inter.noise
HAMBURG 2016

# Unsupervised Ensemble Feature Selection for Underwater Acoustic Target Recognition

Honghui YANG[1]; Anqin GAN; Sheng SHEN; Hanlu CHEN; Yue PAN[2]; Jansheng TANG; Jiangqiao LI

[1] School of Marine Science and Technology, Northwestern Polytechnical University, China

[2] Science and Technology on Underwater Acoustic Antagonizing Laboratory,

Institute of Systems Engineering Research, China

## ABSTRACT

The problems of feature redundancy and label information lacking of underwater acoustic target data and potential local optima of general feature selection algorithms result in poor classification performance and stability in the process of the underwater acoustic target recognition. The unsupervised feature selection using feature similarity (UFSUFS) can effectively remove redundant features without label information, and the ensemble feature selection method can improve the generalization accuracy and stability. Therefore, we proposed an algorithm of unsupervised ensemble feature selection using feature similarity (UEFSUFS) for underwater acoustic target recognition, which primarily generates several sets of results of the feature selection algorithm using diverse training sets by sampling from original training set with replacement, subsequently aggregates those results by majority voting method. The SVM classification performance and stability of the proposed method are examined on the UCI Sonar dataset and a real-world underwater acoustic target dataset. Experimental results on the two datasets show that the proposed method can improve the average classification rate of SVM and the Jaccard index, which indicate that the proposed method can remove redundant features and improve stability of feature selection, thereby leading to better classification performance.

Keywords: Underwater acoustic target recognition; Ensemble; Unsupervised feature selection I-INCE Classification of Subjects Number(s): 74.4

## 1. INTRODUCTION

In order to achieve better underwater acoustic target recognition result, the researchers tend to use a variety of methods to extract multi-domain features of underwater acoustic target radiated noise. However, the increased number of features will have the following issues: (1) underwater acoustic target samples are difficult to be obtained, the contradiction between the small sample size and the large feature size makes the recognition performance of underwater acoustic target recognition system decrease; (2) since some features is redundant for the recognition task, and large feature dimension will inevitably lead to complex identification systems, very slow speed of on-line recognition. Therefore, efficiently remove redundant features is important in the underwater acoustic target recognition.

According to whether the training set is labeled or not, feature selection algorithms can be categorized into supervised, unsupervised and semi-supervised feature selection. In the field of underwater acoustic target recognition, compared to the supervised feature selection methods [1,2], unsupervised feature selection methods[3] relatively scant. Supervised feature selection algorithms assesse the relevance of features guided by the label information. While unsupervised feature selection algorithms work with unlabeled data, it is difficult to evaluate the relevance of features. Pabitra Mitra proposed an unsupervised feature selection method which based on feature similarity(UFSUFS)[4]. This algorithm can quickly select a good feature subset, but it has poor stability and can't accurately select features which number is preset. In this paper, we proposed an unsupervised ensemble feature selection using feature similarity (UEFSUFS). The Support Vector Machine（SVM）classification performance and stability of the proposed method are examined on the UCI Sonar dataset and a

[1] hhyang@nwpu.edu.cn

real-world underwater acoustic target dataset. The results show that the proposed algorithm can solve above problems in underwater acoustic target recognition.

The rest of this paper is structured as follows. Section 2 introduces the principle of the proposed algorithm. Subsequently, we introduced the methodology used to assess stability of algorithms in section3. Section4 present the results of our experiment and our discussion. We conclude with some final remarks in section5.

## 2. UEFSUFS ALGORITHM

### 2.1 The Principle of UEFSUFS Algorithm

Integrated feature selection [5] idea derived from ensemble learning for supervised learning, namely ensemble classifier. In supervised learning, ensemble learning will generate multiple classifiers and then aggregate their classification results. It shows that this aggregated result is usually more accurate than the result from each individual classifier. Similarly, ensemble feature selection is to generate several diverse feature results, and then aggregate these results based on certain criteria.

The unsupervised ensemble feature selection using feature similarity (UEFSUFS) contains two main steps: firstly generates several sets of results of the feature selection algorithm using diverse training sets by sampling from original training set with replacement, subsequently aggregates those results by majority voting method. The procedure of this algorithm is as follows.

### 2.2 The Procedure of UEFSUFS Algorithm

The procedure of this algorithm is as follows:

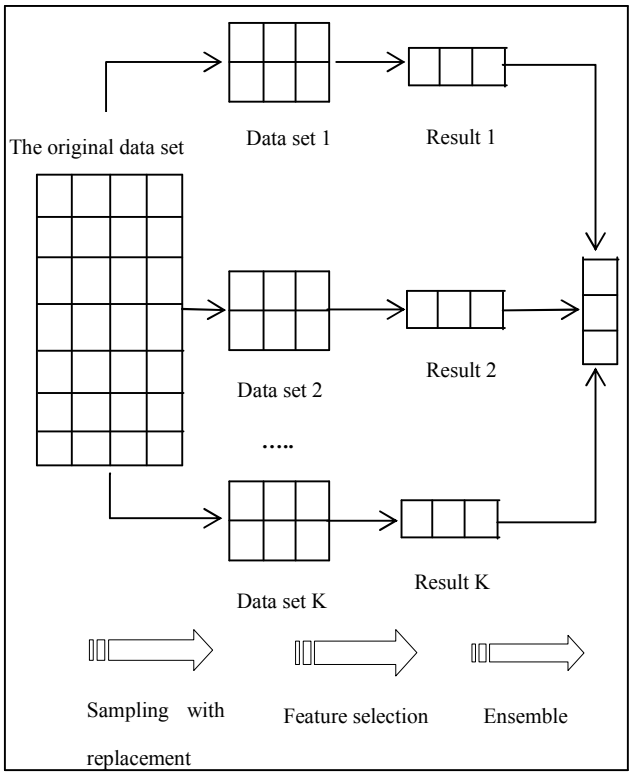



Figure 1 – Principle of UEFSUFS

Input: Training dataset $X$, original feature numbers $m$, feature selection times $K$ sampling sample numbers $p$ (less than the original sample numbers $n$), the value of nearest-neighbor $k$ (less than $m$), feature similarity evaluation function $S$.

Step 1: Obtained $K$ training dada subsets ( $X_1, X_2, ...X_K$ ) by sampling from original training set $X$ with replacement, every subset contains $p$ samples.

Step 2: Initialize the reduced feature subset $R$ to the original feature set.

Step 3: For each feature $F_i \in R$, compute $r_i^k$. (The $r_i^k$ represent the dissimilarity between feature $F_i$ and its $kth$ nearest-neighbor feature in $R$ )

Step 4: Find feature $F_i$ for which $r_i^k$ is minimum. Retain this feature in $R$ and discard $k$ nearest features of $F_i$. (Note: $F_i$ denotes the feature for which removing $k$ nearest-neighbor will cause minimum error among all features in $R$) Let $\varepsilon = r_i^k$.

Step 5: If $k > cardinality(\mathrm{R}) - 1 : k = cardinality(\mathrm{R}) - 1$. Where $cardinality(.)$ be defined to calculate the number of features.

Step 6: If $k = 1$: Go to step 9.

Step 7: While $r_i^k > \varepsilon$ do:

(a) $k = k - 1$. $r_i^k = \inf_{F_i \in R} r_i^k$. ("$k$" is decremented by 1, until the "$kth$ nearest-neighbor" of at least one of the features in R is less than $\varepsilon$-dissimilar with the feature)

(b) If $k = 1$: Go to step 9. (If no feature in R has less than $\varepsilon$-dissimilar "nearest-neighbor" select all the remaining features R.)

End While

Step 8: Go to Step3.

Step 9: Return feature set R as the reduced feature set.

Step 10: Repeat the Step 2 to Step 9, until $K$ training sets are performed feature selection. That will get $K$ feature selection results as $Fea\_K$.

Step 11: According to the $Fea\_K$, vote for each feature. The highest number of votes before $k(k = 1,...m)$ features as the corresponding integrated feature selection result, finally, can get $K$ integrated features selection results, denoted $EFea\_K$.

### 2.3 Feature Similarity Measure

The similarity between two features which is denoted as $S$ could be computed as follows. Let $\sum$ be the covariance matrix of random variables $x$ and $y$. Define maximal information compression index (MICI) as smallest eigenvalue of $\sum$, let $S = MICI(\mathrm{x}, \mathrm{y})$:

$$\begin{aligned} 2Mici(x,y) &= \mathrm{var}(x) + \mathrm{var}(y) \\ &- \sqrt{(\mathrm{var}(x) + \mathrm{var}(y))^2 - 4\,\mathrm{var}(x)\,\mathrm{var}(y)(1 - \rho(x,y)^2)} \end{aligned} \tag{1}$$

The value of $MICI$ is zero when the features are linearly dependent and increases as the amount of dependency decreases.

## 3. STABILITY MEASUREMENT

The stability of feature selection algorithm can be defined as the variation in feature selection results due to small change in the dataset. Stability is the major goal of ensemble learning based feature selection. Literature [6, 7] common stability evaluation methods are mostly based on the subsampling. Suppose the training dataset $X = (\mathrm{x}_1, ..., \mathrm{x}_n)$ contains $n$ samples and $m$ features. Then, generate diverse training sets by sampling from original training set with replacement, $r$ is defined as the ratio of the sample chosen from $X$. The subsampling process will be repeated for $K$ times and thus generate $K$ training data subsets. Subsequently, feature selection is performed on each of the $K$ training data subsets, and then a measure of stability is calculated based on the $K$ feature selection results.

In this paper, we use Jaccard index to evaluate the stability of the proposed method. The Jaccard index can be used as

$$S(FS_i, FS_j) = \frac{\left|FS_i \cap FS_j\right|}{\left|FS_i \cup FS_j\right|} = \frac{\sum_l I(FS_i^l = FS_j^l = 1)}{\sum_l I(FS_i^l + FS_j^l > 0)} \tag{2}$$

Where $FS_i$ represents the feature selection result based on subsample $i(1 \le i \le K)$, and $S(\mathrm{FS}_i, \mathrm{FS}_j)$ represents a similarity measure between $S(\mathrm{FS}_i, \mathrm{FS}_j)$ and $FS_j$. The value of $FS_i^l$ is 1 when $lth$ feature is selected, otherwise the value of $FS_i^l$ is zero. $\sum_l I(.)$ is a count operator.

The total stability is the average over all pairwise similarity comparisons between the different feature selection results [8]:

$$S_{tot} = \frac{2\sum_{i=1}^{K}\sum_{j=i+1}^{K} S(\mathrm{FS}_i, \mathrm{FS}_j)}{K(\mathrm{K}-1)} \tag{3}$$

# 4. EXPERIMENTS and DISCUSSION

## 4.1 Datasets and Classifier

In this section, we use UCI Sonar dataset [9] and a real-world underwater acoustic target dataset to validate performances of the UEFSUFS algorithm experimentally. Classification decision is based upon SVM. Whereas the original problem may be stated in a finite dimensional space, the underwater target datasets to discriminate are not linearly separable in that space. For this reason, in our experiment the original finite-dimensional space must be mapped into a much higher- dimensional space using radial basis function (RBF).

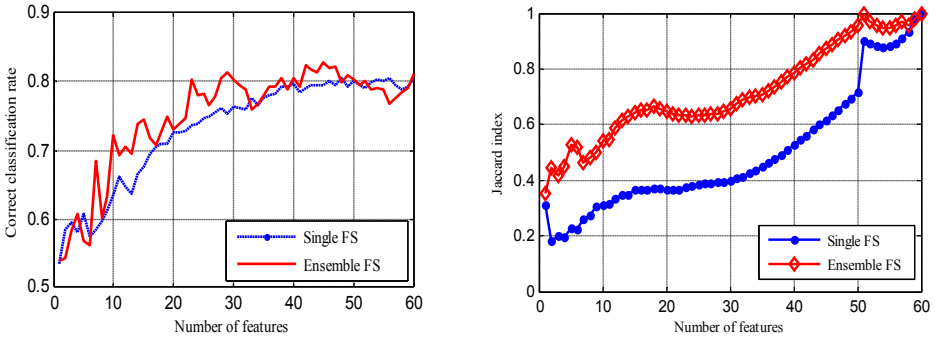
First, the data sets are briefly described. The sonar dataset contains 60 features obtained by bouncing sonar signals off a metal cylinder and rocks at various angles and under various conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. The data set contains signals obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock. Each feature value is in the range 0.0 to 1.0, each value represents the energy within a particular frequency band, integrated over a certain period of time. The integration aperture for higher frequencies occurs later in time, since these frequencies are transmitted later during the chirp. The underwater acoustic target dataset contains 71 features extracted from radiated noises signals of ships obtained by passive sonar, including wavelet analysis features (similar measurements of wavelet signal at all levels and low frequency envelop features of wavelet decomposition), waveform structure features (peak to peak amplitude distribution features, zero-crossing distribution features and wavelength difference distribution features), MFCC (Mel Frequency Cepstrum Coefficient) features and auditory spectrum features. There are totally 1920 samples of 4 classes of underwater acoustic targets, each class contains 480 samples. The detailed information of the datasets is shown in Table 1.

Table 1 – Detailed Information of the Datasets

| Dataset | Number of features | Number of classes | Number of Samples |
|---|---|---|---|
| Underwater acoustics target | 71 | 4 | 1920 |
| Sonar | 60 | 2 | 138 |

## 4.2 Experiments and Results

In experiments, we default $K=100$ , $p=100$ , in accordance with section 2.2 to obtain $Fea\_K$ and $EFea\_K$ . And then, we divide each dataset into 5 parts randomly using 5-fold cross validation method for 10 times according to $Fea\_K$ and $EFea\_K$ . Select a feature subset in the $EFea\_K$ , and define as this subset is the optimum ensemble feature selection result, which makes the 1-norm of the correct classification to be largest. The result comparison is shown in figure2, 3.




(a) (b)

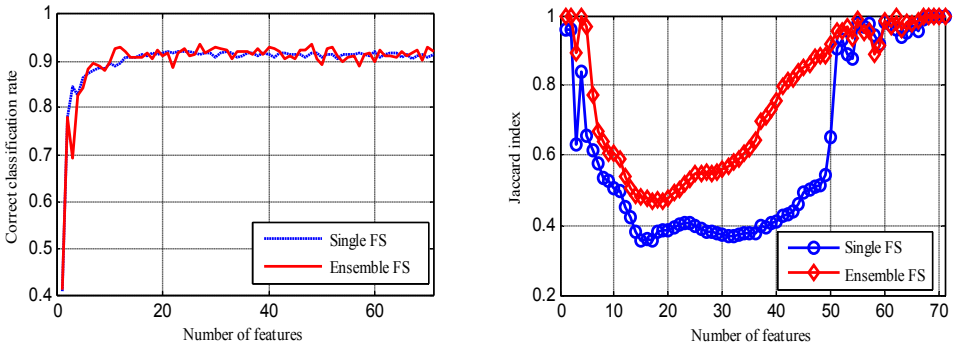
Figure 2 – The average correct classification rate and the stability of sonar dataset

As can be seen from Figure 2-a, in most of number of selected features, the SVM classification accuracy of the integrated feature selection is higher than single feature selection. Ensemble feature selection only needs 23 features to achieve the highest correct classification rate, while the single feature selection needs 27 features. Moreover, the average SVM correct classification rate of ensemble

feature selection is 74.93%, the single feature selection is 71.48%.

As shown in Figure 2-b, the Jaccard index of ensemble feature selection are significantly higher than the single feature selection.

In conclusion, when it is referred to Sonar dataset, UEFSUFS algorithm will not only increase the average correct classification rate but also greatly improve the stability of the algorithm.




(a) (b)

Figure 3 – The average correct classification rate and the stability of underwater acoustic target dataset

As can be seen from Figure 3-a, when it is referred to underwater acoustic target dataset, UEFSUFS algorithm slightly improve the correct classification rate. The ensemble feature selection only use 10 features to achieve the highest classification accuracy rate of 94.7%, while the single feature selection can achieve the highest classification accuracy rate of 94.4% by 28 features.

As shown in Figure 3-b, in most of the selected features, the Jaccard index of ensemble feature selection is significantly higher than the single feature selection. Even in some selected features, Jaccard index is 1, which shows multiple feature selection results are exactly same.

In conclusion, in the application of the underwater acoustic dataset, without reducing the classification accuracy, the stability of the ensemble feature selection significantly higher than single feature selection.

In addition, because of the Bagging integration mechanism, UEFSUFS can accurately select the default number of feature, this characteristic makes it more useful. Analyzing the result of experiment, UEFSUFS not only improve the average classification accuracy but also can greatly improve the stability of the algorithm, so as to improve the performance of the algorithm.

## 5. CONCLUSION

In this paper we proposed an algorithm of unsupervised ensemble feature selection using feature similarity (UEFSUFS) for underwater acoustic target recognition, which primarily generates several sets of results of the feature selection algorithm using diverse training sets by sampling from original training set with replacement, subsequently aggregates those results using voting method. Compared with the results of UFSUFS algorithm on Sonar dataset and underwater acoustic target dataset, UEFSUFS has better classification performance of SVM. In addition, by comparing the Jaccard index of two feature selection methods, prove that the proposed algorithm has better stability. In general, it can be observed that ensemble feature selection provides more stable results than a single feature selection, and can improve the performance of underwater target recognition.

## ACKNOWLEDGEMENTS

This work is supported by grants from National Key Laboratory of Science and Technology on Underwater Acoustic Antagonizing.

## REFERENCES

1. Yang HH, Gan AQ, Chen HL, et al. Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection. 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST) IEEE, 2016.
2. Yang HH, Wang Y, Sun J, et al. An adaboost support vector machine ensemble method with integration of instance selection and feature selection. Hsi-An Chiao Tung Ta Hsueh/Journal of Xi'an Jiaotong University, 2014, 48(12):63-68.
3. Shen S, Yang HH, Yuan S.Unsupervised feature selection approach based on mutual information for

underwater acoustic target classification. Journal of Acoustic technology, 2013, 32(6): 30-33.
4. Mitra P, Murthy C A, Pal S K. Unsupervised Feature Selection Using Feature Similarity[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(3):301--312.
5. Guan D, Yuan W, Lee Y K, et al. A Review of Ensemble Learning Based Feature Selection[J]. Iete Technical Review, 2014, 31(3):190-198.
6. Saeys Y, Abeel T, Peer Y V D. Robust Feature Selection Using Ensemble Feature Selection Techniques[J]. Lecture Notes in Computer Science, 2008:313-325.
7. He Z, Yu W. Stable feature selection for biomarker discovery.[J]. Computational Biology & Chemistry, 2010, 34(4):215-25.
8. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces[J]. Knowledge & Information Systems, 2007, 12(1):95-116.
9. S Hettich, C Blake, C Merz. UCI repository of macheine learning database [EB/OL](1998-7-11)[2014-5-3]http://www.ics.uci.edu/mlearn/MLRepository.html.