Research on Feature Selection Algorithm Based on Relief-F and Genetic Algorithm for Underwater Target

Yong-qiang JI1,*, Gao-zheng XU1, Peng HAN2, Lan XIE1 and Xin ZHANG1

1Systems Engineering Research Institute of CSSC, Beijing, China
2School of Marine Science and Technology, Northwestern Polytechnical University, Xi’an, China
*Corresponding author

Keywords: Underwater target recognition, Relief-F, Genetic algorithms, Feature selection.

Abstract. Aiming at the problem of multi-domain feature selection for underwater target, the feature selection algorithm based on Relief-F and genetic algorithm is proposed. Relief-F algorithm, reduce the search space and speed up the search speed of genetic algorithm. The population initialization of the genetic algorithm is guided by the result of feature sorting. Classification experiments show that the algorithm have a better comprehensive performance, which can get a higher recognition rate with fewer features.

Introduction

Detection and classification of underwater targets is a highly demanding task due to abundant heterogeneous underwater noise sources. Underwater target activity is reflected by acoustic events with each target having its own ‘acoustic signature’. The classification problem concerns the construction of a procedure that will be applied to a variety of acoustic signals, in which each new signal is assigned to one of a set of predefined classes on the basis of observed features [1]. The acoustic signatures for classifying the hydrophone received signals are obtained by applying suitable feature extraction technique, which is then fed to a suitable pattern recognition algorithm for classification. Modern techniques include Artificial Neural Networks (ANN) [2] and Support Vector Machines (SVM) based classification [3,4].

There are several problems in the process of extracting the underwater target characteristics: (1) Underwater target samples are difficult to obtain, the number of available samples is small and high characteristic dimension can affect the performance of classifier. So it is needed to, minimize the dimension of the extracted features ensuring the classifier recognition rate [5]; (2) For a more complete description of the underwater target characteristics, the noise signal is extracted from multiple aspects. The extracted features inevitably contain irrelevant features that will not be conducive to the classification identification task; (3) The extracted features including redundant features can’t improve recognition performance as information that is useful for identifying tasks can be inferred from other features. Aiming at the above questions, it is necessary to study the method of selecting characteristic of underwater target [6].

In this paper, we combine the filtering feature selection algorithm called the Relief-F algorithm and the package feature selection algorithm called BP-GA algorithm as a ReliefF-GA combining feature selection algorithm based on BP network is proposed. The algorithm uses the Relief-F algorithm to remove some of the features with less correlation, reduce the search space for subsequent genetic algorithms, the search speed of genetic algorithm is accelerated. The population initialization of genetic algorithm is guided by feature sequencing making it a better starting point for search. Classification experimental results show that the algorithm has good comprehensive performance because that it can solve the problem of multi-domain feature selection for underwater target.

BP-GA Feature Selection Algorithm

The BP-GA feature selection algorithm for feature selection is available by combine Genetic algorithm with BP neural network. The recognition results of BP neural network classifier are related
to the fitness function of genetic algorithm. Each feature subset is an individual in the genetic algorithm. The problem to be solved by genetic algorithm is to select the optimal subset of features. The setting of fitness function takes account of the magnitude of the characteristic dimension and the recognition rate of BP neural network classifier. The greater the probability that a subset of this trait is passed on to the next generation is followed by the smaller dimension of the subsets in the population and the greater fitness function. By three basic operation of selecting, crossing and mutating, and iterative optimization, ultimately you can output the optimal individual is founded. The characteristic dimension is reduced at the same time.

**Encode**

According to the purpose of reducing the dimension of the feature, each encoding (chromosome) corresponds to a set of selected features. The paper uses binary encoding which is represented by 0 or 1. The length of the encode is equal to the original feature dimension. The corresponding encoding bit is 0 which means that the feature subset does not contain the corresponding sequence number. Meanwhile the corresponding encoding bit is 1 which means that the feature subset contains the corresponding sequence number. Any characteristic combination has a unique corresponding encoding.

**Population Initialization**

An initial population containing \( M \) individuals is randomly generated, and each individual is represented by the coded.

**Fitness Function Design**

The quality of the selected feature subset is related to the recognition rate of the classifier and the dimension of the feature subset. Based of this reason, the paper designs fitness function according to the following two principles.

a. The value of fitness function is larger as the recognition rate of the BP neural network classifier is higher for a feature subset.

b. The value of fitness function is larger as the dimension is lower for a feature subset.

Considering the above two principles, the following fitness function is designed:

\[
f(X) = e^{\alpha \frac{1}{N} R(X)}
\]  

(1)

In the formula, \( X \) represents an individual in a population, that is a subset of features. \( N \) represents the feature dimension of the original sample. \( R(X) \) represents the recognition rate of the BP neural network classifier to the feature subset. \( D(X) \) represents the dimension of the feature subset. \( \alpha \) is a tunable parameter, and the influence of recognition rate on fitness function value is greater when \( \alpha \) is greater.

**Selection Operation**

The proportional selection operators is used to determine the probability of selected individuals. The individual in the population is \( X_k, k = 1, 2, \cdots, M \). The fitness of each individual is \( f(X_k) \). The probability that the individual is chosen is

\[
p(X_k) = \frac{f(X_k)}{\sum_{i=1}^{M} f(X_i)}
\]

(2)

Calculating the cumulative probability of the \( i \) th individual as

\[
q(X_i) = \sum_{k=1}^{i} p(X_k)
\]

(3)
The interval of cumulative probability is $[0,1]$. The roulette selection algorithm is chosen for selection. The random number $r$ in the interval $[0,1]$ is produced. If $r < q_i$, the individual $X_i$ is selected. Otherwise the individual $X_k$ is selected when $X_k$ makes $q(X_k-1) \leq X_k \leq q(X_k)$

**Cross Operation**

Cross operation is achieved by single point crossover operator. The cross operator steps are as follows: Individuals in the contemporary population are paired randomly; The location in the encoding is chosen randomly for crossover operation for each pair of individuals; After encoding the pairs of two individuals, the new individuals are obtained.

**Mutation Operation**

A single point mutation operator is used. According to the setting probability of mutation, an individual is selected to determine the location of the variation point randomly to change the value of the position.

**ReliefF-GA Combining Feature Selection Algorithm**

Relief algorithms have commonly been viewed as feature subset selection methods that are applied in a prepossessing step before the model is learned and are one of the most successful preprocessing algorithms to date\(^7\). The key idea of the original Relief algorithm is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other\(^8\). This method designs a correlation statistic to measure the importance of the characteristics. The correlation statistic is a vector, each component of a vector corresponds to an initial characteristic, the importance of a subset of features is determined by the sum of the relevant statistical components corresponding to each feature. Finally, only a threshold of $\delta$ is specified, and then select the characteristics corresponding to the relevant statistic component of $\delta$ are selected. The method is also realized by specify the number of features noted as $k$ and select the $k$ characteristics of the relevant statistic components.

The Relief algorithm is easy to implement, high efficiency. But the original Relief cannot deal with incomplete data and is limited to two-class problems. Its extension which solves dichotomy and multi-classification problems is called ReliefF\(^9\).

Although ReliefF algorithm cannot remove redundant features, it offers the feature weight value information for each feature. Compared with ReliedF algorithm, BP-GA algorithm is a wrapped feature selection method with low running efficiency and longer running time. However, the BP-GA algorithm has a strong global search ability, and its subset evaluation function (fitness function) is related to the recognition rate of the classifier, so the final classification performance is usually better.

By utilizing these information combined with BP-GA, a newly combination feature selection algorithm based on ReliefF and genetic algorithm is proposed called ReliefF-GA for short. The algorithm uses Relief-F to remove some less correlated features, which reduces the search space of the subsequent genetic algorithm and speeds up the search speed of the genetic algorithm. The results of genetic algorithm are used to guide the population initialization of genetic algorithm, so that it has a good starting search point. The ReliefF-GA algorithm contains two feature selection processes as follows.

**Relief-F Feature Selection Process**

In this process, the algorithm calculates the feature weight and ranks the features according to the correlation between the features and categories. The algorithm has two functions:

a. Remove the characteristics that are uncorrelated related to the category. Because the extracted feature contains many independent features, it can remove the later sorted features, which can reduce the search space of the subsequent genetic algorithm and improve the efficiency of the algorithm.

b. The remaining features can guide the population initialization of the genetic algorithm according to the feature weight. Then the algorithm has a better search starting point, so that the size of the
population can be reduced and the better feature subset can be searched by less evolutionary
generations.

**BP-GA Feature Selection Process**

On the basis of the initial selection of the features, the initialization of the genetic algorithm is
instructed by the result of the sorting of the s weight. When the feature ranking is higher, the
probability that the corresponding encoding bit will be 1 is larger. Meanwhile the feature ranking is
lower, the probability corresponding encoding bit will be 0 is larger.

The specific setting methods for the corresponding probability of each feature are: The probability
of encoding bit will be 1 corresponding to the first feature after sorting is \( f_1(f_1 > 0.5) \), the probability
of the last feature corresponding to the encoding bit will be 1 is \( f_2(f_2 < 0.5) \). The probability of the
remaining features is in the form of arithmetical series in \( (f_2, f_1) \). In this paper, \( f_1 = 0.7 \), \( f_2 = 0.3 \).

The population initialization method of genetic algorithm as follows. The population size is setted
as \( n \), and the feature dimension after Relief-F feature selection is setted as \( m \). The each individual in
the initial population is generated as follows. Randomly generate \( m \) random numbers \( r_i \) distributed
in interval \([0,1]\), where \( i = 1, 2, \ldots, m \). Compare the probability \( f_i \) to the \( i \) th feature and \( r_i \). If \( f_i \geq r_i \), then the \( i \) th feature is selected. And the corresponding individual coding bit is setted as 1, otherwise,
it is setted as 0. Using this method to generate \( n \) individuals, the initial population can be obtained.
This can not only ensure that the initial population of genetic algorithm is random, but also can
effectively use the prior information of the feature weight obtained by the Relief-F algorithm.

Figure 1 shows the flow of the ReliefF-GA algorithm.

**Experimental Results and Analysis**

Three experimental data sets is used for verification of the proposed algorithm in this paper. Two data
sets are obtained by feature selection for the noise signals testing the two types of target and three
types of targets. Each sample in the two data sets has 160 dimensional features (120 dimensional
higher-order spectral characteristics in frequency domain, and 40 dimensional Wigner higher-order
feature in time domain). The third data set is a publicly released Sonar data set from the UCI database
for testing. The following three data sets are called data sets(1), sets(2), and data sets(3) respectively.

In order to facilitate comparison analysis visually, figure 2 and figure 3 show the histogram of
average recognition rate of the feature subset and the feature subset dimension obtained by the feature
selection algorithms.

As can be seen from Figure 2, the average recognition rate of the three feature selection algorithms
is higher than that of the original feature set. It shows that the recognition rate of the classifier can be
effectively improved by using the feature selection algorithm to optimize the original high
dimensional features. From Figure 2 and figure 3, we can see that on each data set, the feature
reduction algorithm based of ReliefF-GA has the highest recognition rate and the lowest feature
subset dimension, it shows that the proposed ReliefF-GA combined feature feature reduction algorithm can better remove the unrelated and redundant features, and the feature subset has a better recognition purpose.

It can be concluded from table 7 that Relief-F algorithm is a filtering feature selection algorithm with low computational complexity and high operation efficiency; The BP-GA algorithm and ReliefF-GA algorithm need to perform search in global feature space, and need to train and test classifier many times to calculate the fitness function value, so the running time is far greater than the Relief-F algorithm. Because the ReliefF-GA algorithm first uses the efficient Relief-F to remove the smaller correlation feature, it greatly compresses the search space of the subsequent genetic algorithm, while the BP-GA algorithm searches the best feature subset in the original high dimensional feature space directly, so the ReliefF-GA algorithm reduces the running time of the BP-GA algorithm, and the efficiency has been greatly improved.

![Figure 2. Average recognition rate obtained by each feature selection algorithm.](image)

![Figure 3. The feature subset dimension of each feature selection algorithm.](image)

**Conclusion**

The ReliefF-GA feature selection algorithm proposed in this paper combines the advantages of the filtering method and the wrapping method. The measured water target data set and the UCI sonar data set are carried out for the classification experiments. The results show that this algorithm can not only guarantee high recognition rate, but also reduce the dimension of feature subset to greater extent. Compared with the operation efficiency of BP-GA algorithm, the proposed ReliefF-GA algorithm
has better comprehensive performance and can better solve the multi-domain feature selection problem in underwater target.

**Reference**


