Advanced Feature Selection Methodology for Automatic Target Recognition

Dale E. Nelson
Wright Laboratory
2010 Fifth Street, Building 23
WPAFB, Ohio 45433-7001
E-mail: nelsonde@aa.wpafb.af.mil

Janusz A. Starzyk
Ohio University
Stocker Center
Athens, Ohio, 45701-2979
E-Mail: starzyk@dolphins.ent.ohiou.edu

Abstract

This paper investigates independent feature selection as used in neural networks for solving classification problems. Radial basis functions and wavelet transforms are used to preprocess the input data. A class of nonorthogonal classifiers is defined and their properties are investigated. It is demonstrated that nonorthogonal classifiers perform better than the orthogonal ones. Feature selection using mutual information is also investigated. Independence of features based on the information content is defined and used to select features for synthesis of ontogenic neural networks. Simulation results using synthetically generated radar returns showed promise for automatic target recognition.

1. Introduction

Automatic target recognition (ATR) is a difficult task. When applied to air-to-air targets using High Range Resolution (HRR) radar, the task becomes even more difficult. Figure 1 shows the typical way the HRR signal is obtained. The returned signal is integrated over range bins, with each range bin containing the total radar return for that time segment. The difficulty of ATR using HRR data lies in the extreme variability in the radar signature with minor changes in azimuth, elevation, and time (the HRR problem). Figure 2 illustrates the variability of two signals. Each signal is an HRR radar return separated by two milliseconds In addition to the variability of the signal, a further complication arises from the use of synthetic data for training and measured data for testing. The most attainable data is synthetically generated. Measured data is used for test. Since measured data is expensive to obtain and usually in short supply, it is not feasible to use it for training. The inherent differences in these two types of data, plus the extreme variability in the data itself, causes poor performance in classification systems.

In this paper we introduce the concept of nonorthogonal feature vectors and mutually independent information which allows the creation of robust classifiers. Classifiers using these features can be implemented in ontogenic neural networks (networks that generate their own topology during training) using a minimal amount of hardware.

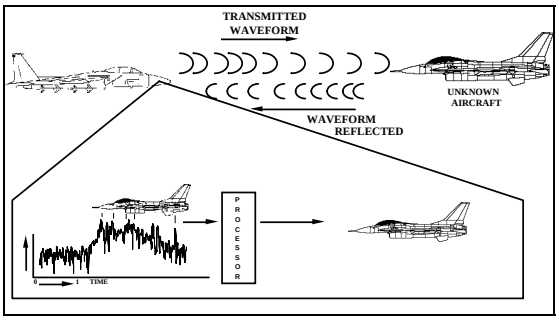

Figure 1. HRR Radar Target Identification

2. Classifiers

A fundamental problem in pattern classification is to determine class membership with the maximum statistical confidence of the correct classification decision. Classification can be performed with 100% probability for the test data, but there is no proof that this classifier yields a similarly high recognition rate for test data. To the contrary, these classifiers are often unable generalize and correctly classify new test data.

A specific transformation of the input is referred to as

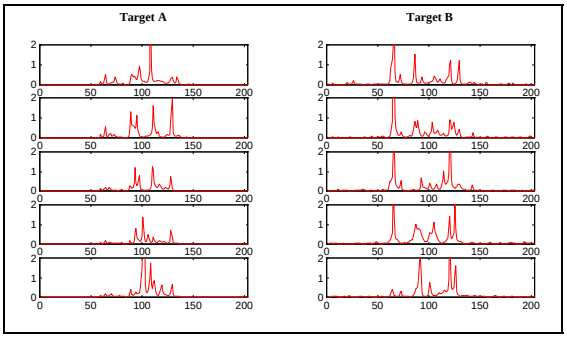

Figure 2. HRR Data for two Target Classes

an input feature. Using an unlimited number of features one can achieve linear separability of any input data. However, this selection may lead to costly classification procedures, expensive hardware, and an inability to generalize and classify new data. By proper selection of the input features we can obtain better classifiers, i.e. resistant to noise, local distortions, and with the ability to generalize.

3. Feature Selection

To learn a nonlinear mapping from the input space to the output space, one needs to consider independent transformations of the input space. Such transformations can be obtained using a complete set of orthogonal functions, in which orthogonality guarantees independence, or using successive approximations of the learned mapping, where successive transformations are found by orthogonalizing the error of the existing fit.

Let us define a **feature *f*** as an ordered pair (F,Ω) of a nonlinear transformation *F* and a proper subset of its output space. We define a **feature domain** *D* as a subset of the domain of the transformation *F* which is mapped into Ω, and a **feature sample set** *S* as a subset of the input training data included in *D*.

For a given transformation *F* we can define infinitely many features by simply modifying the subset Ω. An entire classification task can be based on a single transformation F paired with different sets Ω.

Definition: A feature f_m is ***covered*** by the features f_1, f_2, …, f_k iff $D_m \subset D_1 \cup D_2 \cup ... \cup D_k$.

Definition: A set of features $\Phi = \{ f_1, f_2, \ldots, f_n\}$ is ***independent*** if none of its elements can be covered by others.

Consider a set of training data *T* used for pattern recognition. This set is composed of subsets of vectors from different classes $T = C_1 \cup C_2 \cup ... \cup C_c$, where C_i is a set of vectors from the class *I*. For simplicity let the same symbol represent a class and its set of input vectors. Let us assume that all classes are disjoint, i.e. $C_i \cap C_j = \varnothing$ for $i \neq j$.

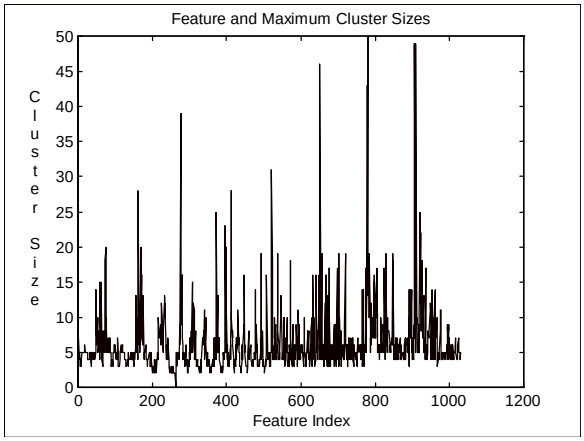

Figure 3. Features and Cluster Sizes

Definition a feature $f(C_i)$ is a ***distinguishing feature*** of class C_i if its domain includes only the input vectors from class C_i.

Whether a feature is distinguishing or not depends on the complete training set *T*. Adding or deleting training data may change a distinguishing feature to a non-distinguishing feature or vice versa.

Definition: A ***dominating distinguishing feature*** is a distinguishing feature with the highest ratio of the cardinality of its feature set over the cardinality of the associated class set.

Definition: A set of distinguishing features $\Phi(C_i) = \{f_1(C_i), f_2(C_i), \ldots, f_n(C_i)\}$ is an ***orthogonal classifier*** for class C_i if the sum of the domains of its features includes the set C_i. The classifier is a ***minimal classifier*** if the distinguishing features are independent.

To classify input data, we can use a set of orthogonal classifiers defined for all classes. Orthogonal classifiers yield 100% recognition for the training data. However, the recognition rate for new data may be significantly less, as often the orthogonal classifiers are based on many independent features, leading to small feature domains and poor generalization ability.

To design a simple classifier, we need a set of independent features of minimum cardinality which differentiates all classes. In this paper, we introduce feature selection based on a ***sequential classifier***. A sequential classifier is obtained as follows: A differentiating feature is selected and its sample set is removed from the input space. Then, another differentiating feature is selected and its sample set removed. This process repeats until all training samples are classified. A sequential classifier is a nonorthogonal classifier. We demonstrated that nonorthogonal classifiers are able to obtain correct classification of trained data with better generalization ability than orthogonal classifiers. Sequential classifier implementation will result in a multilevel neural net structure where the number of neurons, processing layers and the overall organization is a function of the input data, a feature of the ontogenic neural net [2].

Since an input vector may or may not exhibit an individual feature, we can design combinational classifiers in which a decision regarding an input classification is expressed by a combinational logic function which depends on several features. This leads to the construction of pattern recognition neural networks in which classification decisions are made by a network of logic gates. Such classifiers will be extremely hardware efficient. Probabilistic or fuzzy classifiers can also be

designed by using fuzzy logic instead of binary logic.

4. Simulation Results

In this work, sequential classifiers demonstrate the potential of nonorthogonal classifiers for ATR. A simple feature selection method was used to demonstrate that the nonorthogonal classifiers work even with the simplest thresholding features. It is expected that, when used with more elaborate feature selection processes, nonorthogonal classifiers will maintain their advantage over the orthogonal classifiers. The experiments described used synthetically generated HRR data. The data set has six targets, each target has 1071 profiles of 128 range bins. The training and test sets consisted of 60 randomly selected profiles for each target, giving 360 profiles for each set. In actual usage, much larger training and test sets would be used.

The input space includes the original signal, its amplitude range, average value, standard deviation, Shannon entropy, log energy, and l^p norms. A Haar wavelet transform of the original signal was used to enhance the input space. Figures 1 and 2 show examples of raw signal data. Figure 4 shows an example of a signal and its Haar transforms.

Feature selection was based on the *slicing approach* in which input data are projected onto one dimensional subspaces. In these subspaces, intervals which include input vectors from different classes were found. These intervals define slices in the multidimensional input space which contain only the samples of a single class. Features were selected based on the cardinality of input vectors in these slices. Figure 3 illustrates maximum cluster sizes for different transformation functions of the input data. Notice that the best features are based on the signal transforms rather than the raw data (represented by features 7-134.) This is in agreement with other research results in ATR, which indicate that preprocessing may enhance classifiers' recognition ability [1].

	Training Data			
	Orthogonal		Sequential	
Target Number	Recognition Rate %	Error Rate %	Recognition Rate %	Error Rate %
1	87	0	83	0
2	53	0	60	3
3	50	0	63	0
4	37	0	67	3
5	28	0	78	2
6	28	0	48	2

Table 1. Classification Training Results

	Test Data			
	Orthogonal		Sequential	
Target Number	Recognition Rate %	Error Rate %	Recognition Rate %	Error Rate %
1	83	0	85	0
2	28	18	32	8
3	40	2	73	3
4	25	18	53	15
5	12	0	48	10
6	18	7	35	15

Table 2. Classification Test Results

5. Mutual Information Measure

A sequential classifier is a special case of a more general combinational classifier. In a combinational classifier, independent features can be selected in a number of ways. It is our aim to select these features in a way that yields optimum neural network structures. A natural way of achieving this is to select a classifier which contains a minimum number of features satisfying a specified selection criteria. In this section, feature selection based on the mutual information measure is investigated. It yields a maximally informative set of features which minimizes the initial uncertainty in the object class.

It is expected that this approach yields the minimum number of features to perform the classification task. The mutual information between a feature f and a set of classes G is defined as follows [3]:

$$I(f,C) = \sum_{C \in G} P(C,f) \log \frac{P(C,f)}{P(C)P(f)}$$

where P(C,f) stands for the joint probability for class C

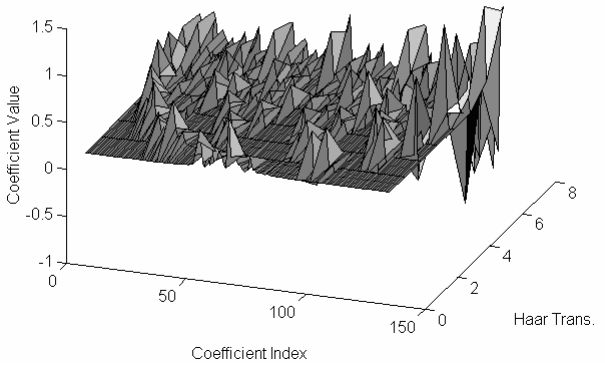

Figure 4. Haar Wavelet of a Signal

and feature f, P(C) is the class probability, and P(f) is the feature probability. In general, the conditional entropy, which measures the uncertainty in the object class, will be reduced when a new feature is added to the classifier. It remains unchanged if and only if P(C,f)=P(C)P(f), in which case the feature does not bring new information and should not be used in the classifier. Batti presented convincing arguments why the mutual information measure is useful in feature selection for object classification. His selection procedure uses a sequence of features f_i which maximizes I($f_p C$). He realized that this selection may introduce features which are strongly dependent and in spite of having large I($f_p C$) values their contribution to the classification problem may be much less than expected. To alleviate this problem he uses a feature selection that maximizes:

$$I(f,C) - \beta \sum_{f_p \in F_p} I(f, f_p)$$

where f_p is the set of the previously selected features, I(f,f_p) is the mutual information (which measures dependence) between candidate feature f and the already selected features f_p and β is a parameter between 0 and 1 which regulates the relative importance of the mutual information between features f and f_p.

The problem with this approach is that β is arbitrarily selected and cannot correctly remove the mutual information between features. The major reason why no single value of β can be found is that all previously selected features may be mutually dependent. The more dependence there is between previously selected features, the smaller the β value must be used. But this, in turn, is very much case dependent. What is more important, features selected using this criteria may be completely dependent on features previously selected and will not contribute to information increase. As a result, the obtained classifier is not minimal from an information theory point of view. In addition, the method gives no clue as to how many features should be selected to reach the optimum level of information accuracy possible with a given training set.

Motivated by the above deficiencies, a new method called the maximum information increase feature selection (MIIFS) was developed and tested on a selected set of input features. The method is computationally efficient and provides optimum feature selection based on the exact information measure. In this method, both the feature domain and its complement are considered in reaching the classification decision. As a consequence, each feature partitions the input space into two subspaces. If several features are considered the input space is partitioned into a number of subspaces. Each subspace is included in a unique combination of feature domains or their complements. In the MIIFS approach, the mutual information is computed based on these subspaces. The mutual information between a set of features P and the set of classes G is computed from:

$$I(\Phi, G) = \sum_{s \in S} \sum_{C \in G} P(C,s) \log \frac{P(C,s)}{P(C)P(s)}$$

where S is the orthogonal sum of all the subspaces created by the intersection of feature domains and their complements. First, the mutual information is computed for each feature with $S = D \oplus \overline{D}$, and $\Phi = f$. A feature f with the largest value of information $I(\Phi,G)$ is selected. The input space is divided into two subspaces. Next the feature space is searched for a new feature which maximizes the information increase

$$\Delta I = I(\Phi_n, G) - I(\Phi_{n-1}, G)$$

where $I(\Phi_n,G)$ is the mutual information between a new set of features Φ_n and the set of classes G. Φ_{n-1} stands for the previously selected set of features. $I(\Phi_n,G)$ is computed on a new set of subspaces which are created by intersecting a new feature domain and its complement with the previously obtained set of subspaces. Notice, that if Φ_n has n features, then up to 2^n subspaces are created. Feature selection continues until the maximum value of ΔI is less than a specified threshold. This method produces a minimum set of independent features optimized from an information theory point of view. If the information threshold is set to zero, the method produces features capable of 100% recognition of the trained data. Since the method minimizes the number of features, it is also capable of good generalization. Although the method can achieve 100% recognition, statistical confidence in the classification may be lowered by small increments of information added by features selected later in the process. This is a direct result of the error in estimating the mutual information I which differs from the true value of information $\bar{I}$ represented by the set of features, where the error of the mutual information estimate is:

$$\Delta I = I - \bar{I} \approx \frac{1}{2N}(K_C K_S - K_C - K_S)$$

and N is the total number of training samples, K_c is the number of classes, K_s is the number of subspaces. An additional issue, that must be investigated to determine the confidence in the classification result, is to determine the statistical likelihood that a new sample may be of a different class than the samples represented by the feature domain (subspace) which correspond to the given combination of features. This is directly related to the number of training data in the given subspace.

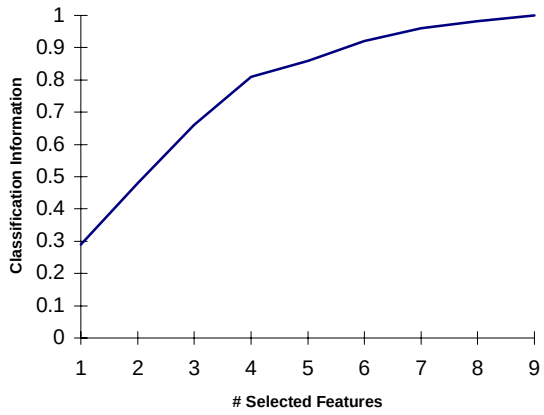

Figure 4. Information Content in a Sequence of Features

The maximum information increase for feature selection was used to select a small number of independent features for target classification. Using this approach the initial class uncertainty may be reduced to zero with a relatively small number of features. These features, in the combinational classifier, will give 100% recognition rate for the training data. In addition, since a small number of features is selected, the obtained combinational classifier should have a good recognition rate for the test data. Figure 4 illustrates the total information which can be obtained from a sequence of features selected using the MIIFS method. We see that the information content of the selected features quickly saturates to 100%. Further increase in the number of features will not add new information about the existing set of training signals. It may, however, increase the robustness of the classifier to recognize test data. (considering that the trained system could have been affected by noise).

Target Number	Training Data		Test Data	
	Recognition Rate %	Error Rate %	Recognition Rate %	Error Rate %
1	100	12	93	28
2	81	7	65	23
3	95	12	95	17
4	75	13	65	28
5	95	15	87	17
6	88	7	73	8

Table 3. Classification Results Using MIIFS

6. Conclusion

In this work, feature selection methods for use in neural network classifiers for HRR target recognition was investigated. Sequential and combinational classifiers were used as an example of nonorthogonal classifiers to select distinguishing features for use in synthesizing ontogenic neural networks. An orthogonal classifier, based on the dominating distinguishing features was selected to compare the classification performance. Wavelet transforms and other signal transformations were used to preprocess the radar signal data. Simulation results demonstrate the potential benefit of using nonorthogonal classifiers for ATR.

7. References

[1] Q. Zhao and Z. Bao, “Radar Target Recognition Using a Radial Basis Function Neural Network,” Neural Networks, vol. 9, no. 4, pp. 709-720, 1996

[2] D. Ensley and D. Nelson, “Applying Cascade Correlation to the Extrapolation of Chaotic Time Series,” Proceedings of the Third Workshop on Neural Networks: Academic, Industrial, NASA, Defense 92; (Auburn AL, Feb. 1992).

[3] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning.” IEEE Transactions on Neural Networks, vol 5, pp. 537-550, July 1996.