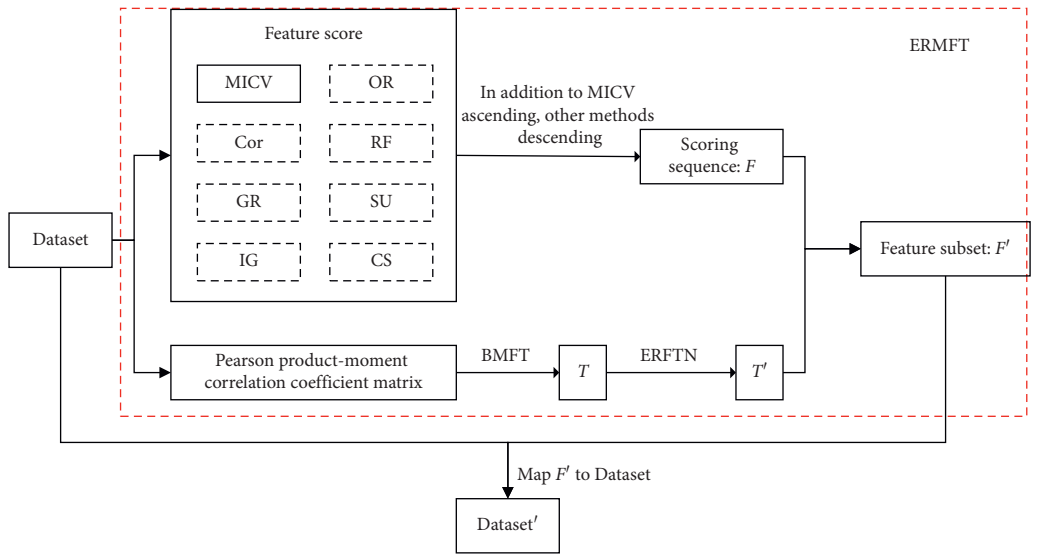

Figure 9: Flowchart of experiment of MICV-ERMFT feature selection.

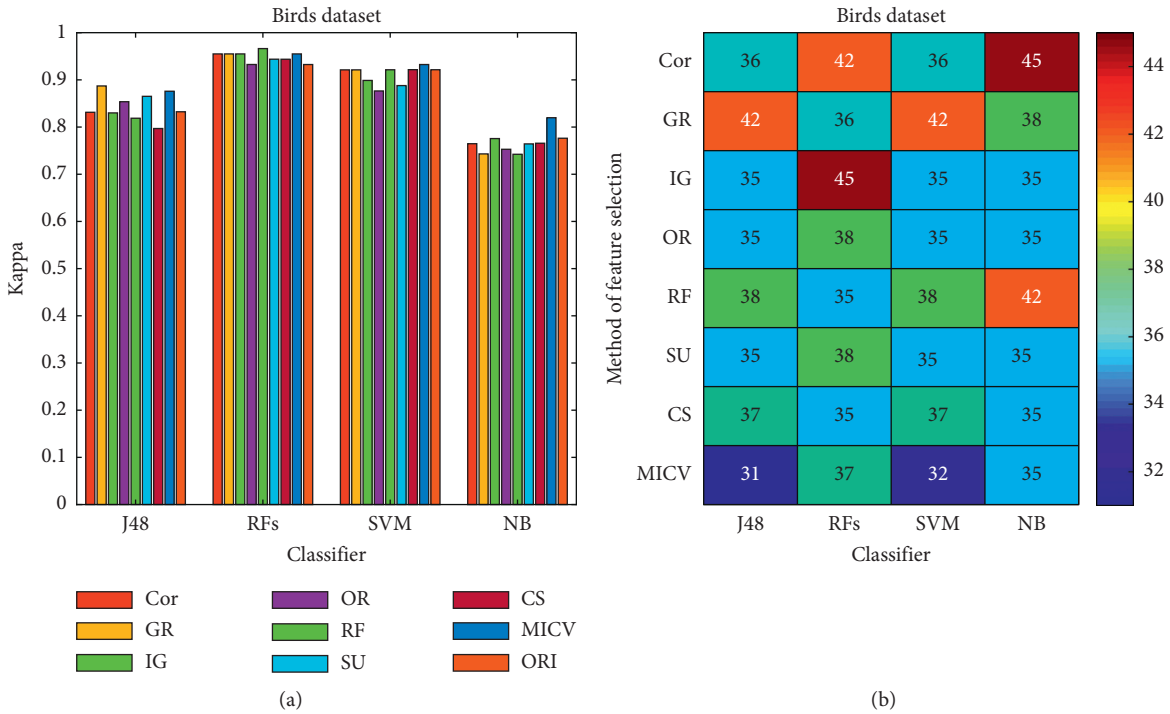

Figure 10: Continued.

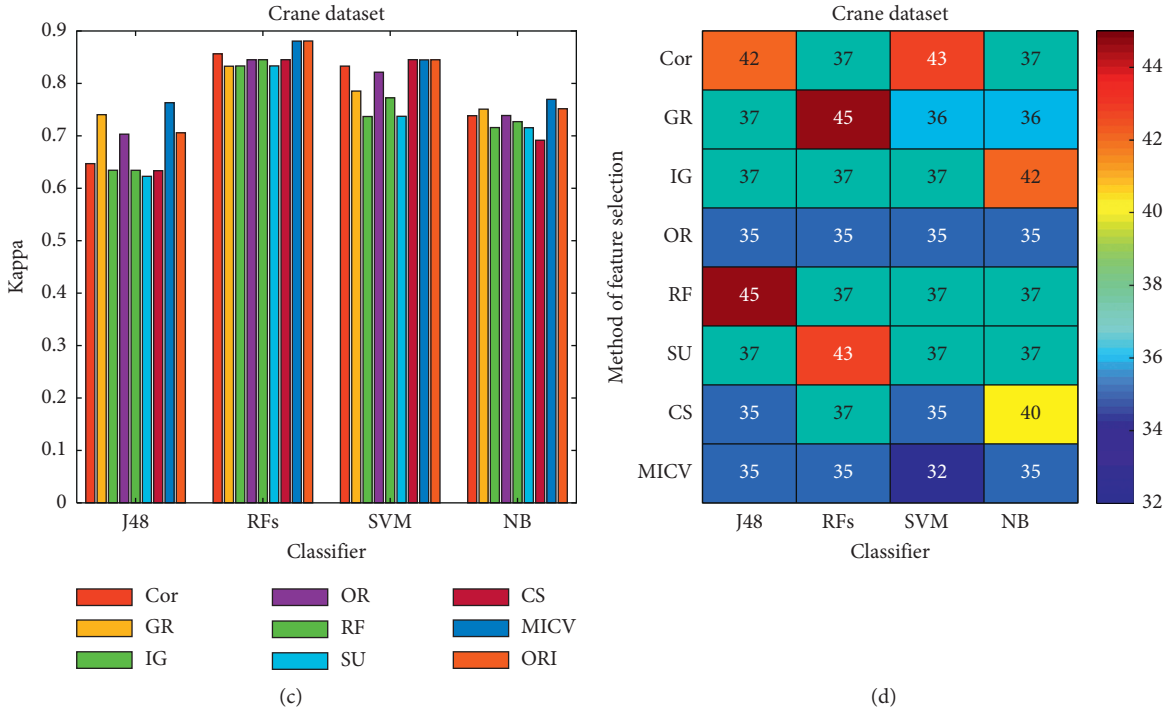

Figure 10: Experimental results of MICV-ERMFT method. (a) Kappa obtained with different classifiers by using different feature evaluation methods in Birds dataset. (b) Heat map of selected feature in (a). (c) Kappa obtained with different classifiers by using different feature evaluation methods in Crane dataset. (d) Heat map of selected features in (c).

Table 3: The ratio of the number of selected features with different values of λ.

Dataset	λ								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Birds dataset	**0.293**	**0.293**	0.466	0.453	0.493	0.440	0.480	0.666	0.560
Crane dataset	0.240	0.360	**0.160**	0.333	0.706	0.506	0.826	0.933	0.866

$$\mathrm{DRR} = \left(1 - \frac{F_n'}{F_n}\right) * 100\%. \tag{16}$$

In equation (16), F_n' is the number of selected features and F_n is the number of all features of each dataset. The larger the DRR value, the stronger the ability to reduce dimensions.

3.3.2. Experiment of MICV-ERMFT Results and Analysis. Figure 10 shows the experimental results obtained from four different classifiers using eight different feature evaluation methods combined with ERMFT. Figures 10(a) and 10(b) are the results of the Birds dataset; Figures 10(c) and 10(d) are the results of the Crane dataset. In Figure 10(a), four histograms represent the results under the four classifiers, and 9 elements in the group of histograms are Kappa values calculated from the eight methods with ERMFT and the original data (ORI). The heat map of Figure 10(b) shows the number of selected features when the Kappa reaches a certain value in each method and similarly so do Figures 10(c) and 10(d). In Figure 10(a), it can be clearly observed that the MICV-ERMFT method has a slightly higher Kappa than other methods and the J48 classifier in Figure 10(c) is more effective. Besides, the Kappa of the MICV-ERMFT method is higher than the original data. Looking at Figures 10(a) and 10(b) at the same, it is evident that the MICV-ERMFT method achieves a good modeling effect using a small number of features' time, comparing with the other methods. Figures 10(c) and 10(d) show a similar result.

In conclusion, compared with the other seven methods, the MICV-ERMFT method demonstrates good abilities in dimensionality reduction and feature interpretation.

Combining Figures 8(b) and 8(d) with Table 6, it is obvious that the MICV-ERMFT method has a significant dimensionality reduction effect and model performance effect for the Birds dataset and the Crane dataset. In Table 6, Kappa value and DRR performance are very good for J48, NB, and SVM classmates on Birds dataset. Particularly for the NB classifier, the other seven comparison methods' Kappa value does not exceed ORI, while the MICV-ERMFT method exceeds 0.4. In the Crane dataset, the MICV-ERMFT outperforms other methods. Table 7 shows the

Table 4: Comparison of Kappa, accuracy, and F_1 scores with feature selection methods in Birds dataset.

Evaluation indicator	Classifier	Feature selection method: Cor	GR	IG	OR	RF	SU	CS	MICV
		Number of features|the highest value							
Kappa	J48	72 | 0.85	60 | 0.85	59 | 0.86	57 | 0.86	73 | 0.83	68 | 0.84	51 | 0.84	**38 | 0.88**
	NB	68 | 0.79	37 | 0.86	36 | 0.87	39 | 0.86	70 | 0.77	**36** | 0.87	58 | 0.79	46 | **0.88**
	SVM	63 | 0.93	**44** | 0.93	52 | 0.93	61 | 0.93	64 | 0.95	46 | 0.93	65 | 0.95	51 | **0.95**
	RFs	71 | 0.97	50 | 0.97	36 | 0.97	73 | 0.97	72 | 0.95	70 | 0.97	53 | 0.96	**30 | 0.97**
Accuracy	J48	72 | 88.18	60 | 88.18	59 | 89.09	57 | 89.09	73 | 86.36	65 | 87.27	52 | 87.27	**36 | 90.96**
	NB	64 | 83.63	37 | 89.09	36 | 88.18	39 | 89.09	70 | 81.81	36 | 90.00	58 | 83.63	**34 | 90.90**
	SVM	63 | 94.54	44 | 94.45	**41** | 94.00	61 | 94.45	64 | 96.3**6**	46 | 94.45	65 | **96.36**	51 | 93.63
	RFs	71 | 98.12	50 | 98.12	36 | 98.12	73 | **98.18**	72 | 96.36	70 | 98.12	53 | 97.27	**30** | 98.12
F_1 score	J48	72 | 0.88	60 | 0.88	59 | **0.89**	57 | 0.89	73 | 0.86	65 | 0.87	51 | 0.87	**38** | 0.87
	NB	64 | 0.83	37 | 0.89	36 | 0.89	39 | 0.89	70 | 0.81	37 | 0.90	58 | 0.83	**34 | 0.90**
	SVM	63 | 0.94	49 | 0.94	**41** | 0.94	61 | 0.94	64 | **0.96**	46 | 0.94	65 | 0.96	51 | 0.93
	RFs	71 | 0.96	50 | 0.98	55 | 0.98	73 | 0.98	73 | 0.96	70 | 0.98	54 | 0.97	**30 | 0.98**

Table 5: Comparison of Kappa, accuracy, and F_1 scores with feature selection methods in Crane dataset.

Evaluation indicator	Classifier	Feature selection method: Cor	GR	IG	OR	RF	SU	CS	MICV
		Number of features|the highest value							
Kappa	J48	73 | 0.72	69 | 0.74	71 | 0.68	70 | 0.72	68 | 0.71	71 | 0.69	**22** | 0.68	25 | **0.75**
	NB	73 | 0.75	69 | 0.75	73 | 0.75	73 | 0.75	71 | 0.75	73 | 0.75	53 | 0.79	**43 | 0.79**
	SVM	73 | 0.84	69 | 0.84	73 | 0.84	73 | 0.84	72 | 0.86	73 | 0.84	73 | 0.84	**69 | 0.84**
	RFs	66 | 0.89	51 | 0.88	69 | 0.89	73 | 0.89	68 | 0.89	63 | 0.88	**36** | 0.90	41 | **0.90**
Accuracy	J48	73 | 77.00	69 | 78.00	71 | 73.00	70 | 77.00	68 | 76.00	71 | 73.00	22 | 73.00	**18 | 79.00**
	NB	73 | 79.00	69 | 79.00	73 | 79.00	73 | 79.00	71 | 79.00	73 | 79.00	53 | 81.00	**43 | 83.00**
	SVM	72 | 87.00	**68** | 87.00	73 | 87.00	73 | 87.00	72 | **89.0**0	73 | 87.00	73 | 87.00	69 | 87.00
	RFs	66 | 91.00	51 | 90.00	69 | 91.00	73 | 91.00	58 | 91.00	63 | 90.00	**36** | 90.00	41 | **91.00**
F_1 score	J48	73 | 0.77	69 | 0.78	71 | 0.73	70 | 0.77	67 | 0.77	71 | 0.73	25 | 0.73	**25 | 0.79**
	NB	73 | 0.79	69 | 0.79	73 | 0.79	73 | 0.79	73 | 0.79	73 | 0.79	53 | 0.81	**43 | 0.82**
	SVM	72 | 0.86	**68** | 0.87	73 | 0.86	73 | 0.86	72 | **0.88**	73 | 0.86	73 | 0.87	69 | 0.86
	RFs	69 | 0.91	51 | 0.89	69 | 0.91	73 | 0.90	68 | 0.90	66 | 0.90	72 | 0.91	**41 | 0.91**

Table 6: MICV-ERMFT compared to other methods of Kappa and DRR.

Dataset	Classifier	Method: Cor	GR	IG	OR	RF	SU	CS	MICV	ORI
		Kappa|DRR (%)								
Birds	J48	0.83 | 52	**0.88** | 44	0.83 | 53	0.85 | 53	0.81 | 49	0.86 | 53	0.79 | 50	0.87 | **58**	0.83 | 0
	NB	0.76 | 40	0.74 | 49	0.77 | 53	0.75 | 53	0.74 | 44	0.76 | 53	0.76 | 53	**0.81 | 53**	0.77 | 0
	SVM	0.92 | 52	0.92 | 44	0.89 | 53	0.87 | 44	0.92 | 53	0.88 | 53	0.92 | 53	**0.93 | 57**	0.92 | 0
	RFs	0.95 | 44	0.95 | 52	0.95 | 40	0.93 | 49	**0.96 | 53**	0.94 | 49	0.94 | 53	0.93 | 50	0.93 | 0
Crane	J48	0.64 | 44	0.74 | 50	0.63 | 50	0.70 | 53	0.63 | 40	0.62 | 50	0.63 | 53	**0.76 | 53**	0.70 | 0
	NB	0.73 | 50	0.75 | 52	0.71 | 44	0.73 | 53	0.72 | 50	0.71 | 50	0.69 | 46	**0.76 | 53**	0.75 | 0
	SVM	0.83 | 42	0.78 | 52	0.73 | 50	0.82 | 53	0.77 | 50	0.73 | 50	0.84 | **53**	**0.84** | 52	0.84 | 0
	RFs	0.85 | 50	0.83 | 40	0.83 | 50	0.84 | 53	0.84 | 50	0.83 | 42	0.84 | 50	**0.88 | 53**	0.88 | 0

Table 7: The time used by different feature evaluation methods.

Dataset	Classifier	Method: Cor	GR	IG	OR	RF	SU	CS	MICV	ORI
Birds	J48	2.1035	**1.8026**	2.7100	4.5227	2.3204	2.7069	2.3268	2.1074	3.1728
	NB	1.3208	2.2001	2.8666	6.2036	**1.2039**	1.6001	1.8569	1.8257	3.5810
	SVM	2.1580	3.1500	4.2689	3.6028	5.6244	3.6028	**2.0789**	2.5104	5.1568
	RFs	3.1829	4.2626	5.1698	2.7853	3.6952	4.1524	6.1236	**2.1568**	8.1732
Crane	J48	1.9326	2.3825	2.8624	2.2596	3.2632	4.1069	2.6547	**1.1638**	5.1629
	NB	1.8624	**1.6527**	1.6549	3.9326	4.3829	5.2806	4.1026	1.9628	4.6258
	SVM	6.3426	7.1869	6.5826	6.3429	5.3440	3.3651	4.6458	**3.0824**	7.3496
	RFs	4.9637	5.3746	6.0689	4.0547	4.1968	3.1906	6.6504	**3.1869**	8.3048

running time cost by the MICV-ERMFT method and the other seven feature selection methods. It is not too time-consuming than other methods.

In experiments of Birds dataset and Crane dataset, Kappa metrics using different classifiers with the MICV-ERMFT method are generally superior to the other methods. The MICV-ERMFT method remains excellent for the most part and is more stable than the other methods, although other methods surpass the MICV-ERMFT method in some classifiers. Besides, the MICV-ERMFT method improves the Kappa value compared to the original data. Although the improvement is minimal in some cases, the MICV-ERMFT method only uses about half of the characteristic features compared to the original data.

In conclusion, MICV-ERMFT has better performance in dimensionality reduction and model performance improvement.

4. Conclusion

Feature selection is an important preprocessing step in data mining and classification. In recent years, researchers have focused on feature contribution evaluation and redundancy reduction, and different optimization algorithms have been proposed to address this problem. In this paper, we measure the contribution of features to the classification from the perspective of probability. Combined with the maximum feature tree to remove the redundancy, the MICV-ERMFT method is proposed to select the optimal features and applied in the automatic recognition of bird sounds.

To verify the MICV-ERMFT method's effectiveness in automatic bird sounds recognition, two datasets are used in the experiments: data of different genera (Birds dataset) and data of the same genera (Crane dataset). The results of experiments show that the Kappa indicator of the Birds dataset reaches 0.93, and the dimension reduction rate reaches 57%. The Kappa value of the Crane dataset is 0.88, the dimension reduction rate reached 53%, and good results were obtained.

This study shows that the proposed MICV-ERMFT feature selection method is effective. The bird audio selected in this paper is noise filtered, and further research should test this method's performance using a denoising method. We will continue to explore the performance of MICV-ERMFT in the dataset with a larger number of features and instances.

Data Availability

All the data included in this study are available upon request by contact with the corresponding author.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China under Grants nos. 61462078, 31960142, and 31860332.

References

[1] C. A. Ruiz-Martinez, M. T. Akhtar, Y. Washizawa, and E. Escamilla-Hernandez, "On investigating efficient methodology for environmental sound recognition," in *Proceedings of the ISPACS 2013—2013 International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 210–214, Naha, Japan, November 2013.

[2] P. Jancovic and M. Köküer, "Bird species recognition using unsupervised modeling of individual vocalization elements," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 932–947, 2019.

[3] A. D. P. Ramirez, J. I. De La Rosa Vargas, R. R. Valdez, and A. Becerra, "A comparative between mel frequency cepstral coefficients (MFCC) and inverse mel frequency cepstral coefficients (IMFCC) features for an automatic bird species recognition system," in *Proceedings of the 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pp. 1–4, Gudalajara, Mexico, November 2018.

[4] D. Griffin and J. Jae Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[5] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 917–924, 1992.

[6] E. Tsau, S.-H. Kim, and C.-C. J. Kuo, "Environmental sound recognition with CELP-based features," in *Proceedings of the ISSCS 2011—International Symposium on Signals, Circuits and Systems*, pp. 1–4, Iasi, Romania, July 2011.

[7] C. Collberg, C. Thomborson, and D. Low, "A taxonomy of obfuscating transformations," Technical Reports 148, The University of Auckland, Auckland, New Zealand, 1997.

[8] S. García, J. Luengo, F. Herrera, S. García, J. Luengo, and F. Herrera, "Feature selection," *Intelligent Systems Reference Library*, vol. 72, pp. 163–193, 2015.

[9] V. Kumar and S. Minz, "Feature selection: a literature review," *Smart Computing Review*, vol. 4, 2014.

[10] Y. Zhang, Q. Wang, D.-W. Gong, and X.-F. Song, "Non-negative Laplacian embedding guided subspace learning for unsupervised feature selection," *Pattern Recognition*, vol. 93, pp. 337–352, 2019.

[11] S. Zhao, Y. Zhang, H. Xu, and T. Han, "Ensemble classification based on feature selection for environmental sound recognition," *Mathematical Problems in Engineering*, vol. 2019, Article ID 4318463, 7 pages, 2019.

[12] S. H. Zhang, Z. Zhao, Z. Y. Xu, K. Bellisario, and B. C. Pijanowski, "Automatic bird vocalization identification based on fusion of spectral pattern and texture features," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275, Calgary, Canada, April 2018.

[13] A. V. Bang and P. P. Rege, "Recognition of bird species from their sounds using data reduction techniques," in *Proceedings of the 7th International Conference on Computer and Communication Technology*, pp. 111–116, Allahabad, India, November 2017.

[14] M. Mafarja, I. Aljarah, A. A. Heidari et al., "Binary dragonfly optimization for feature selection using time-varying transfer functions," *Knowledge-Based Systems*, vol. 161, pp. 185–204, 2018.

[15] Q. Wu, Z. Ma, J. Fan, G. Xu, and Y. Shen, "A feature selection method based on hybrid improved binary quantum particle swarm optimization," *IEEE Access*, vol. 7, pp. 80588–80601, 2019.

[16] H. W. Wang, Y. Meng, P. Yin, and J. Hua, "A model-driven method for quality reviews detection: an ensemble model of feature selection," in *Proceedings of the Fifteenth Wuhan International Conference Electric Buses*, pp. 573–581, Wuhan, China, 2016.

[17] H. Rao, X. Shi, A. K. Rodrigue et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing*, vol. 74, pp. 634–642, 2019.

[18] D. A. A. Gnana, "Literature review on feature selection methods for high-dimensional data," *International Journal of Computer Applications*, vol. 136, no. 1, pp. 9–17, 2016.

[19] G. I. Sayed, A. Darwish, and A. E. Hassanien, "A new chaotic whale optimization algorithm for features selection," *Journal of Classification*, vol. 35, no. 2, pp. 300–344, 2018.

[20] A. E. Hegazy, M. A. Makhlouf, and G. S. El-Tawel, "Improved salp swarm algorithm for feature selection," *Journal of King Saud University—Computer and Information Sciences*, vol. 32, no. 3, pp. 335–344, 2020.

[21] M. Khamees, A. Albakry, and K. Shaker, "Multi-objective feature selection: hybrid of salp swarm and simulated annealing approach," in *Proceedings of the International Conference on New Trends in Information and Communications Technology Applications*, pp. 129–142, Baghdad, Iraq, January 2018.

[22] M. Sadeghi and H. Marvi, "Optimal MFCC features extraction by differential evolution algorithm for speaker recognition," in *Proceedings of the 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, pp. 169–173, Shahrood, Iran, December 2017.

[23] A. V. Bang and P. P. Rege, "Automatic recognition of bird species using human factor cepstral coefficients," *Smart Computing and Informatics*, vol. 77, pp. 363–373, 2018.

[24] R. H. D. Zottesso, Y. M. G. Costa, D. Bertolini, and L. E. S. Oliveira, "Bird species identification using spectrogram and dissimilarity approach," *Ecological Informatics*, vol. 48, pp. 187–197, 2018.

[25] J. Stastny, M. Munk, and L. Juranek, "Automatic bird species recognition based on birds vocalization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–7, 2018.

[26] S. Fagerlund, *Automatic Recognition of Bird Species by Their Sounds*, Helsinki University of Technology, Espoo, Finland, 2004.

[27] L. Ptáček, *Birds individual automatic recognition*, PhD thesis, University of West Bohemia, Pilsen, Czechia, 2012.

[28] A. B. Labao, M. A. Clutario, and P. C. Naval, "Classification of bird sounds using codebook features," in *Proceedings of the Asian Conference on Intelligent Information and Database Systems*, pp. 223–233, Dong Hoi City, Vietnam, March 2018.

[29] D. Lepage, "Avibase—the world bird database," 2020, https://avibase.bsc-eoc.org/avibase.jsp.

[30] G. A. Pereira, "Xeno-canto—sharing birds sounds from around the world," 2003, https://www.xeno-canto.org.

[31] J. Stastny, V. Skorpil, and J. Fejfar, "Audio data classification by means of new algorithms," in *Proceedings of the 2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 507–511, Rome, Italy, July 2013.

[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.