Feature Selection: A Neuro-fuzzy Approach

Sankar K. Pal Jayanta Basak Rajat K. De[1]

Machine Intelligence Unit,

Indian Statistical Institute,

Calcutta 700035, INDIA.

E-mail: sankar@isical.ernet.in

jayanta@isical.ernet.in

res9318@isical.ernet.in

1. Introduction

The main objective of *feature selection*, is to retain the optimum salient characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms can be devised for efficient classification. Several methods based on probability theory [1], fuzzy set theory [2,3] and Artificial Neural Network (ANN) [6,7,8] have been reported. Incorporation of fuzzy set theory enables one to deal with uncertainties in a system, arising from vagueness, incompleteness in available information, in an efficient manner. ANNs, having the capability of fault tolerance, adaptivity and generalization, and scope for massive parallelism, are widely used in dealing with optimization tasks. Recently, attempts are being made to integrate the merits of fuzzy set theory and ANN under the heading "neuro-fuzzy computing" for making the systems artificially more intelligent. This article is an attempt in this line, and has two parts. In the first part a method of ranking the features (or subsets of features) is described using a new fuzzy set theoretic feature evaluation index and its performance with an existing one [1,2] is compared. The second part provides a neuro-fuzzy approach where a new connectionist model has been designed in order to optimize the aforesaid fuzzy evaluation index which incorporates weighted distance for computing class membership values. This optimization process results in a set of weighting coefficients representing the importance of the individual features. These weighting coefficients lead to a transformation of the feature space for better modeling the class structures. The effectiveness of the algorithms is demonstrated on a speech recognition problem.

2. Evaluation Index and Feature Subset Selection

Let the pth pattern vector be represented as $\mathbf{f}^{(p)} = [f_1^{(p)}, f_2^{(p)}, \ldots f_i^{(p)}, \ldots f_n^{(p)}]$, where n is the number of features in M (set of measurable quantities). Let $prob_k$ and $d_k(\mathbf{f}^{(p)})$ stand for the *a priori* probability for the class C_k and the distance of the pattern $\mathbf{f}^{(p)}$ from the kth mean vector $\mathbf{m}_k$ $(= [m_{k1}, m_{k2}, \ldots m_{ki} \ldots m_{kn}])$ respectively. λ_{ki} is the bandwidth for the class C_k along the direction of the ith feature.

[1]R. K. De is a Dr. K. S. Krishnan Senior Research Fellow, Department of Atomic Energy, Government of India.

0-7803-3210-5/96 $4.00©1996 IEEE

The feature evaluation index is defined as,

$$E = \sum_{k} \sum_{\mathbf{f}^{(p)} \in C_k} \frac{s_k(\mathbf{f}^{(p)})}{\sum_{k' \neq k} s_{kk'}(\mathbf{f}^{(p)})} \times \alpha_k, \tag{1}$$

where

$$s_k(\mathbf{f}^{(p)}) = \mu_{C_k}(\mathbf{f}^{(p)}) \times (1 - \mu_{C_k}(\mathbf{f}^{(p)})) \tag{2}$$

and

$$s_{kk'}(\mathbf{f}^{(p)}) = \frac{1}{2}[\mu_{C_k}(\mathbf{f}^{(p)}) \times (1 - \mu_{C_{k'}}(\mathbf{f}^{(p)}))] + \frac{1}{2}[\mu_{C_{k'}}(\mathbf{f}^{(p)}) \times (1 - \mu_{C_k}(\mathbf{f}^{(p)}))]. \tag{3}$$

$\mu_{C_k}(\mathbf{f}^{(p)})$ and $\mu_{C_{k'}}(\mathbf{f}^{(p)})$ are the membership values of the pattern $\mathbf{f}^{(p)}$ in classes C_k and $C_{k'}$ respectively. α_k is the normalizing constant for class C_k.

Note that, s_k is zero (minimum) if $\mu_{C_k} = 1$ *or* 0, and is 0.25 (maximum) if $\mu_{C_k} = 0.5$. On the other hand, $s_{kk'}$ is zero (minimum) when $\mu_{C_k} = \mu_{C_{k'}} = 1$ *or* 0, and is 0.5 (maximum) for $\mu_{C_k} = 1$, $\mu_{C_{k'}} = 0$ or *vice-versa.*

Therefore, the term $\frac{s_k}{\sum_{k' \neq k} s_{kk'}}$ is minimum if $\mu_{C_k} = 1$ and $\mu_{C_{k'}} = 0$ for all $k' \neq k$ *i.e.*, if the ambiguity in the belongingness of a pattern $\mathbf{f}^{(p)}$ to classes C_k and $C_{k'}$ $\forall k' \neq k$ is minimum (the pattern belongs to only one class). It is maximum when $\mu_{C_k} = 0.5$ for all k. In other words, the value of E decreases as the belongingness of the patterns increases for only one class (*i.e.*, compactness of individual classes increases) and at the same time decreases for other classes (*i.e.*, separation between classes increases). E increases when the patterns tend to lie at the boundaries between classes (*i.e.*, $\mu \rightarrow 0.5$). Our objective is, therefore, to select those features for which the value of E is minimum.

In order to achieve this, $\mu_{C_k}(\mathbf{f}^{(p)})$ is defined with a multi-dimensional π-function [5], and is given by,

$$\begin{aligned} \mu_{C_k}(\mathbf{f}^{(p)}) &= 1 - 2d_k^2(\mathbf{f}^{(p)}) && 0 \leq d_k(\mathbf{f}^{(p)}) < \tfrac{1}{2}, \\ &= 2[1 - d_k(\mathbf{f}^{(p)})]^2 && \tfrac{1}{2} \leq d_k(\mathbf{f}^{(p)}) < 1, \\ &= 0 && otherwise, \end{aligned} \tag{4}$$

where

$$d_k(\mathbf{f}^{(p)}) = \left[\sum_{i} \left(\frac{f_i^{(p)} - m_{ki}}{\lambda_{ki}} \right)^2 \right]^{\frac{1}{2}}. \tag{5}$$

$\lambda_{ki} = 2 \max_{\mathbf{f}^{(p)} \in C_k} \left[|f_i^{(p)} - m_{ki}| \right]$ and $m_{ki} = \frac{\sum_{\mathbf{f}^{(p)} \in C_k} f_i^{(p)}}{|C_k|}$ are measured from the data set.

In Eqn. (1), the effect of class size is normalized by introducing a multiplicative factor α_k corresponding to the class C_k. Here we have chosen $\alpha_k = 1 - prob_k$. However, other expressions like $\alpha_k = \frac{1}{|C_k|}$ or $\alpha_k = \frac{1}{prob_k}$ could also have been used.

3. Weighted Membership Function and Feature Evaluation using Neural Network

In the previous section, the class structures are modeled before computing the index and it is kept fixed throughout the computation. Instead of rigidly modeling the class structures, a weighted membership function is defined where the feature space is suitably transformed depending on the weighting factors. The new weighted membership function is also given by Eqn. (1) where $d(\mathbf{f}^{(p)})$ is defined as

$$d_k(\mathbf{f}^{(p)}) = \left[\sum_i w_i^2 \left(\frac{f_i^{(p)} - m_{ki}}{\lambda_{ki}}\right)^2\right]^{\frac{1}{2}}, \quad w_i \in [0,1]. \tag{6}$$

The compactness of the individual classes and the separation between the classes as measured by E (Eqn. (1)) is now essentially a function of $\mathbf{w}$ $(= [w_1, w_2, \ldots w_n])$. The problem of feature selection/ranking thus reduces to finding a set of w_is for which E becomes minimum. The task of minimization has been performed by using gradient descent technique in a connectionist framework (because of its massive parallelism, fault tolerance, adaptivity etc.). A new connectionist model is developed for this purpose.

Note that, the method of individual feature ranking, explained in Section 2, is not identical to that described in this section. The later one finds the set of w_is (for which E is minimum) considering the effect of inter-dependencies of the features, whereas in the case of former one, each feature is considered individually independent of other. Let us now present the network model and its dynamics.

The network (Fig. 1) has two layers : input layer accepting the feature values, and the output layer providing the class membership values. Auxiliary nodes modulate the activation of the output nodes. An output node, after receiving input from the input layer through the feedforward links, can become activated only when the corresponding auxiliary node is made active. The weight of the feedback link from the auxiliary node, (corresponding to the kth output node) to the ith input node is equated to $-m_{ki}$. The weight of the feedforward link from the ith input node to the kth output node provides the degree of importance of the feature f_i, and is $W_{ki} = (\frac{w_i}{\lambda_{ki}})^2$. w_is are updated during the training process in order to minimize $E(\mathbf{w})$ (Eqn. (7)). Note that, λ_{ki}s and m_{ki}s are directly computed from the training set and kept fixed during updating of w_is. During training, the patterns are presented at the input layer and the membership values are computed at the output layer.

The auxiliary nodes are activated (*i.e.* activation values are equated to unity) one at a time while the others are made inactive (*i.e.*, the activation values are fixed at 0), and thus during training, only one output node is allowed to get activated at a time. Whenever an auxiliary node is activated, it sends the feedback to the input layer. The input nodes in turn send the resultant activations to the output nodes. The activation of the output node (connected to the active auxiliary node) provides the membership value of the input pattern to the corresponding class.

Thus, the membership values of the input pattern corresponding to all the classes are computed by sequentially activating the auxiliary nodes one at a time.

When the kth auxiliary node is activated, input node i has an activation value is $u_{ik}^{(p)} = (x_{ik}^{(p)})^2$ where $x_{ik}^{(p)} = f_i^{(p)} - m_{ki}$ is the total activation received by the ith input node for the pattern $\mathbf{f}^{(p)}$. $f_i^{(p)}$ is the external input and $-m_{ki}$ is the feedback activation from the kth auxiliary node to the ith input node. The activation value of the kth output node is $v_k^{(p)} = g(y_k^{(p)})$, where $g(.)$, the activation function of each output node, is a π-function as given in Eqn. (4). $y_k^{(p)} = \left(\sum_i u_{ik}^{(p)} \times (\frac{w_i}{\lambda_{ki}})^2\right)^{\frac{1}{2}}$ (same as d_k in Eqn. (5)) is the total activation received by the kth output node for the pattern $\mathbf{f}^{(p)}$. $v_k^{(p)}$ is equal to the membership value of the input pattern $\mathbf{f}^{(p)}$ in the class C_k.

The expression for $E(\mathbf{w})$ (from Eqn. (1)), in terms of the output node activations, is

$$E(\mathbf{w}) = \sum_k \sum_{\mathbf{f}^{(p)} \in C_k} \frac{v_k^{(p)}(1 - v_k^{(p)})}{\sum_{k' \neq k} \frac{1}{2}\left[v_k^{(p)}(1 - v_{k'}^{(p)}) + v_{k'}^{(p)}(1 - v_k^{(p)})\right]} \times \alpha_k. \tag{7}$$

w_is are updated by minimizing $E(\mathbf{w})$ by gradient-descent technique, *i.e.*,

$$\triangle w_i = -\eta \frac{\partial E}{\partial w_i}, \forall i, \tag{8}$$

where η is the learning rate.

For the computation of $\frac{\partial E}{\partial w_i}$, the following expressions are used.

$$\frac{\partial s_{kk'}(\mathbf{f}^{(p)})}{\partial w_i} = \frac{1}{2}\left[[1 - 2v_{k'}^{(p)}]\frac{\partial v_k^{(p)}}{\partial w_i} + [1 - 2v_k^{(p)}]\frac{\partial v_{k'}^{(p)}}{\partial w_i}\right], \tag{9}$$

$$\frac{\partial s_k(\mathbf{f}^{(p)})}{\partial w_i} = [1 - 2v_k^{(p)}]\frac{\partial v_k^{(p)}}{\partial w_i}, \tag{10}$$

$$\begin{aligned} \frac{\partial v_k^{(p)}}{\partial w_i} &= -4d_k(\mathbf{f}^{(p)})\frac{\partial d_k(\mathbf{f}^{(p)})}{\partial w_i}, & 0 \leq d_k(\mathbf{f}^{(p)}) < \tfrac{1}{2} \\ &= -4[1 - d_k(\mathbf{f}^{(p)})]\frac{\partial d_k(\mathbf{f}^{(p)})}{\partial w_i}, & \tfrac{1}{2} \leq d_k(\mathbf{f}^{(p)}) < 1 \\ &= 0, & otherwise \end{aligned} \tag{11}$$

$$\frac{\partial d_k(\mathbf{f}^{(p)})}{\partial w_i} = \frac{w_i}{d_k(\mathbf{f}^{(p)})}(\frac{f_i^{(p)} - m_{ki}}{\lambda_{ki}})^2. \tag{12}$$

After convergence, $E(\mathbf{w})$ attains a local minima, in that case, the weights of the feedforward links indicate the order of importance of the features.

5. Results

The above-mentioned algorithms were tested on a speech recognition problem. The data consists of a set of 437 Indian Telugu vowel sounds collected by trained personnel. These were uttered in a consonant-vowel-consonant context by three male speakers in the age group of 30 to 35 years. The data set has three features, f_1, f_2 and f_3 corresponding to the first, second and third vowel formant frequencies obtained through spectrum analysis of the speech data. Fig. 2 shows a 2-D projection of the 3-D feature space of the six vowel classes (∂, a, i, u, e, o) in the $f_1 - f_2$ plane (for ease of depiction).

Table 1 indicates the orders of different subsets of features based on the values of E (Eqn. (1)). These orders are also compared with those obtained by Pal *et al.* [2,3]. Table 1 shows that the subset $\{f_2\}$ is the best and $\{f_1, f_2\}$ is the second best using Eqn. (1) while the corresponding order is $\{f_1, f_2\}$ and $\{f_2\}$ in the case of Pal *et al.* However, in both the methods, the difference in index values for the subsets $\{f_2\}$ and $\{f_1, f_2\}$ is insignificant. f_3 stands at the bottom of the order list. Note also that, the inclusion of f_2 in a subset improves its characterization/discrimination ability. This further justifies the importance of f_2 in characterizing vowel classes. These results conform to the earlier findings [4] on speech recognition.

Table 2 provides the degrees of importance (**w**) of different features of the data set, obtained by the neural network model. The effect of three different weight initializations is shown in Table 2). It is found from Table 2 that the order of the features, in all the cases, is f_2, f_3, f_1.

References

1. P. A. Devijver and J. Kittler, *Pattern Recognition, A Statistical Approach.* London: Prentice Hall International Inc., 1982.

2. S. K. Pal and B. Chakraborty, "Fuzzy set theoretic measures for automatic feature evaluation," *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-16, no. 5, pp. 754–760, 1986.

3. S. K. Pal, "Fuzzy set theoretic measures for automatic feature evaluation: II," *Information Sciences*, vol. 64, pp. 165–179, 1992.

4. S. K. Pal and D. K. DuttaMajumder, *Fuzzy Mathematical Approach to Pattern Recognition.* New York: John Wiley (Halsted Press), 1986.

5. S. K. Pal and P. K. Pramanik, "Fuzzy measures in determining seed points in clustering," *Pattern Recognition Letters*, vol. 4, pp. 159–164, 1986.

6. L. M. Belue and J. K. W. Bauer, "Determining input features for multilayer perceptrons," *Neurocomputing*, vol. 7, no. 2, pp. 111–121, 1995.

7. D. W. Ruck, S. K. Rogers and M. Kabrisky, "Feature selection using a multilayer perceptron," *Journal of Neural Network Computing*, pp. 40–48, Fall, 1990.

8. R. Battiti, "Using mutual information for selecting features in supervised neural network," *IEEE Trans. on Neural Networks*, vol. 5, no. 4, pp. 537–550, July, 1994.

Feature subset	Order obtained using	
	Eqn. (1)	*FEI* of Pal *et al.* [2,3]
$\{f_1\}$	5	3
$\{f_2\}$	1	2
$\{f_3\}$	7	6
$\{f_1, f_2\}$	2	1
$\{f_1, f_3\}$	6	7
$\{f_2, f_3\}$	4	4
$\{f_1, f_2, f_3\}$	3	5

Table 1: Values of FEI for every feature subset.

Feature	Initial w					
	$= 1.0$		in $[0, 1]$		$= 0.5 \pm \epsilon$	
	$E = 225.679036$ after 152 epochs		$E = 217.328403$ after 49 epochs		$E = 220.536857$ after 60 epochs	
	w	Rank	w	Rank	w	Rank
1	0.001194	3	0.000048	3	0.001037	3
2	0.342003	1	0.337536	1	0.342621	1
3	0.192297	2	0.001745	2	0.092156	2

Table 2: Degrees of importance of different features.

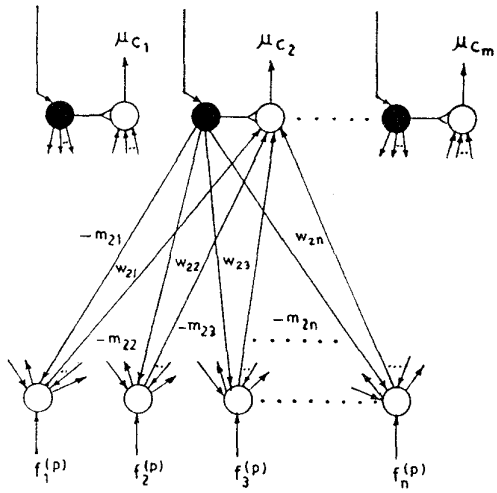

Fig. 1. Neural network model.

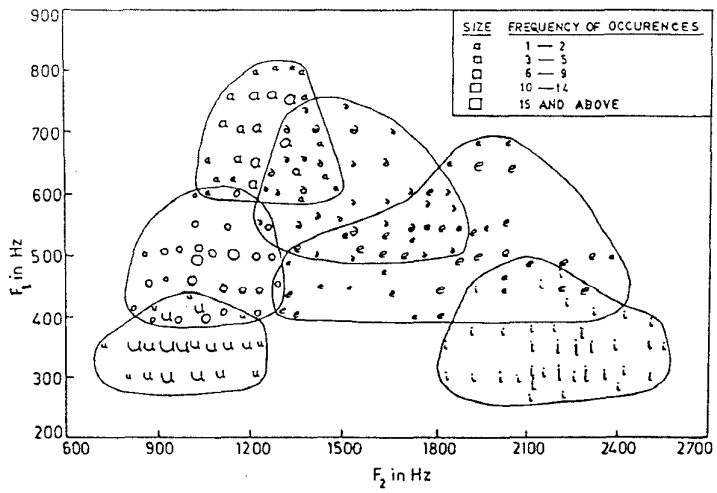

Fig. 2. 2-D projection of the data set.